Exploratory data analysis and checking assumptions

Experimental Design and Data Analysis for Biologists

G. Quinn and M. Keough

Design by M. Logan
2004

About

This presentation is a brief revision of basic statistical concepts.

Links…

Throughout the presentation, marks in the form of (qk2002, …) provide references to sections within the recommended statistical text;

> Quinn, G. P. and Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Words and phrases in purple type face provide tooltip-style extra information, while blue type face provide links to popups that contain additional information and or definitions.

Navigation…

Navigation buttons on the right hand side of each page provide (from top to bottom) 'Previous Page', 'Next Page', 'First Page', 'Last Page', 'Go Back' and 'Quit' navigational shortcuts.

Purpose of graphical displays

- **Exploratory data analysis (EDA)**
  - checking assumptions of parametric statistical analyses (normality, equal variance)
  - identifying unusual values (outliers)
  - evaluating the appropriateness of a particular statistical model
- **Analysis**
  - model fitting
- **Presentation & communication of results**

Type of analysis

- Many statistical analyses fit a model to data
- Statistical models usually contain
  - a response (dependent) variable (Y)

    continuous
  - predictor (independent) variable(s) (X_1, X_2)

    continuous and/or categorical

Linear models take the following form:

$$Y = \text{constant} + \text{coefficient}_1 \times X_1 + \text{coefficient}_2 \times X_2 + \ldots + \text{error}$$

Linear models - comparing groups

- Analysis of variance (ANOVA)
- *Y* is continuous, X_1, X_2… are categorical factors comprising 2 or more groups
- Partition variance in *Y* into
  - explained by model (between groups)
  - not explained by model (within groups or residual or error)
- Statistical Null Hypothesis (H_0):
  - *Y* population means of each group are equal ($\mu_1=\mu_2$ =...)

Example – comparing groups

Medley and Clements (1998) recorded the species diversity of diatom communities from between 4 and 7 stations at each of 4 zinc levels in streams in the Rocky Mountains (qk2002, Box 8.1).

- Is there a difference in the mean diatom species diversity between stream stations with different zinc levels?
  - *Y* is diatom species diversity
  - *X* is zinc-level groups (Background, Low, Medium, High)
  - Replicate units are the stations on the streams

Linear models - bivariate relationships

- Regression analysis
- *Y* is continuous
- X_1, X_2… are continuous
- Partition variance in *Y* into
  - explained by model (linear relationship with *X*s)
  - not explained by model (residual or error)
- Statistical Null Hypotheses:
  - no linear relationships between *Y* and X_1 or X_2.., i.e. the population slope between *Y* and X_1 and/or *Y* and X_2.. = 0

Example – bivariate relationships

Christenson et al. (1996) measured the density of riparian trees and the basal area of course woody debris (CWD) on the shoreline of 16 lakes in North America (qk2002, Box 5.3).

- Is there a linear relationship between CWD basal area and the riparian tree density?
  - *Y* is CWD basal area
  - *X* is riparian tree density
  - Replicate units are the 16 lakes

Assumptions of analysis

- Apply to response variable *Y* at each *X*
- Normality of observations (*Y*) at each *X*
  - symmetrical distribution is important
  - positive skew common with biological variables
- Similar variances of *Y* at each *X* (homogeneity of variances)
  - variances independent of means
- Independence of observations
  - design and data collection issue

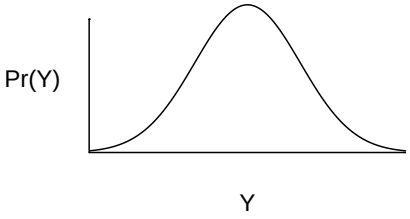
Checking assumptions – (EDA)

Explore the:

- Shape of sample (and therefore population)
  - is *Y* normally distributed (symmetrical) or skewed at each *X*?
- Spread of sample (and therefore population)
  - are *Y* variances similar for different *X*?

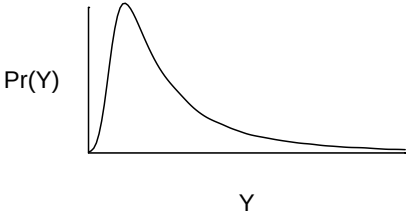
Exploring sample data

Distributions of biological data



Bell-shaped symmetrical distribution:

- Normal (Gaussian) distribution



Common skewed asymmetrical distributions:

- Log-normal
- Poisson

Common skewed distributions

Log-normal distribution:

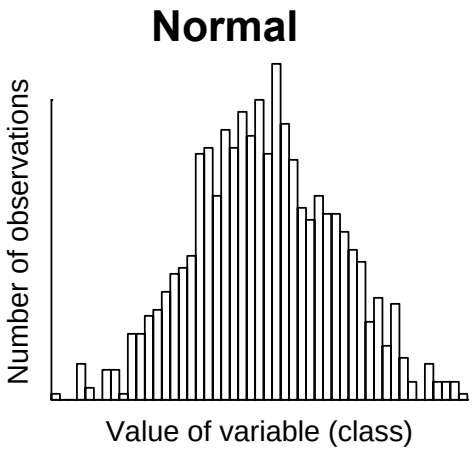
- when μ proportional to σ
- measurement data, e.g. length, weight ...

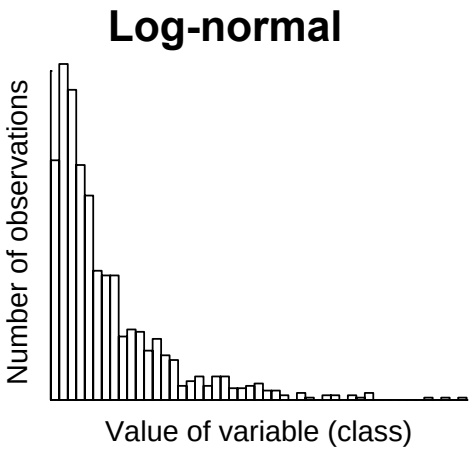
Poisson distribution:

- when $\mu = \sigma^2$
- count data, e.g. numbers of individuals

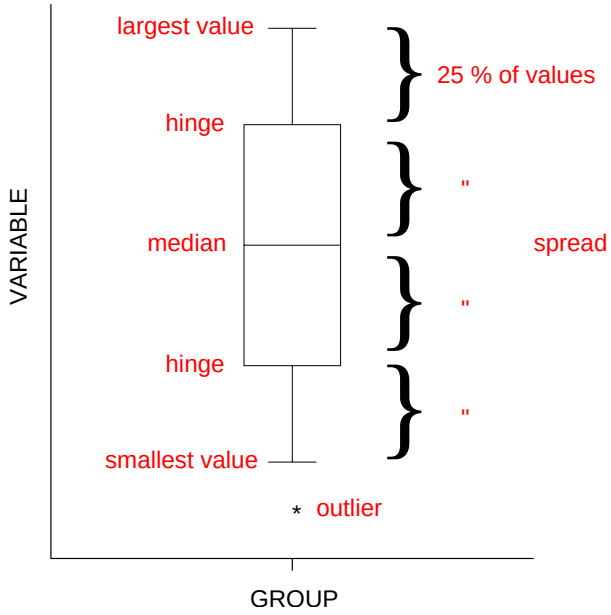
Frequency distributions

Observations grouped into classes (e.g. size or number).

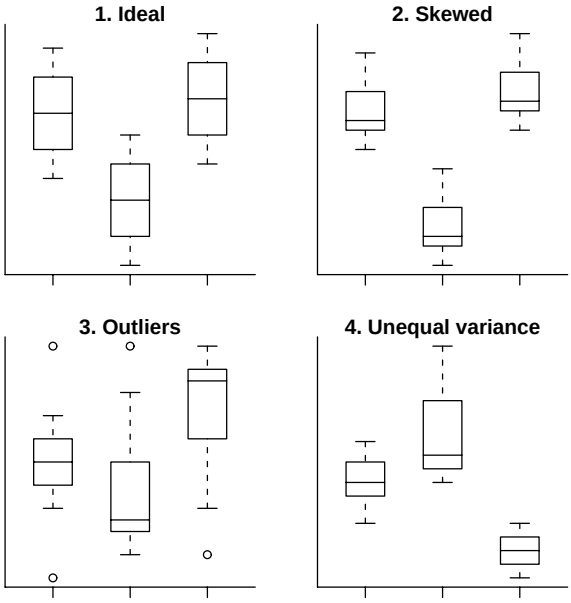




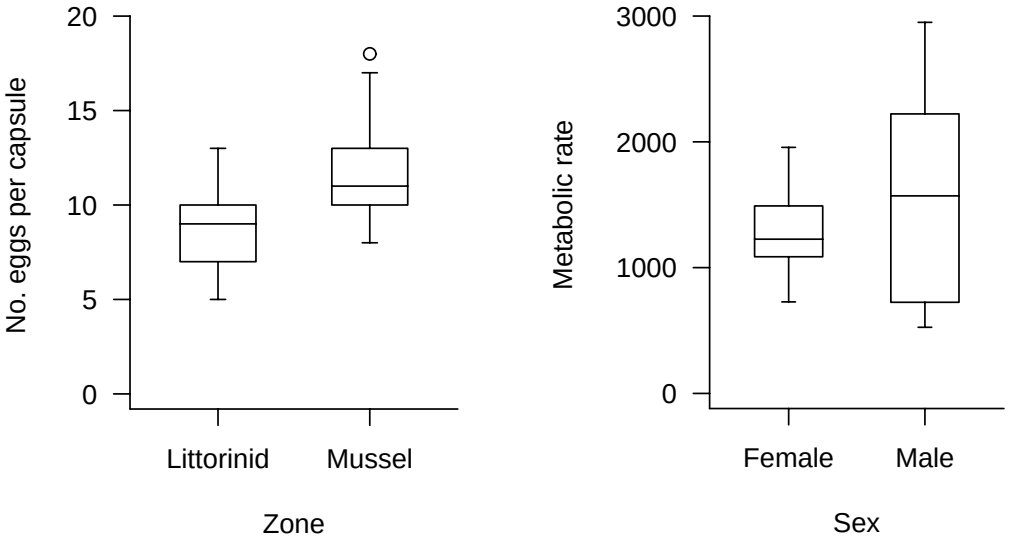
Boxplots


largest value
hinge
median
hinge
smallest value
* outlier
25 % of values
"
"
"
spread
VARIABLE
GROUP

Boxplots



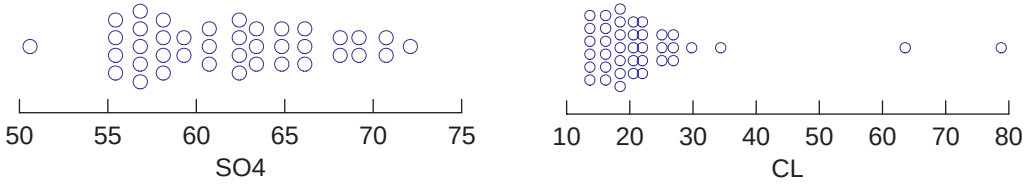
Examples of Boxplots



(qk2002, Fig 4.5)

Dotplots

- Each observation is represented by a dot
- For example, concentration of SO_4^{2-} and Cl^- for 39 sites from forested watersheds in North America (qk2002, Fig 4.5)

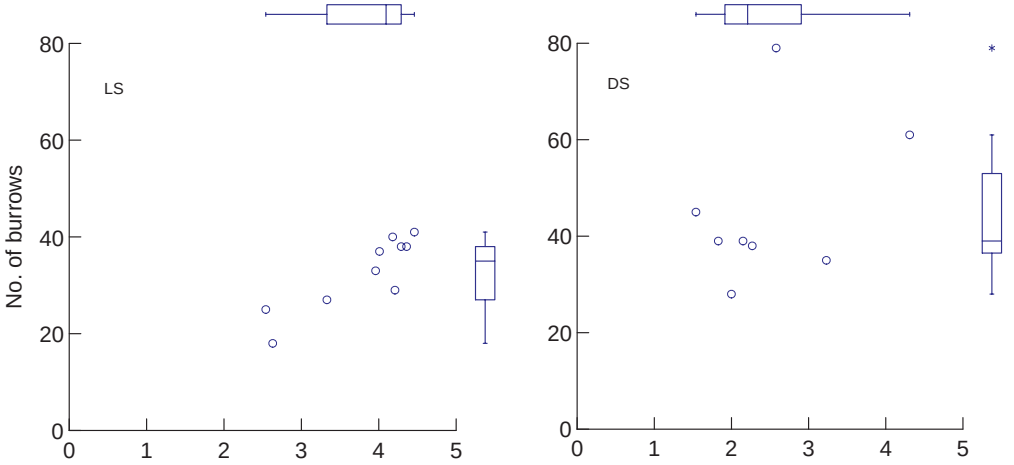


Scatterplots

- For plotting bivariate data
- Value of two variables are recorded for each observation
- Each variable is plotted on one of the axes (X or Y)
- Symbols (points) represent each observation
- Used to assess the relationship between two variables

Scatterplots

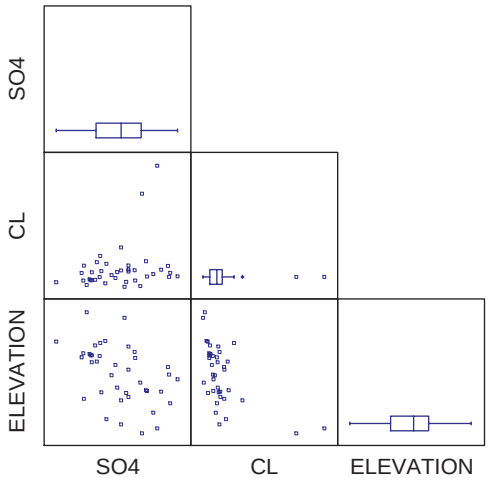
Relationship between the number of burrows and crab density for two sites on Christmas Island (qk2002, Fig 4.5)



Scatterplot matrix (SPLOM)

- Extension of scatterplot
- For plotting relationships between three or more variables on single graph
- Pairwise bivariate plots in multiple SPLOM panels
- Univariate plots (boxplots, histograms) in diagonal panels

SPLOM



SO4
- concentration of SO_4^{2-}

CL
- concentration of Cl^-

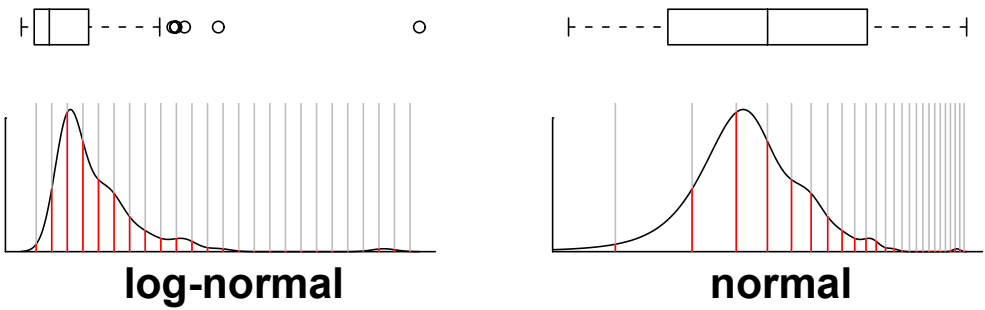
- ELEVATION
  - site elevation

- (qk2002, Fig 4.6)

Transformations

- Mainly used for response variable
- Improve normality
- Remove relationship between mean and variance, therefore make variances more similar in different populations
- Reduce influence of outliers
- Make relationships between variables more linear (regression analysis)
  - sometimes transform both response and predictor variables

Transformations



Vertical lines on both graphs represent corresponding data values and thus the relative spacing of data on the *X*-axis. Note that transformations only alter the scale of the data, they DO NOT change the order of the data

Transformations

Log transformations

Lognormal  Normal

$Y = \log(Y)$

Log transformations are often useful for normalizing measurement data – use log(Y+c) where c is constant for data with zeros.

Power transformations

Poisson  Normal

$Y = \sqrt[n]{Y}$ (i.e. $Y = \sqrt{Y}$, $Y = \sqrt[4]{Y}$)

Power transformations are often useful for normalizing count data – use 4th root to correct more extreme skew

Arcsine √ transformations

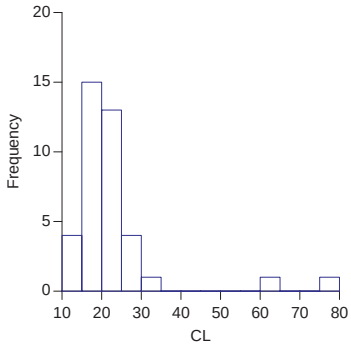
Square  Normal

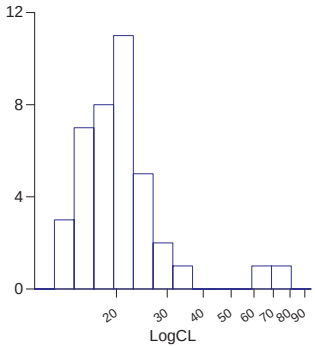
$Y = \sin^{-1}(\sqrt{Y})$

Arcsine √ transformations are occasionally useful for normalizing proportions

Example – Log-transformation

Frequency distributions for raw and transformed (log_{10}) Cl^- concentration from forested North American streams. (qk2002, Fig 4.9)





Log scale of *X*-axis mimics a log transformation

Outliers

- **Observations very different from rest of the sample**
  - identified as $*$ or o in boxplots (see slide 15)
- **Observations that are far from fitted model**
  - identified as large residuals from fitted model
  - can have large influence on estimates of model parameters and statistical tests
- **Two types of outliers might be different**
  - latter more important for linear models

Dealing with outliers

- **Check if outliers are mistakes**
  - error in data entry or measuring equipment
  - if so, omit value
- **Extreme values in a skewed distribution**
  - transform data
- **Alternatively, run analysis twice**
  - outliers included vs outliers excluded
  - if outcome and conclusions differ then outliers are influential – consider robust analyses

If assumptions are not met

- **Check and deal with outliers**
- **Transformation**
  - might fix non-normality (and outliers) and unequal variances
- **Alternative robust analyses**

Alternative analyses

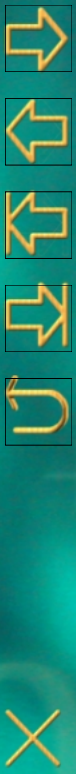
- **Generalised linear models (qk…)**
  - models that can handle a range of distributions
  - mormal, log-normal, poisson etc.
  - require specific software (R, S-Plus etc.)
- **Robust parametric tests (qk…)**
  - can handle unequal variances
  - only for simple models (single predictors)






Alternative analyses (cont.)

- **Non-parametric rank tests (qk…)**
  - do not assume any specific distribution (e.g. normality)
  - usually assume equal distribution shapes between groups (i.e. equal variances)
  - only suitable for simple analyses – do not deal well with interactions or complex models
- **Non-parametric randomization tests (qk…)**
  - do not assume any specific distribution (e.g. normality)



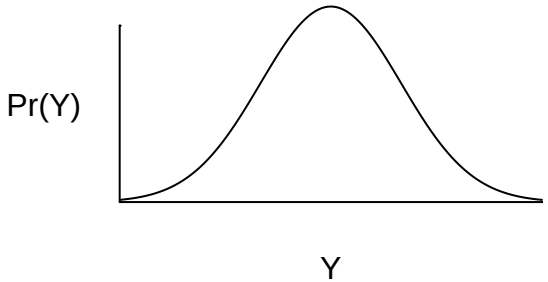
- only suitable for simple analyses – do not seem to deal well with interactions or complex models
- useful for non-random sampling situations or unusual data types
- require specific software






Normality

Normality refers to the state of a variable that is normally distributed. The normal (or Gaussian) distribution is a symmetrical probability distribution with a characteristic bell-shape. Statistical procedures that use sample means to characterize populations, assume that the observations that make up the sample (and thus the population) are normally distributed. Likewise, measures of the spread of data (often based on deviation from the center –mean) assume equal spread either side of the mean.



Click anywhere to close

Homogeneity (equality) of variance

hypothesis tests for linear models (based on ordinary least squares estimation), assume that the response variable is equally variable at each level of the predictor variable (or combination of levels when multiple predictor variables). If this assumption is not met, interval estimates and statistical tests may be unreliable.

Unequal variances are often the result of non-normality. When the response variable follows a lognormal or a Poisson distribution, a relationship between mean and variance (of the response variable at each level of predictor variable) is expected. Often, appropriate normalizing transformations will also improve the degree of homogeneity. Alternatively, unequal variances can also be caused by unusual values (ouliers).

Model fitting

Statistical models are fitted to data to test the effect of one or more predictor variables (continuous or categorical) on a response (dependent) variable. Such models take on the form:

$$response\ variable = model + error$$

where the *model* component incorporates the predictor variable(s) and the parameters that relate each predictor variable to the response variable.

Click anywhere to close

Linear models

The term 'linear' in linear model **does not** refer to the shape of the relationship between predictor variable(s) and the response variable. That is, it does not imply a straight-line relationship. The term 'linear' refers to the linear combination of parameters in the statistical model. That is, the parameters that relate the predictor variable(s) to the response variable are neither exponents, nor are they multiplied by or divided by any other parameter.

$Y = \alpha \times X + error$ and $Y = X^{\alpha} + error$ are linear models

$Y = 2^{\alpha} \times X + error$ and $Y = {}^{\alpha}/_{\beta} \times X + error$ are not

Click anywhere to close

Statistical null hypothesis

Logically, predictions cannot be proved –only disproved. Therefore to investigate a hypothesis, we attempt to disprove a null hypothesis. A null hypothesis therefore covers all possible outcomes except the prediction in the hypothesis.

A statistical null hypothesis is a null hypothesis expressed in terms of a statistical hypothesis test. For example a regression analysis tests the null hypothesis that there is no relationship between two (or more) continuous variables. Statistically, regression analysis tests whether the population regression slope between the variables is equal to 0.

Click anywhere to close

Linear relationship

Unlike, in the phrase 'linear model', the term 'linear' in linear relationship or linear regression **does** refer to the shape of the relationship between predictor variable(s) and the response variable. A significant outcome from a linear regression, suggests that the response variable is linearly related to the predictor variable(s). Hence the relationship between one predictor variable and a response variable can be characterized by a straight line with a uniform slope and a *y*-intercept:

$$Y = intercept + regression\ slope \times X$$

Click anywhere to close

Parametric vs non-parametric analyses

Parametric tests make distributional assumptions about the population(s) from which the data were sampled. Hence, they can only be applied to data from which the probability distribution(s) for the sampled population(s) can be specified. As an example, statistical tests that are based on the *t* distribution assume that the populations from which the samples were collected are normal. Contrastingly, **nonparametric** rank based tests do not make any distributional assumptions since they generate their own probability distribution for the particular test statistic. Observations are ranked and the ranks are then randomized a large number of times (each time recalculating the test statistic) to generate a probability distribution of the rank-based test statistic.