Data Analysis

Useful references:

http://www.biostathandbook.com/analysissteps.html

Quinn and Keogh
Experimental Design and Data Analysis for Biologists
http://www.lacbiosafety.org/wp-content/uploads/2011/09/experimental-design-and-data-analysis-for-biologists1.pdf

Step-by-Step

1. Specify the biological question
2. Construct a NULL and ALTERNATIVE hypothesis
3. Identify your response/dependent and explanatory/independent variables
4. Identify what kind of variable they are
5. Summarise and analyse the results
6. Present the results

Biological Question

are trees bigger on south facing slopes?

are males more aggressive than females?

Write some more examples here

Hypothesis H0:H1

H0 (null): Trees on the south facing slope are the same height as trees on the north facing slope.

H1 (alt): Trees on the south facing slope are a different height to trees on the north facing slope.

Write some more examples here

Identify Variables

Response or Dependent variable. What you measure, e.g.
Tree height

Independent or Explanatory variable. The variable that is influencing the response, e.g. Slope orientation

Write some more examples here

What kind of variable

To decide on the statistical test you need to first identify what sort of variables you have

Numeric/Continuous

take any number (-/+), whole numbers or decimals

Nominal/Discrete

take discrete categories; sex (M or F), tree growth form (tree, shrub, vine), presence/absence, count data (non-negative whole numbers; 1,2,3).

Ordinal/Ranked

observations are put into order smallest-largest, relative ranking

Why is the variable type important?

Determines which statistical test is appropriate for analysis

Parametric

Makes assumptions about the parameters you are estimating

Not Parametric

Assumption-free, makes no assumptions about the parameters you are estimating

What is the deal with parametric and non-parametric?

Parametric statistical tests have three main assumptions

1. Normal distribution
2. Homogeneous variance
3. Independent observations

What is the deal with parametric and non-parametric?

Parametric statistical tests have three main assumptions

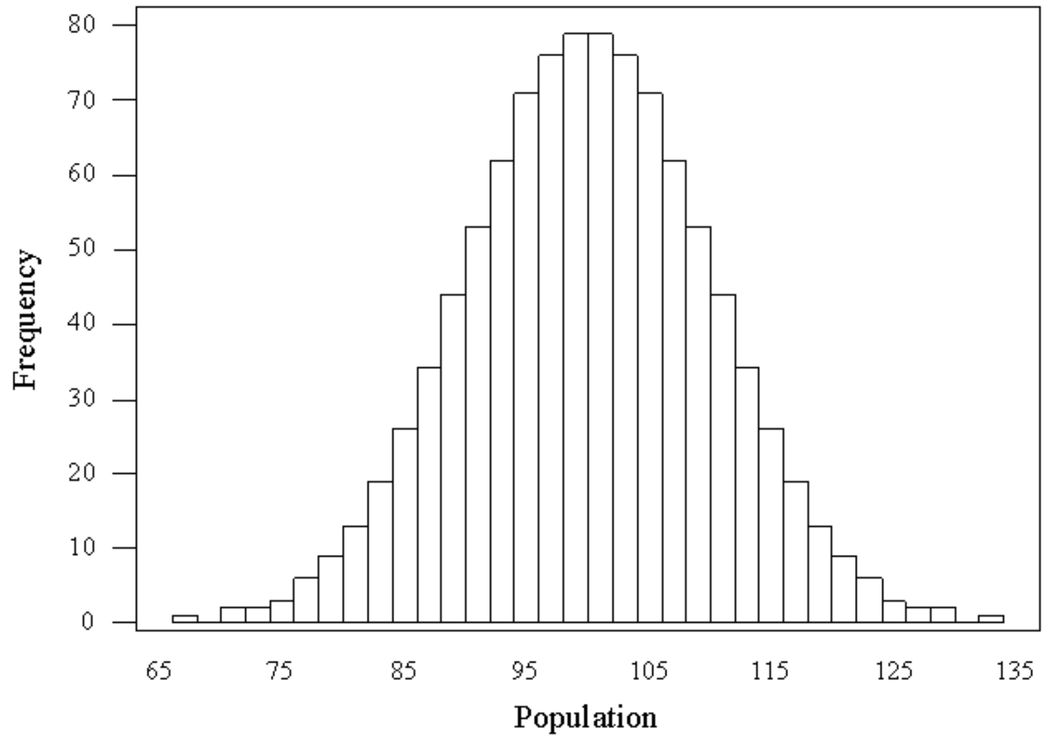
1. Normal distribution
2. Homogeneous variance
3. Independent observations

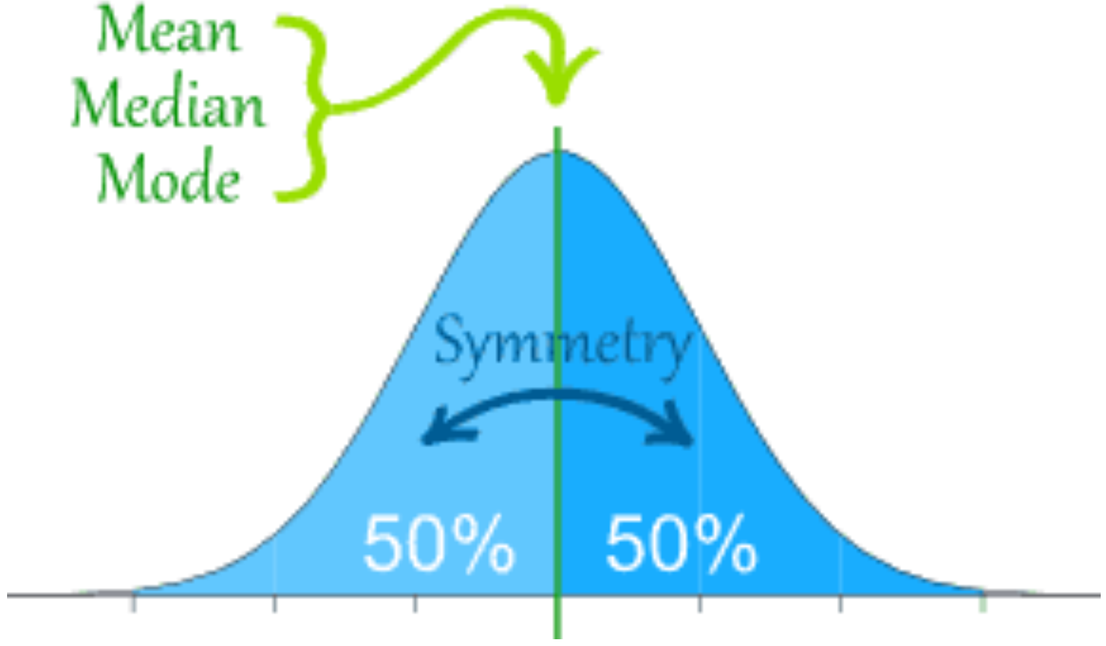
If any one of these assumptions are violated you can NOT use parametric statistical tests

What is the deal with parametric and non-parametric?

1. Normal distribution

common for numeric/continuous variables

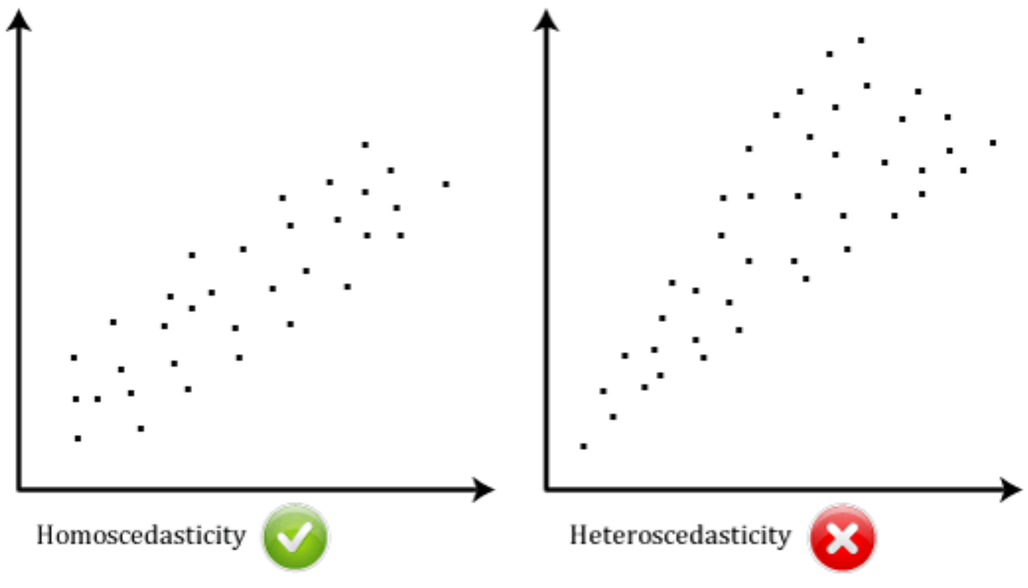


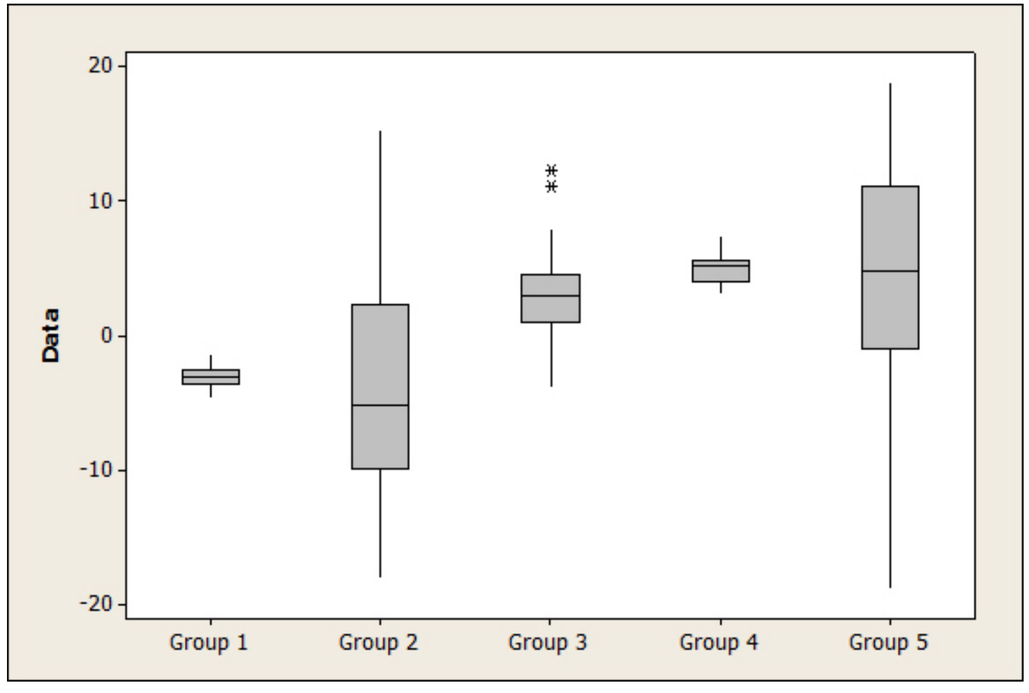


What is the deal with parametric and non-parametric?

2. Homogeneous variation (homoscedasticity)

variance between groups is the same





Unequal variance between groups = heteroscedasticity

What is the deal with parametric and non-parametric?

2. Independent observations

each observation is not related or correlated with another

Violation: observations taken close together in space or time, or if they are taken from related individuals or the same individual, they will be more similar to each other than observations taken further apart or taken from unrelated/different individuals

e.g. eggs in a clutch, observations across days, repeated measures of an individual, measurements taken on trees close together

Examples of tests for different types of variables

Independent Variable	Continuous	Discrete, 2 groups	Discrete, >2 groups
Response Variable			
Numeric/Continuous (P or NP)	Pearson Correlation (P) Kendall/spearman correlation (NP) Linear Model (P)	Student t-test (P) Mann-Whitney U (NP) Sign test (NP)	anova (P) Kruskal-Wallis (NP)
Nominal/Discrete (NP)	Generalised linear model	Chi-square test	GLM Chi-square test
Ordinal/Rank (NP)	Spearman rank correlation	Kruskal-Wallis Wilcox signed rank Kendell's W	Kruskal-Wallis Wilcox signed rank Kendell's W

P = Parametric: Normal distribution, homogeneous variance, independent observations

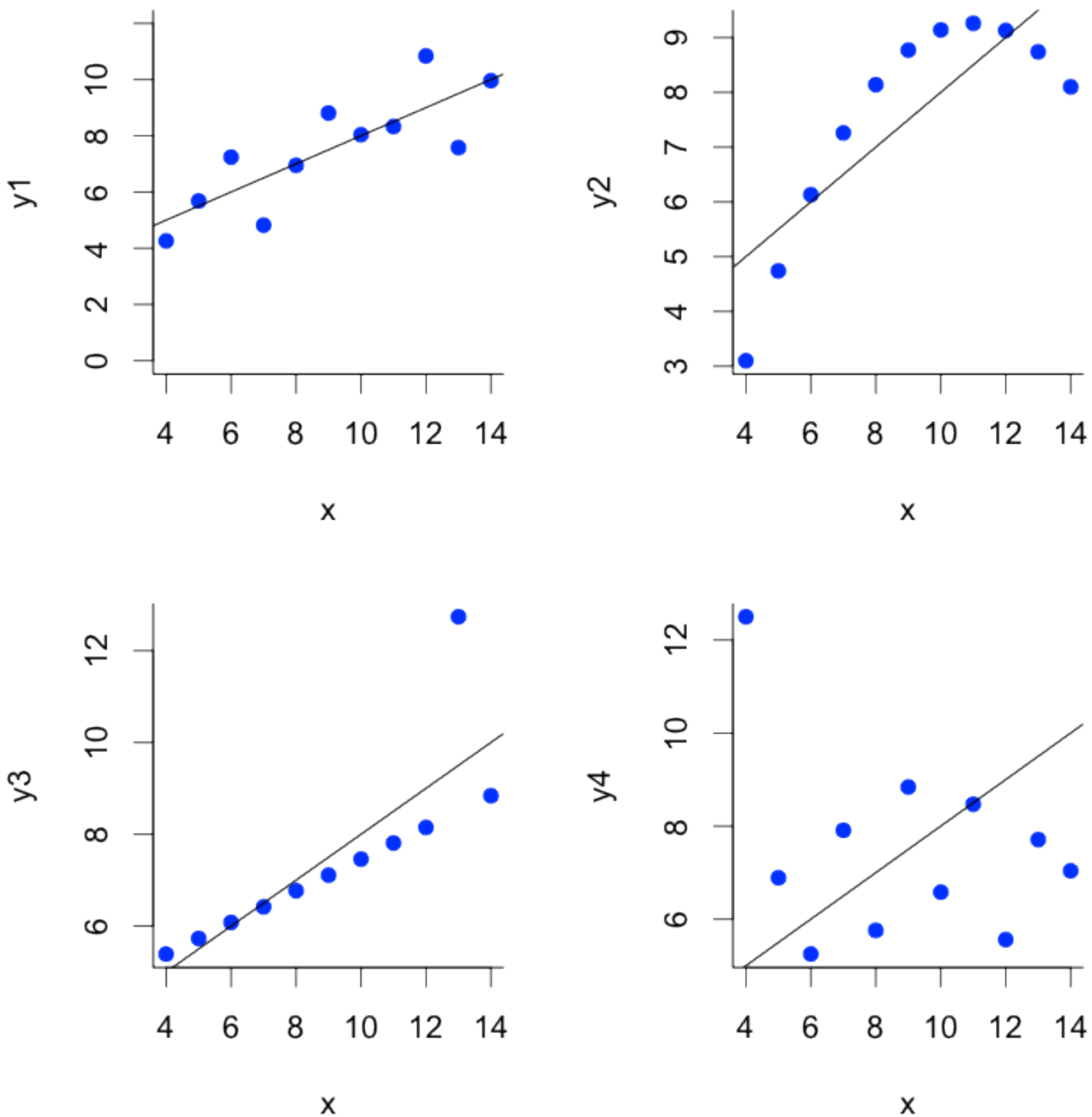
NP = Non-Parametric: non-normal distribution

Summarise and analyse your data

1. plot your data - is there a clear pattern? are there outliers?

scatterplot, boxplot, bar plot of means ± standard error

PLOT YOUR DATA



ALWAYS plot your data
Look for outliers
Look at the spread

Describing your data in terms of means and variance only can be misleading

All these graphs come from data with the same mean and variance
mean = 7, var=4.12

Summarise and analyse your data

1. plot your data - is there a clear pattern? are there outliers?

    scatterplot, boxplot, bar plot of means±se

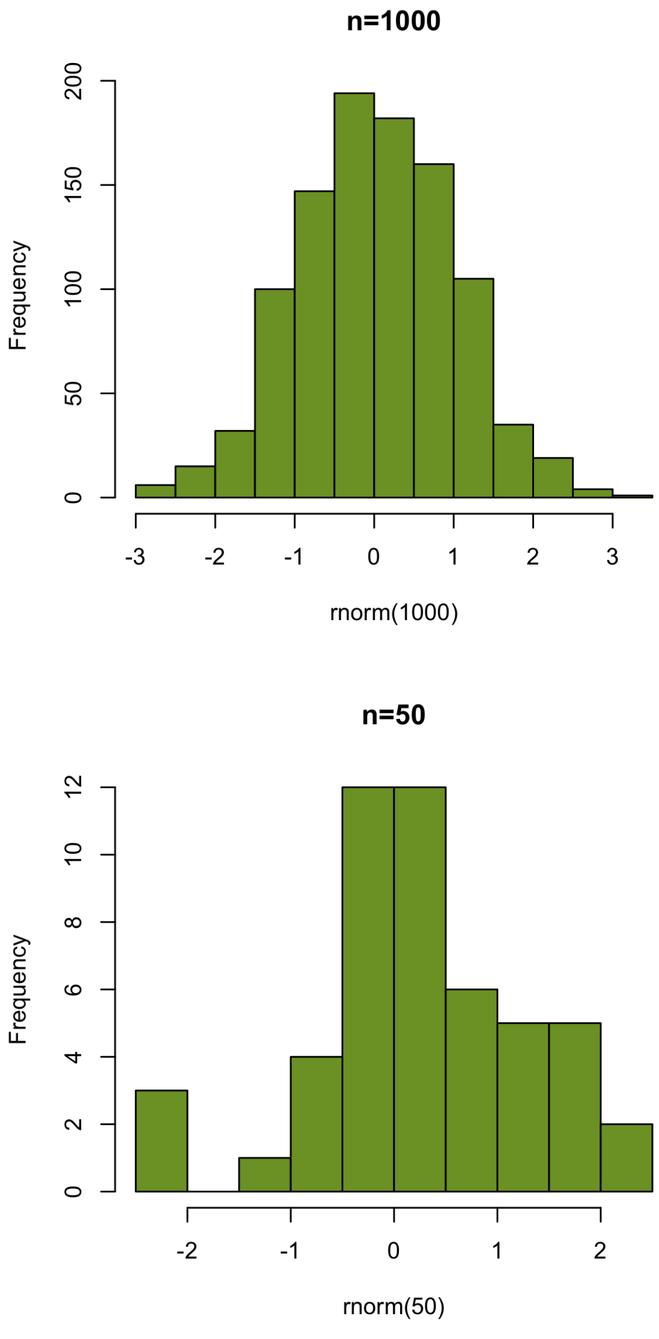
2. are your data parametric?

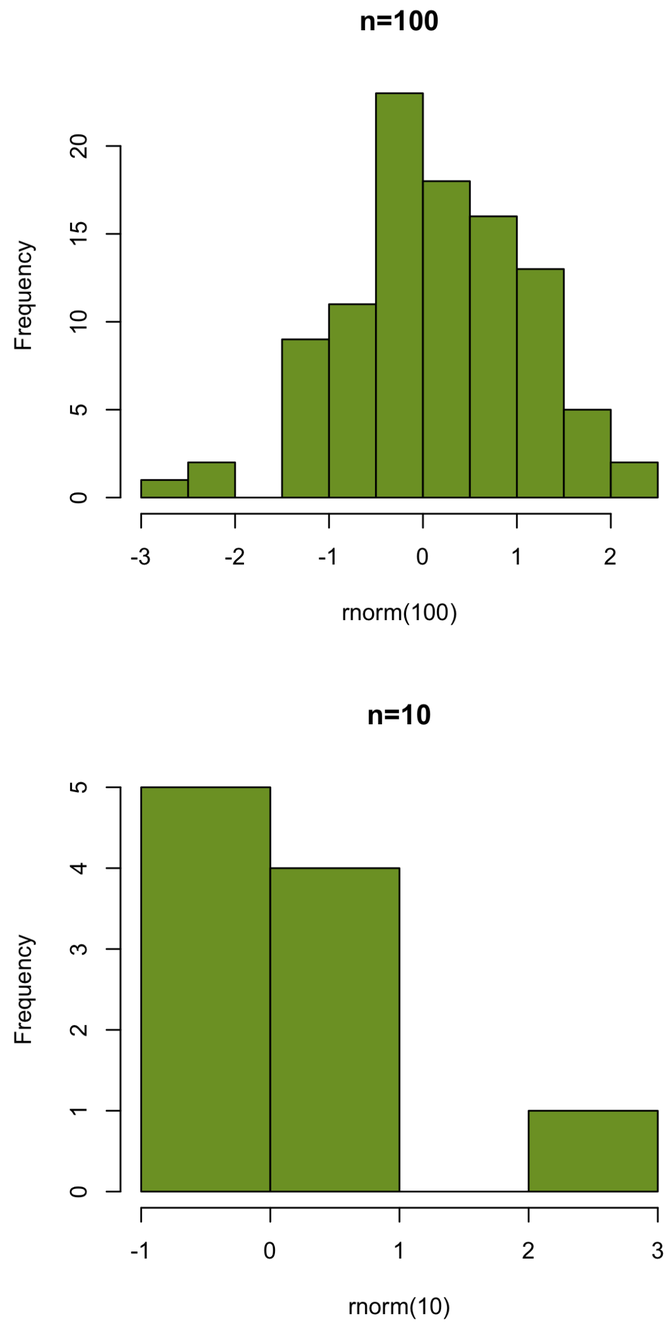
    normally distributed

    equal variance

    independent observations

Distribution and sample size





It is difficult to determine the distribution of your data when you have few observations (n is small)

These data all come from a normal distribution with mean=0 and var=1

Present the results

1. make a graph OR a table

   don't present the same data twice - once in a table and once in a graph

2. clearly summarise the results in the text and refer to the figures/graph/table

   e.g. Trees on the south side of the mountain were taller than trees on the north side of the mountain (Figure 1, $t(8) = 3.064$, $p=0.012$).