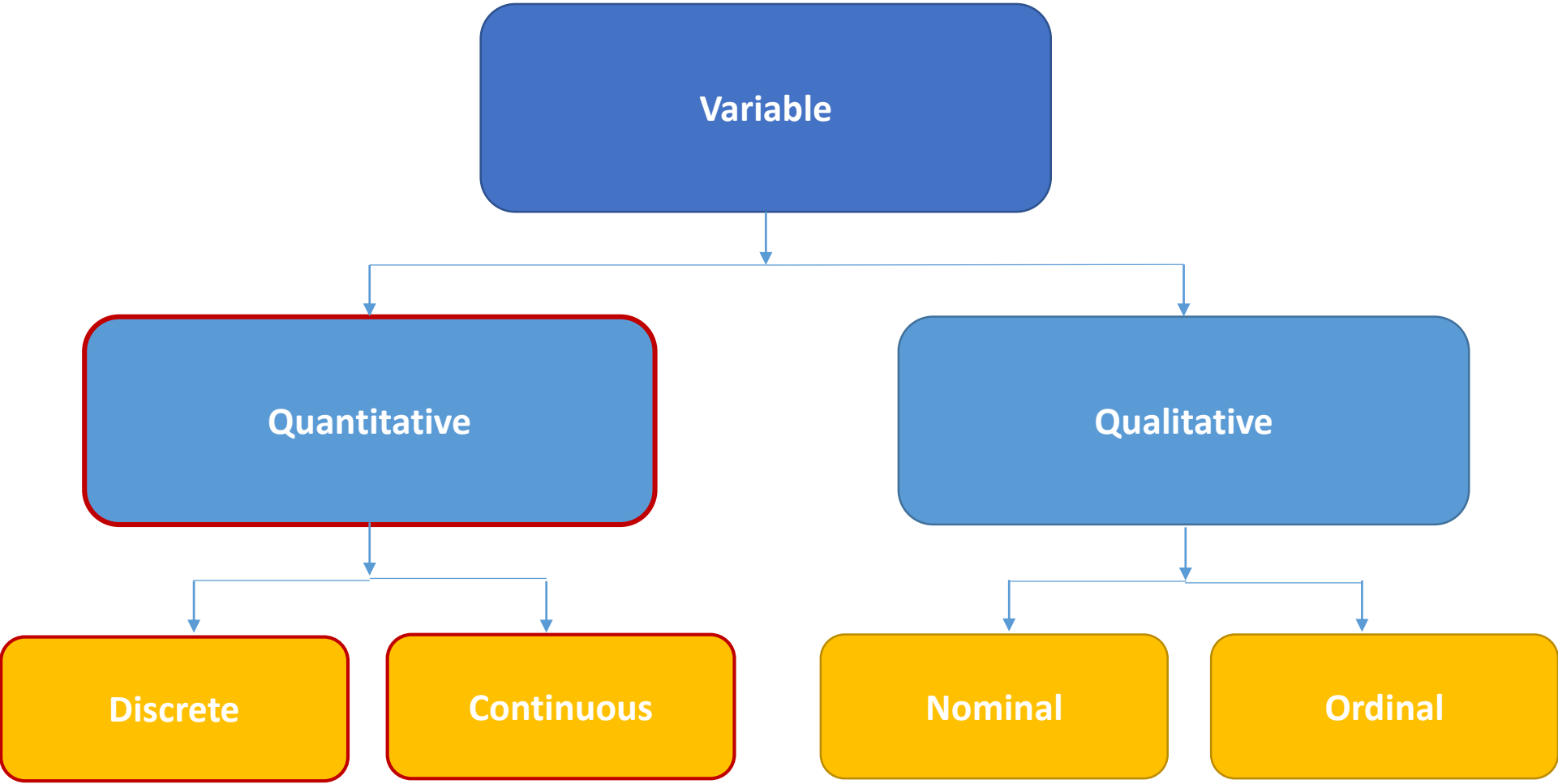




Descriptive Stats and Data Exploration

Anne Segonds-Pichon
v2020-09





Quantitative data

- They take **numerical values** (units of measurement)
- Discrete: obtained by counting
 - Example: number of students in a class
 - values vary by finite specific steps
- or continuous: obtained by measuring
 - Example: height of students in a class
 - any values



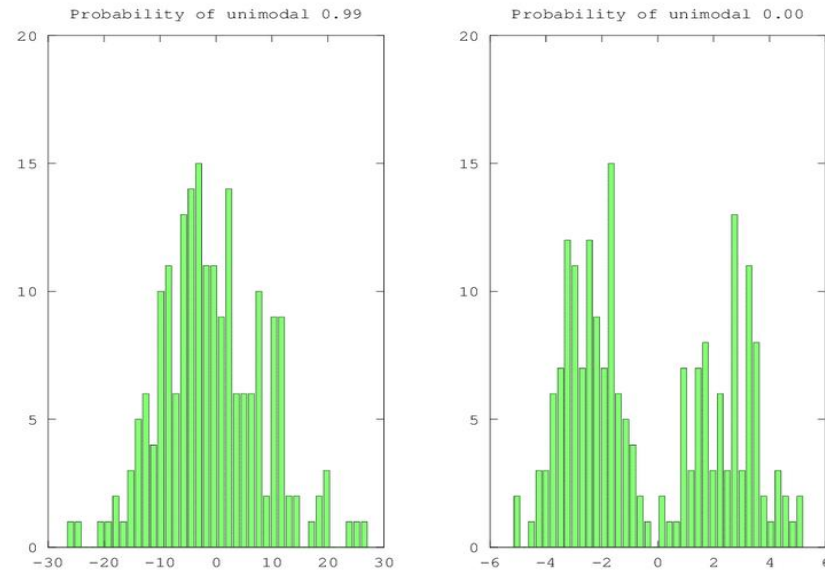
<https://github.com/allisonhorst/stats-illustrations#other-stats-artwork>

- They can be described by a series of parameters:
 - **Mean, variance, standard deviation, standard error and confidence interval**

Measures of central tendency

Mode and Median

- **Mode:** most commonly occurring value in a distribution



- **Median:** value exactly in the middle of an ordered set of numbers

Example 1: 18 27 34 52 54 59 61 68 78 82 85 87 91 93 100, Median = 68

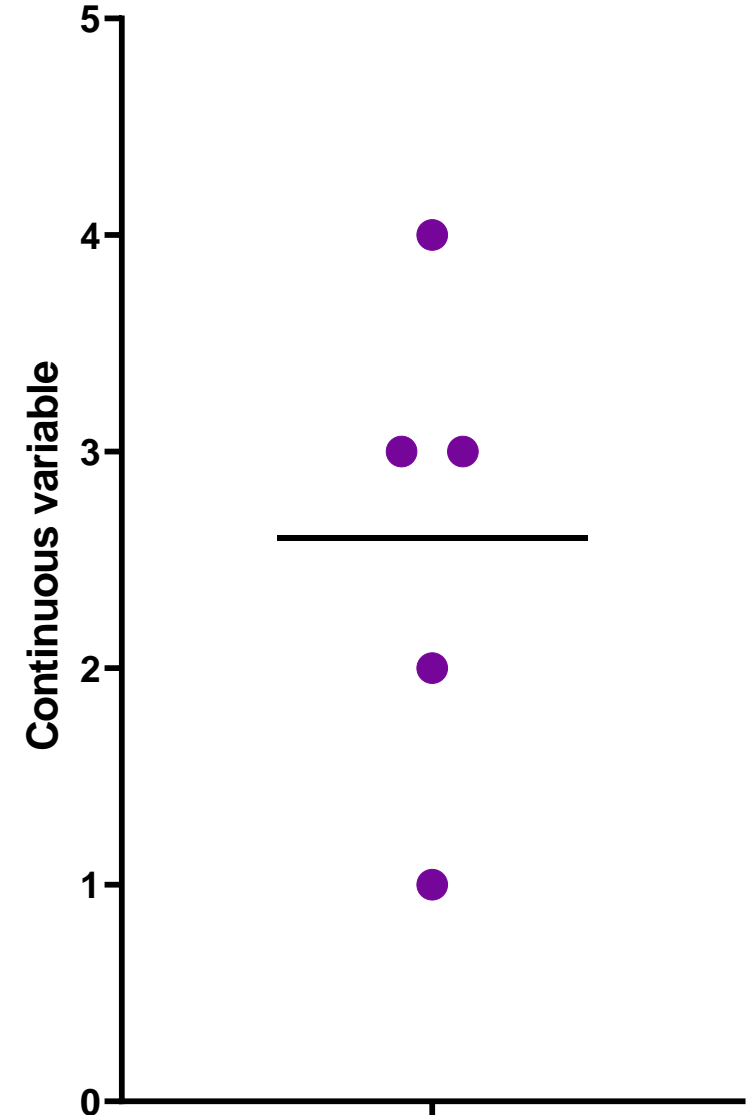
Example 2: 18 27 27 34 52 52 59 61 68 68 85 85 85 90, Median = 60



Measures of central tendency

Mean

- Definition: **average of all values in a column.**
- Example: mean of: 1, 2, 3, 3 and 4
 - $(1+2+3+3+4)/5 = 2.6$
- The mean is a **model** because it summarizes the data.
- How do we know that it is an **accurate model**?
 - Difference between the real data and the model created



Measures of dispersion

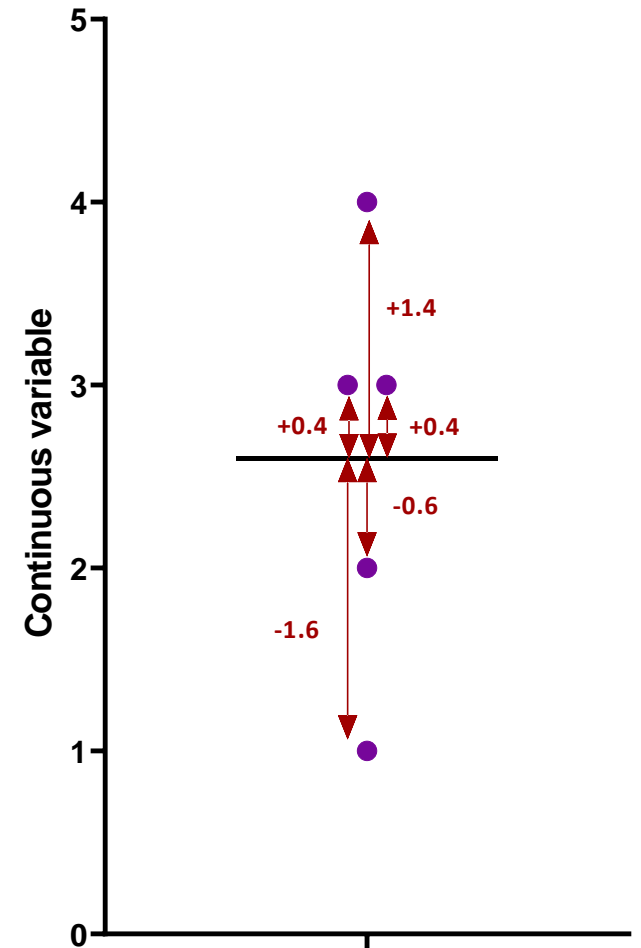
- Calculate the magnitude of the differences between each data and the mean

- Total error = sum of differences

$$= \sum(x_i - \bar{x}) = -1.6 - 0.6 + 0.4 + 0.4 + 1.4 = 0$$

No errors !

- Positive and negative: they cancel each other out.



Sum of Squared errors (SS)

- To solve that problem: we square errors

– Instead of sum of errors: **sum of squared errors (SS)**:

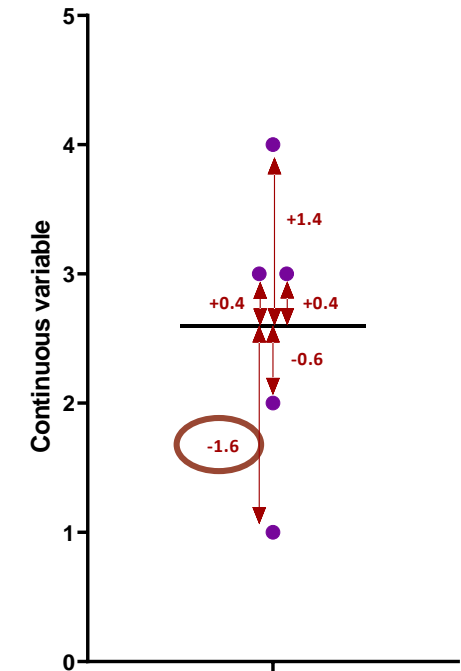
$$\begin{aligned}(SS) &= \sum(x_i - \bar{x})(x_i - \bar{x}) \\ &= (-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2 \\ &= 2.56 + 0.36 + 0.16 + 0.16 + 1.96 \\ &= 5.20\end{aligned}$$

- SS gives a good measure of the accuracy of the model

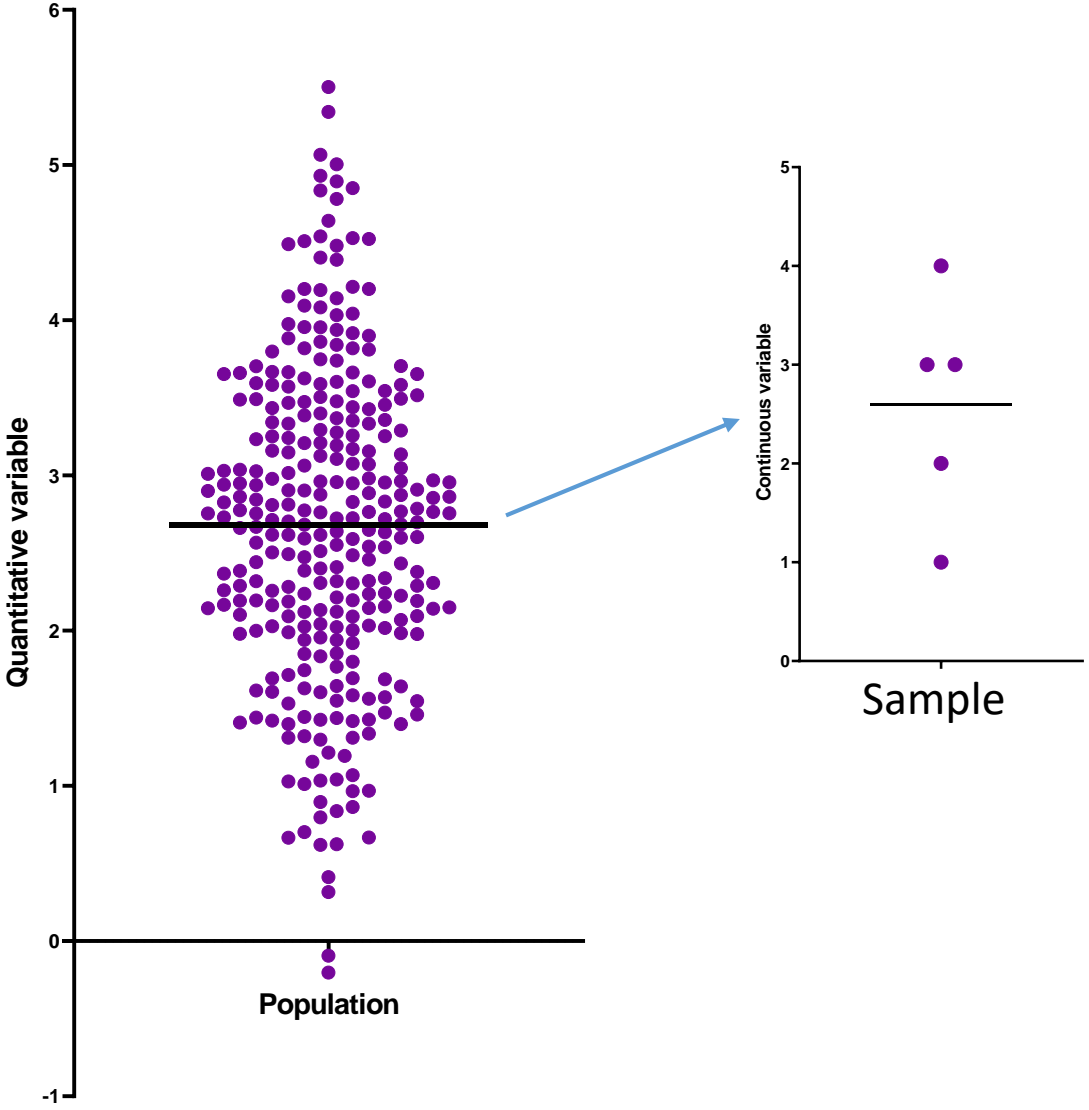
– But: dependent upon the amount of data: the more data, the higher the SS.

– Solution: to divide the SS by the number of observations (N)

- As we are interested in measuring the error in the sample to estimate the one in the population, we divide the SS by N-1 instead of N and we get the **variance (S^2)** = SS/N-1



Degrees of freedom



Mean Population (μ) = Mean Sample (\bar{x}) = 2.6

$$\bar{x} = 2.6 = (1+2+3+3+4)/5 = 2.6$$

First (n-1) values: whatever

nth value: fixed

n - 1 degrees of freedom

Variance and standard deviation

- $variance (s^2) = \frac{SS}{N-1} = \frac{\sum (x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.3$

- Problem with variance: measure in squared units

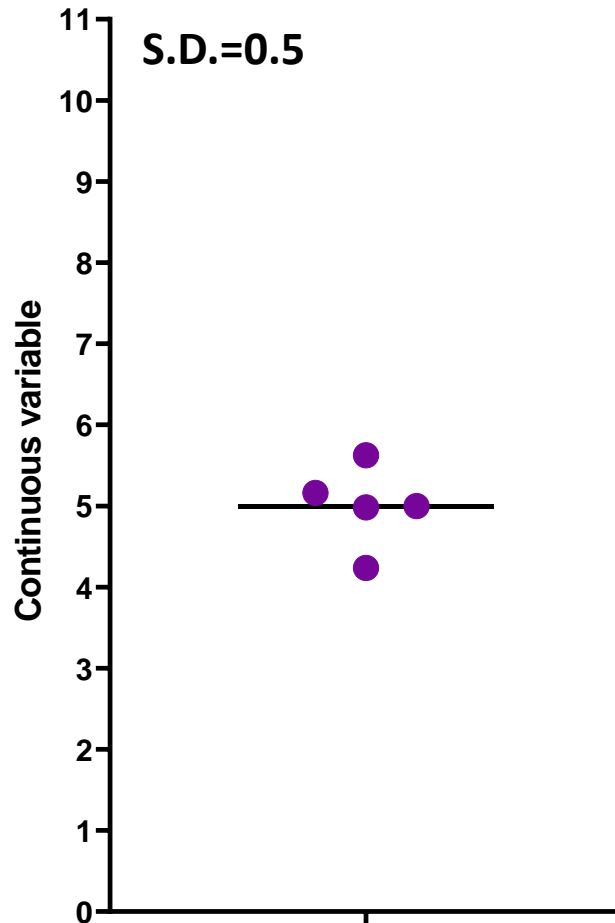
- The square root of the variance is taken to obtain a measure in the same unit as the original measure:

- the **standard deviation**

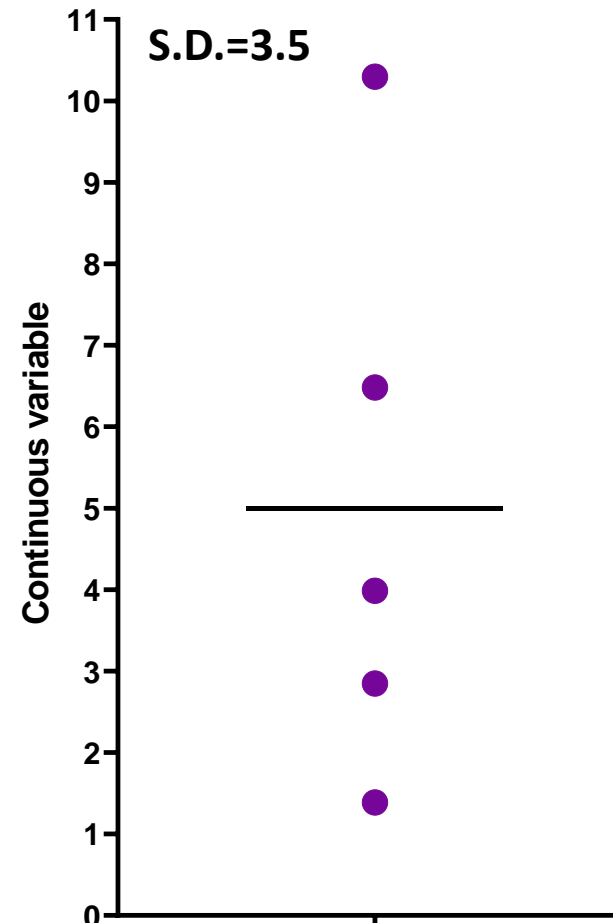
- S.D. = $\sqrt{SS/N-1} = \sqrt{s^2} = s = \sqrt{1.3} = 1.14$

- The **standard deviation** is a measure of how well the mean represents the data.

Standard deviation

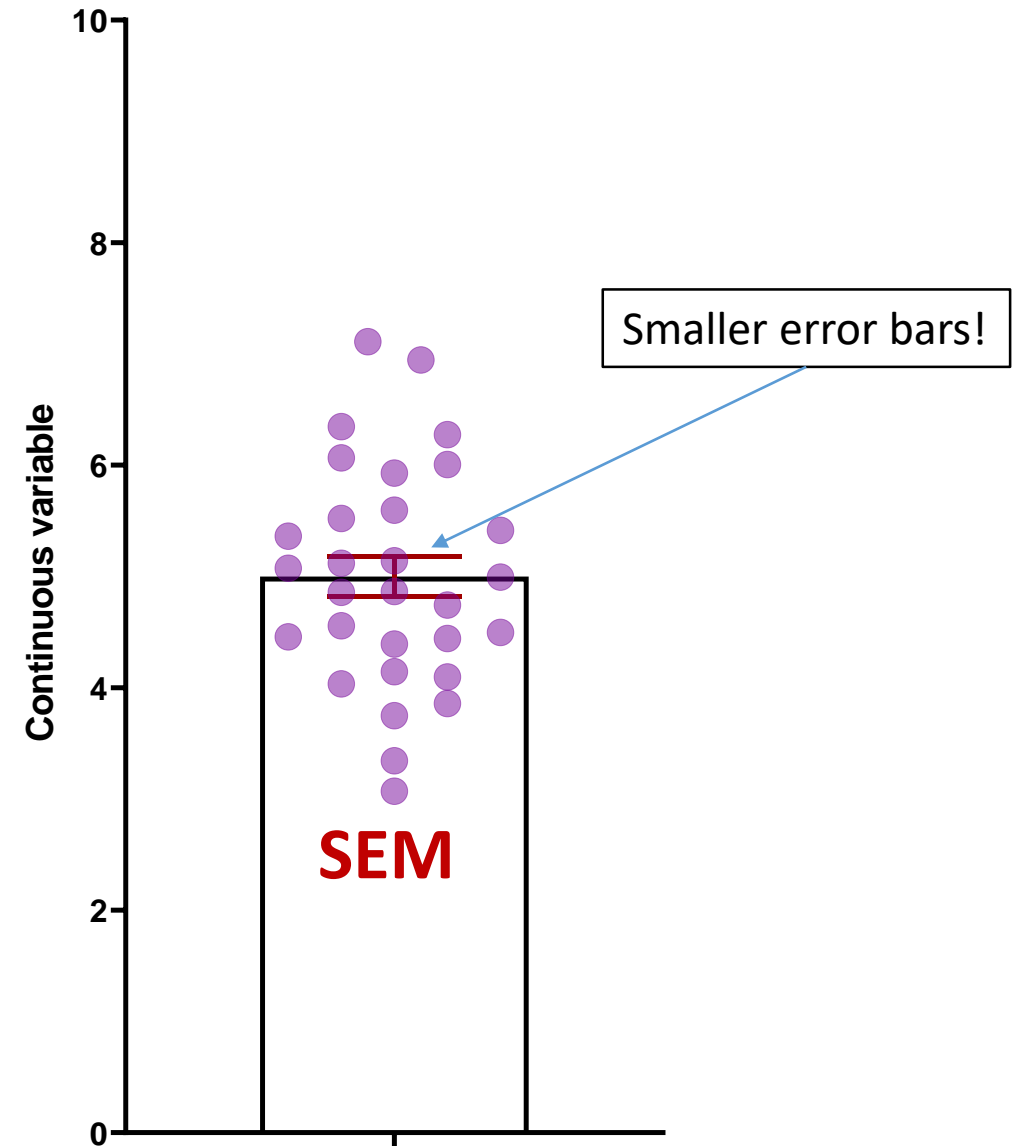
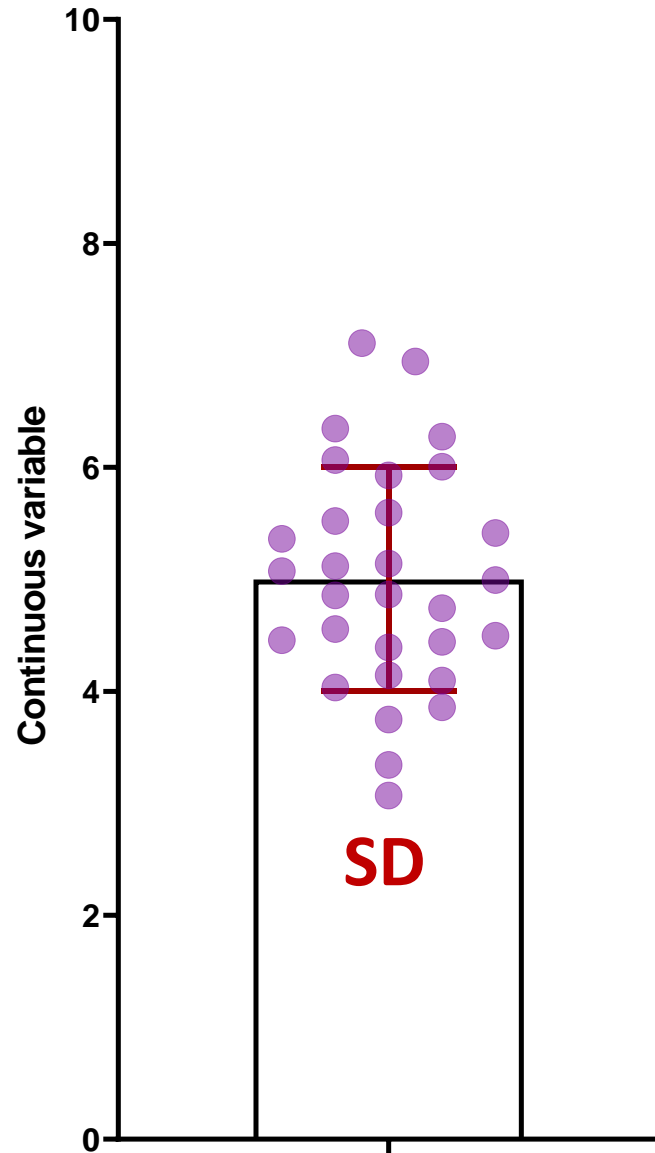


Small S.D.:
data close to the mean:
mean is a **good fit** of the data



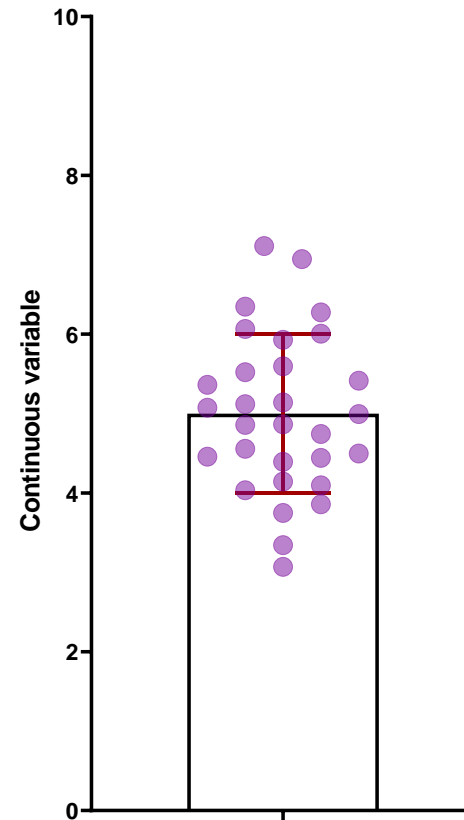
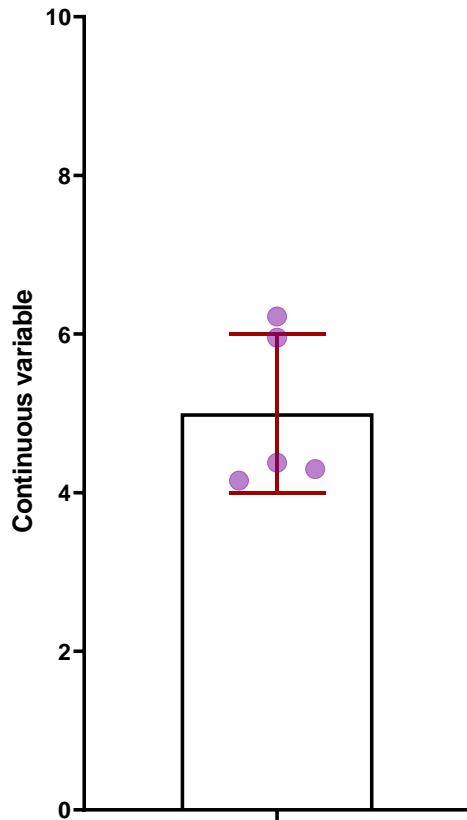
Large S.D.:
data distant from the mean:
mean is **not an accurate representation**

Standard Deviation (SD) or Standard Error Mean (SEM)?



Standard Deviation

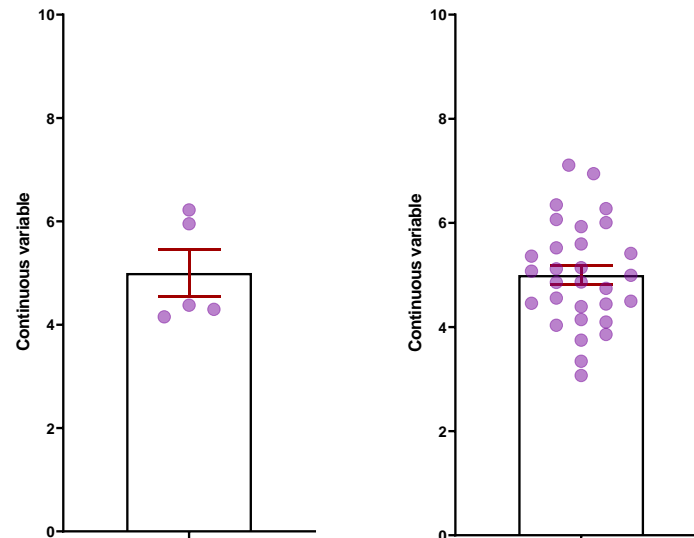
- The **SD** quantifies **how much the values vary** from one another
 - **scatter or spread**
 - The **SD** does not change predictably as you acquire more data.



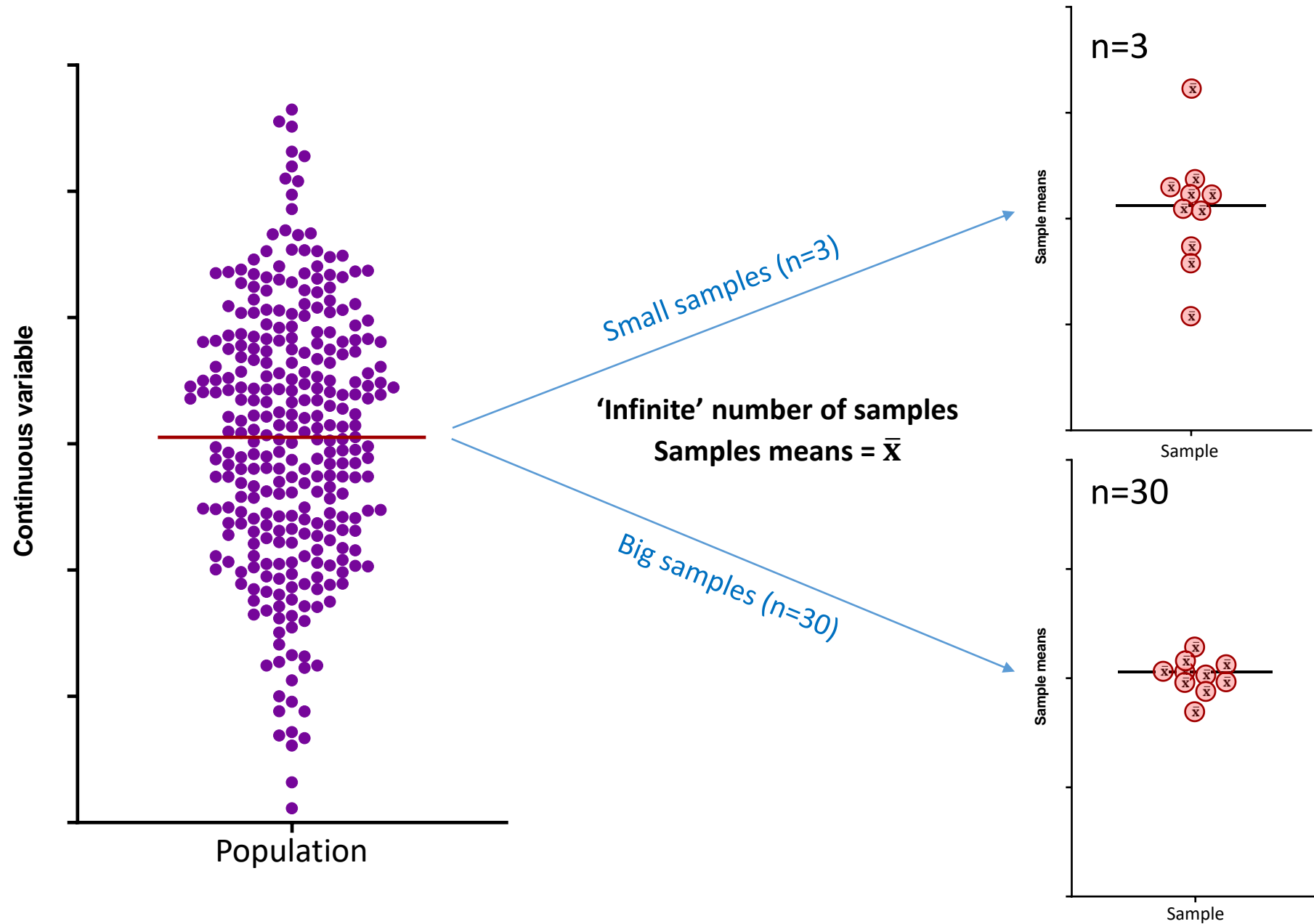
Standard Error Mean

$$\text{SEM} = \frac{\text{SD}}{\sqrt{N}}$$

- The **SEM** quantifies **how accurately** we know the **true mean** of the population.
 - Why? Because it takes into account: **SD + sample size**
- The SEM gets smaller as your sample gets larger
 - Why? Because the mean of a large sample is likely to be closer to the true mean than is the mean of a small sample.



The SEM and the sample size



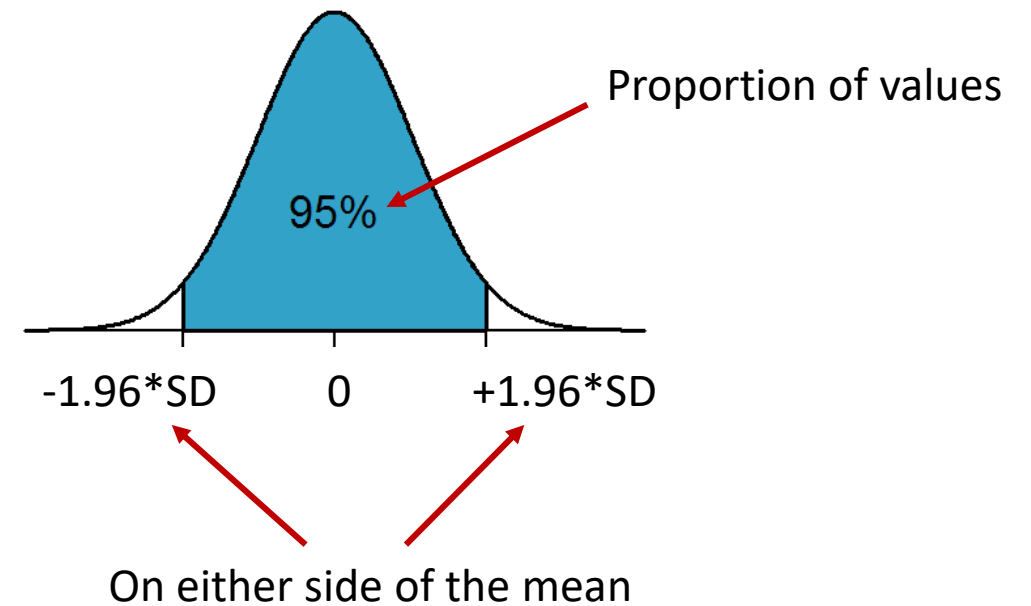
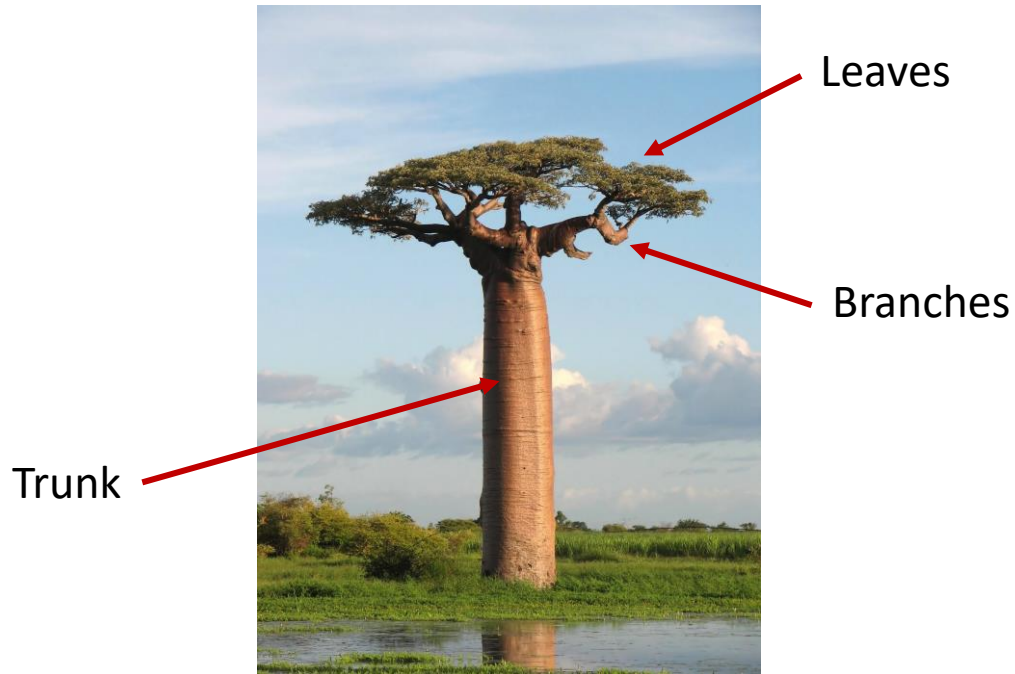
SD or SEM ?

- If the scatter is caused by **biological variability**, it is important to show the variation.
 - **Report the SD** rather than the SEM.
 - Better even: show a graph of all data points.
- If you are using an in vitro system with no biological variability, the scatter is about **experimental imprecision** (no biological meaning).
 - **Report the SEM** to show how well you have determined the mean.

Confidence interval

- Range of values that we can be 95% confident contains the true mean of the population.
 - Limits of 95% CI: **[Mean - 1.96 SEM; Mean + 1.96 SEM]** (SEM = SD/ \sqrt{N})

A distribution is not something made, it is something observed.

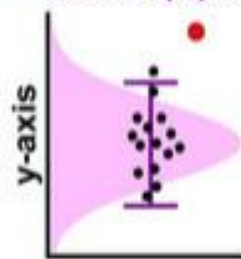


To recapitulate

- The Standard Deviation is **descriptive**
 - Just about the sample.
- The Standard Error and the Confidence Interval are **inferential**
 - Sample → General Population

Standard Deviation(SD) (Descriptive)

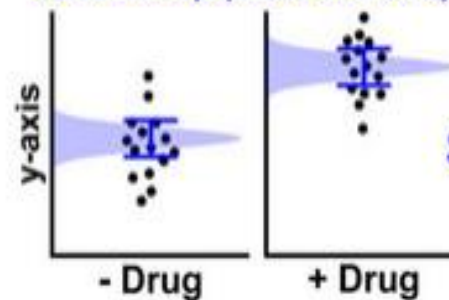
Q's w/n a population: *Is this "normal"?*



$$SD = \sqrt{\frac{\sum (y - \bar{y})^2}{(n-1)}}$$

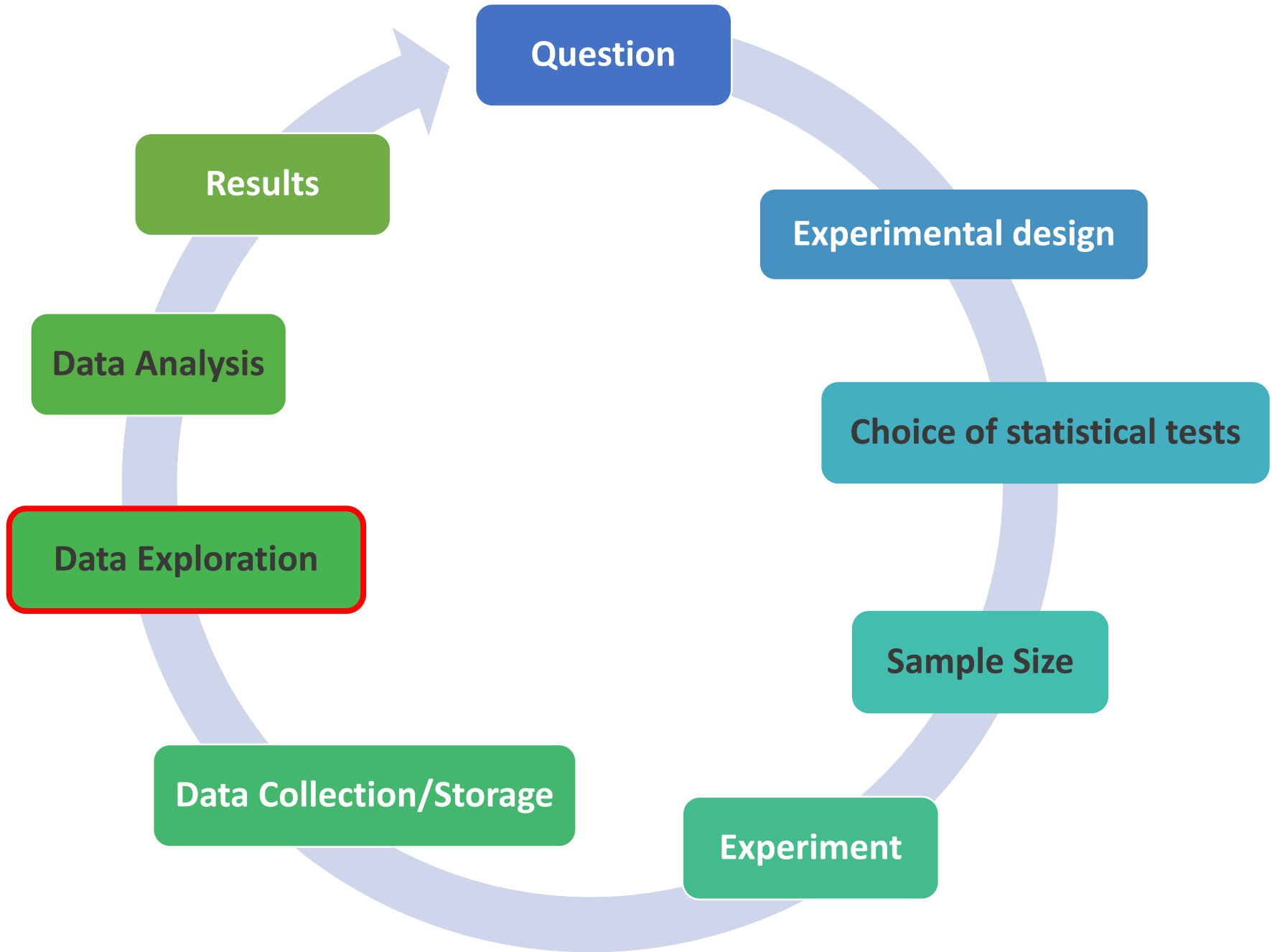
Standard Error(SE) (Inferential)

Q's between populations: *Are they "different"?*



$$SE = \frac{SD}{\sqrt{n}}$$

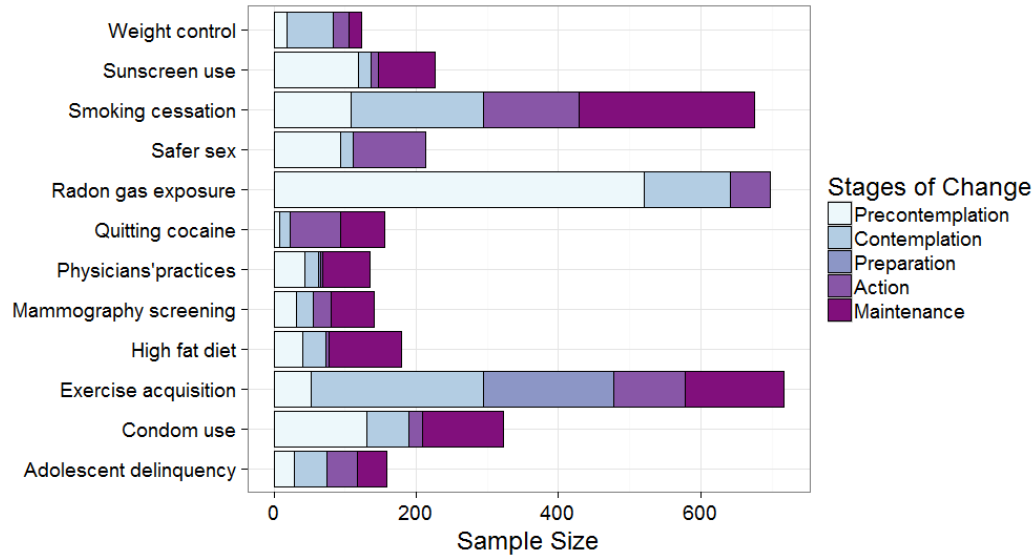
Graphical exploration of data



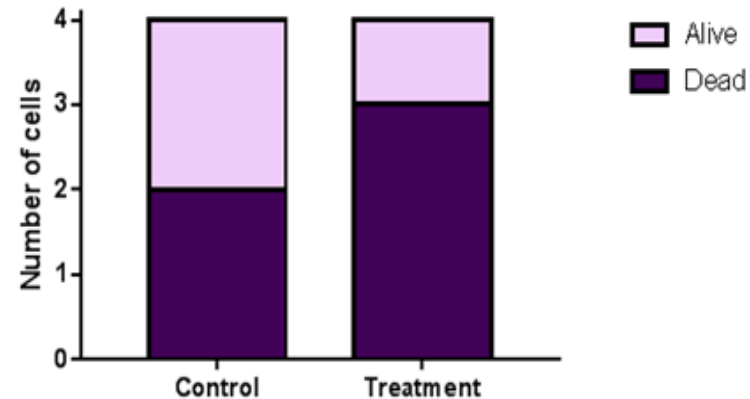
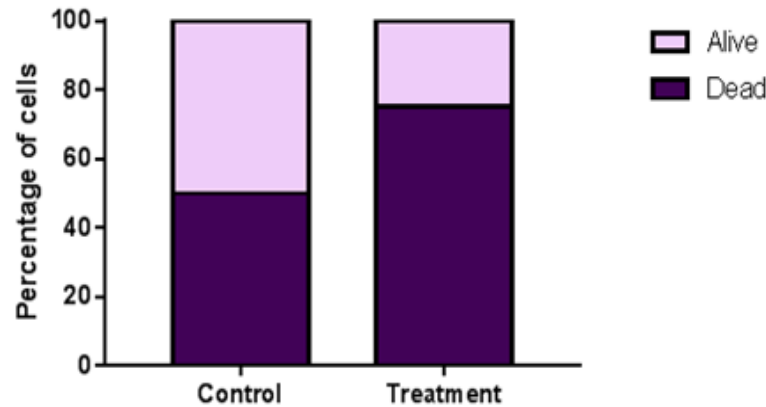
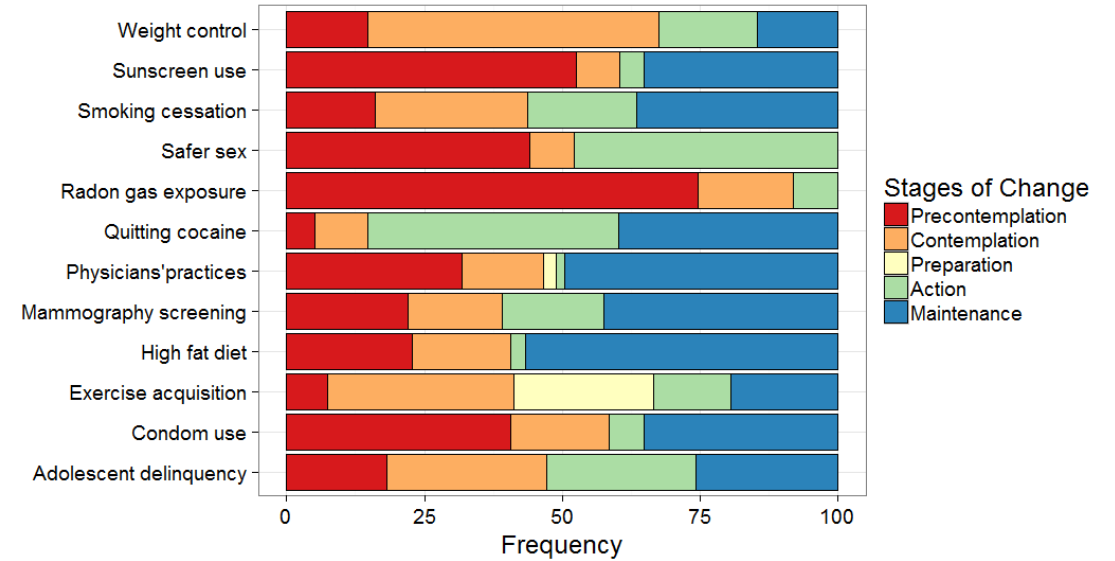
Data Exploration

Categorical data

Stages for Each of the 12 Problem Behaviours

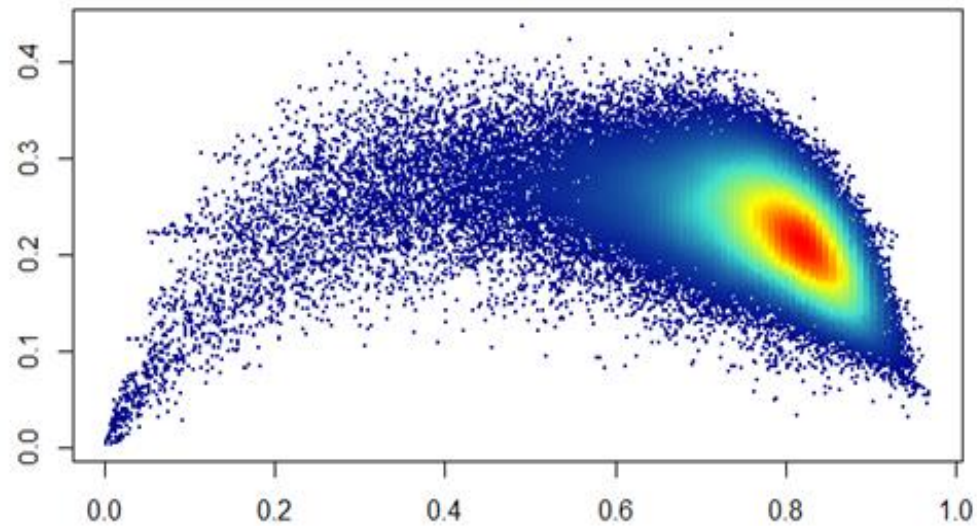
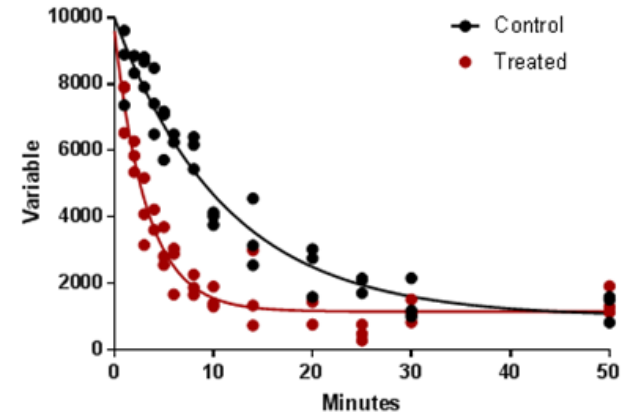
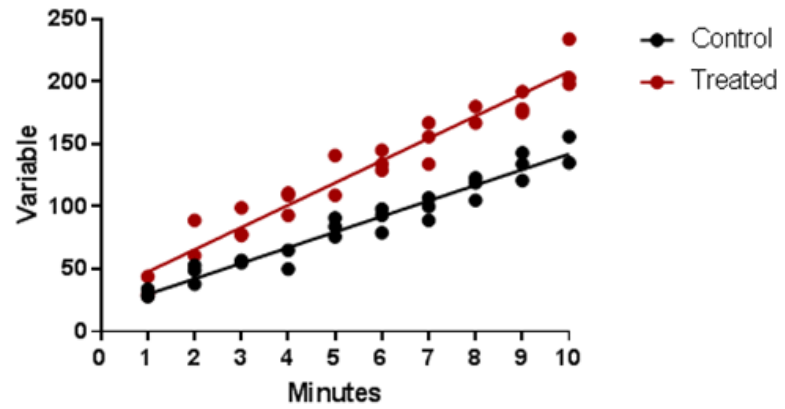


Stages for Each of the 12 Problem Behaviours



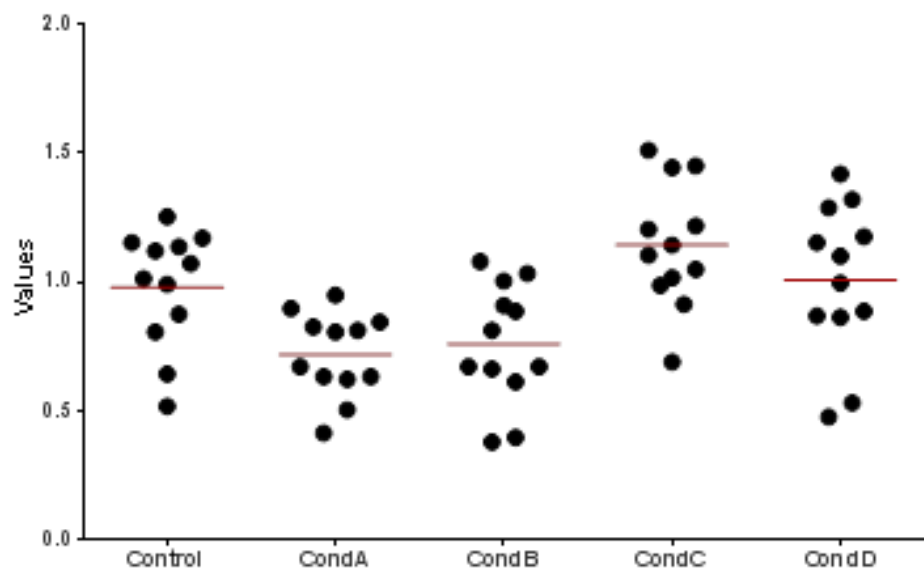
Data Exploration

Quantitative data: Scatterplot

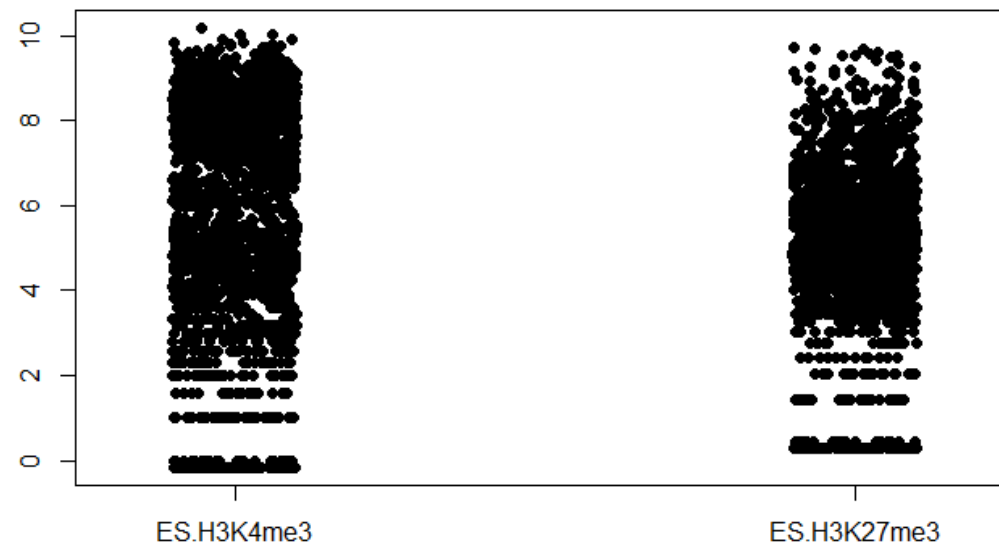


Data Exploration

Quantitative data: Scatterplot/stripchart



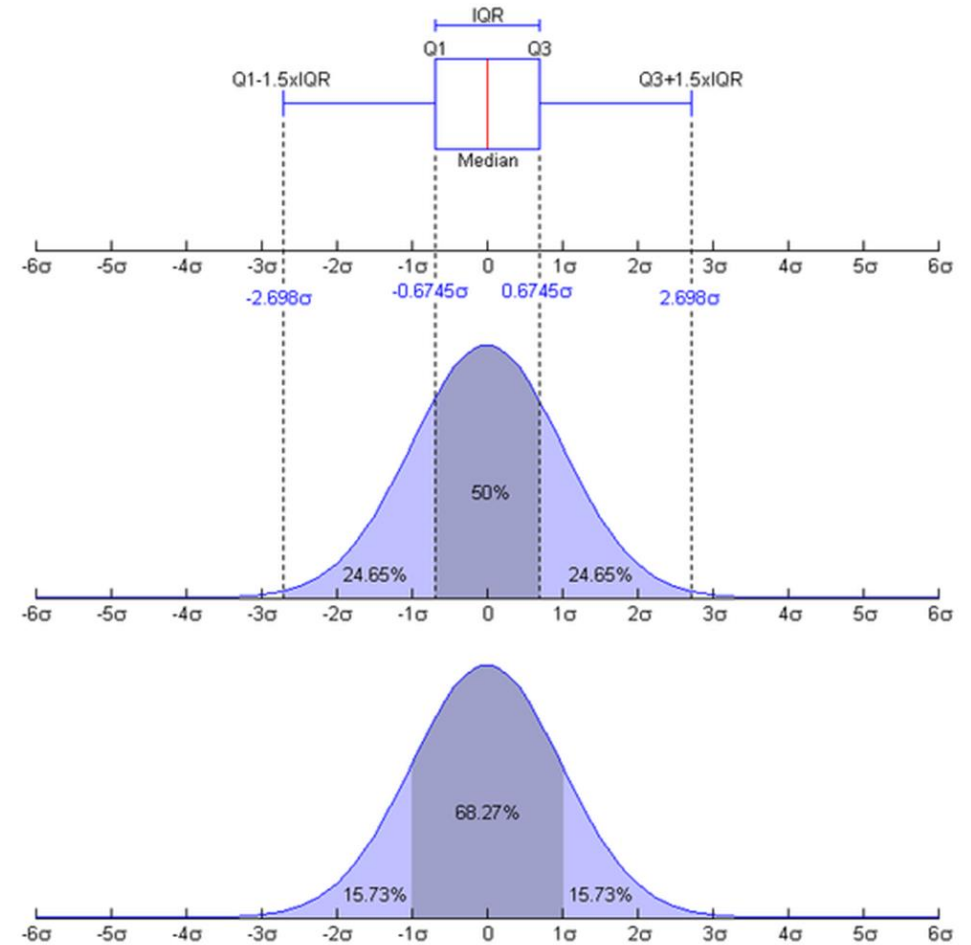
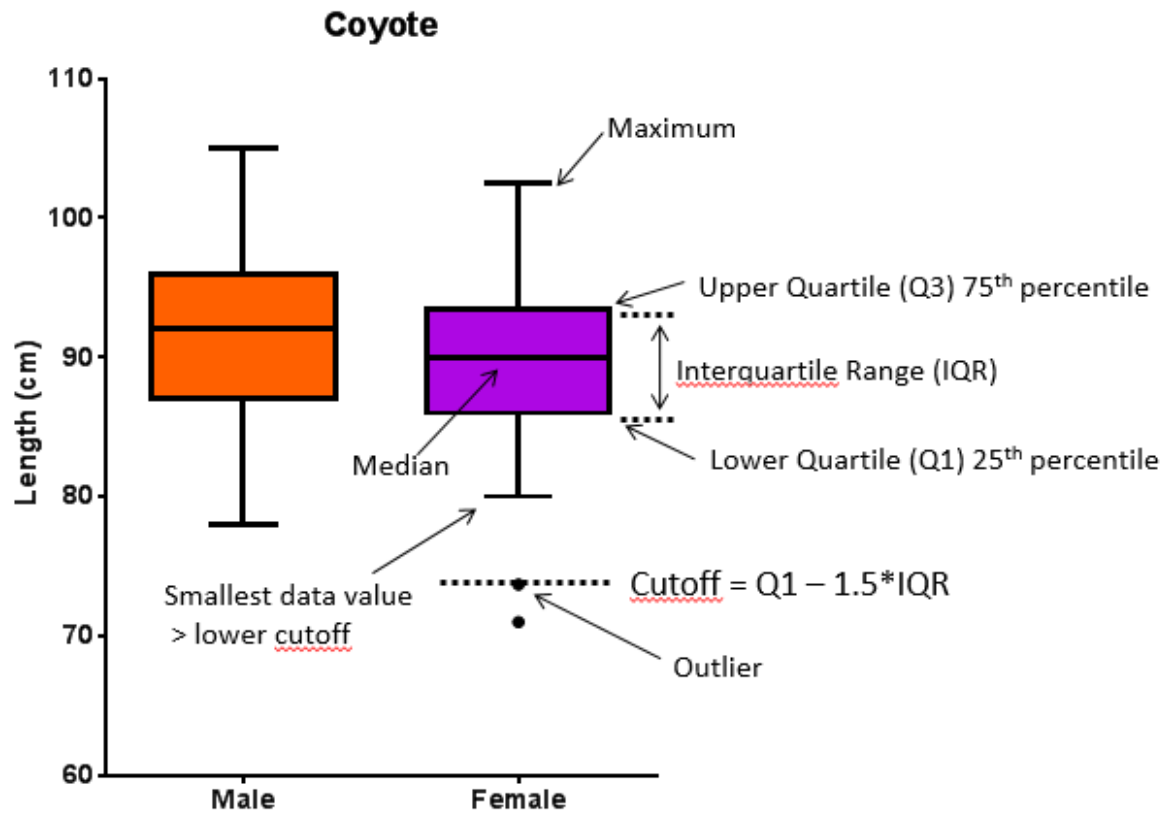
Small sample



Big sample

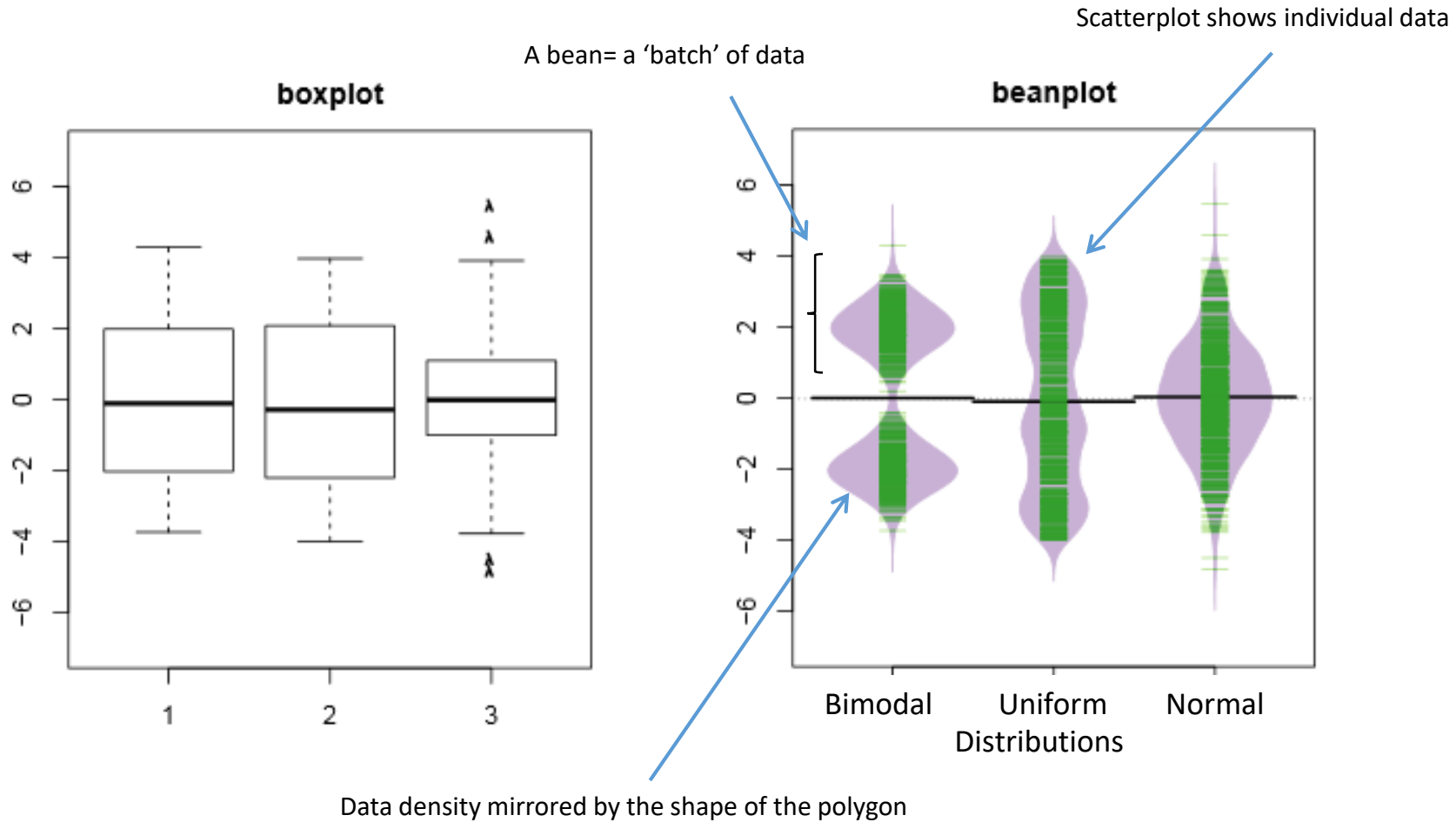
Data Exploration

Quantitative data: Boxplot



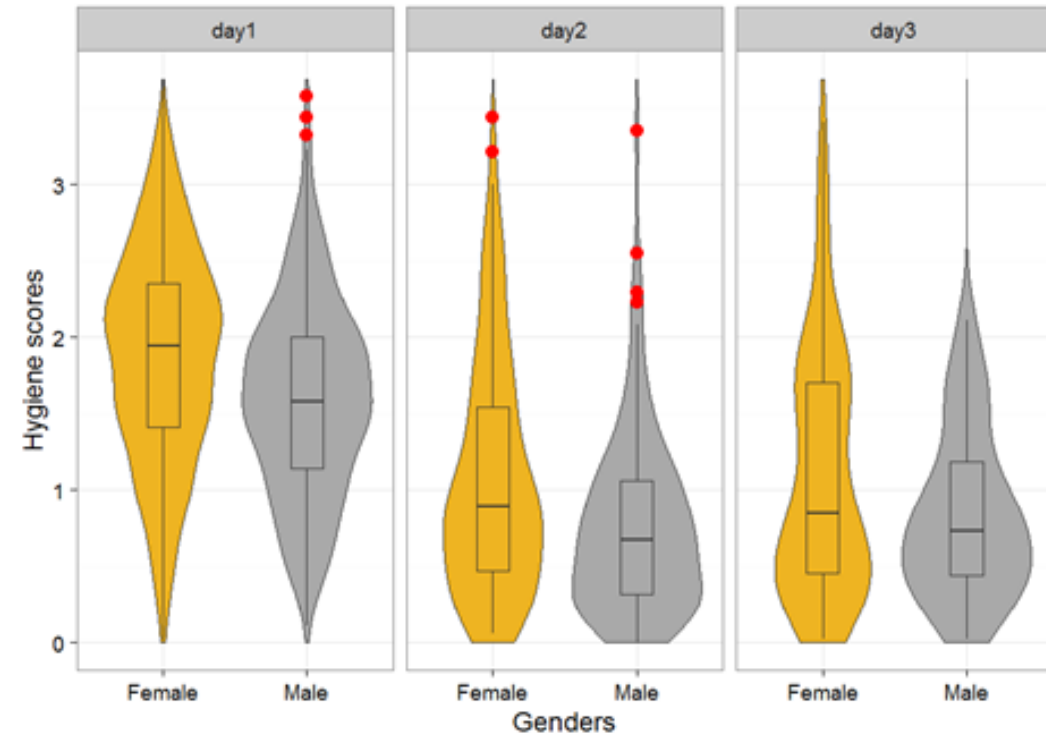
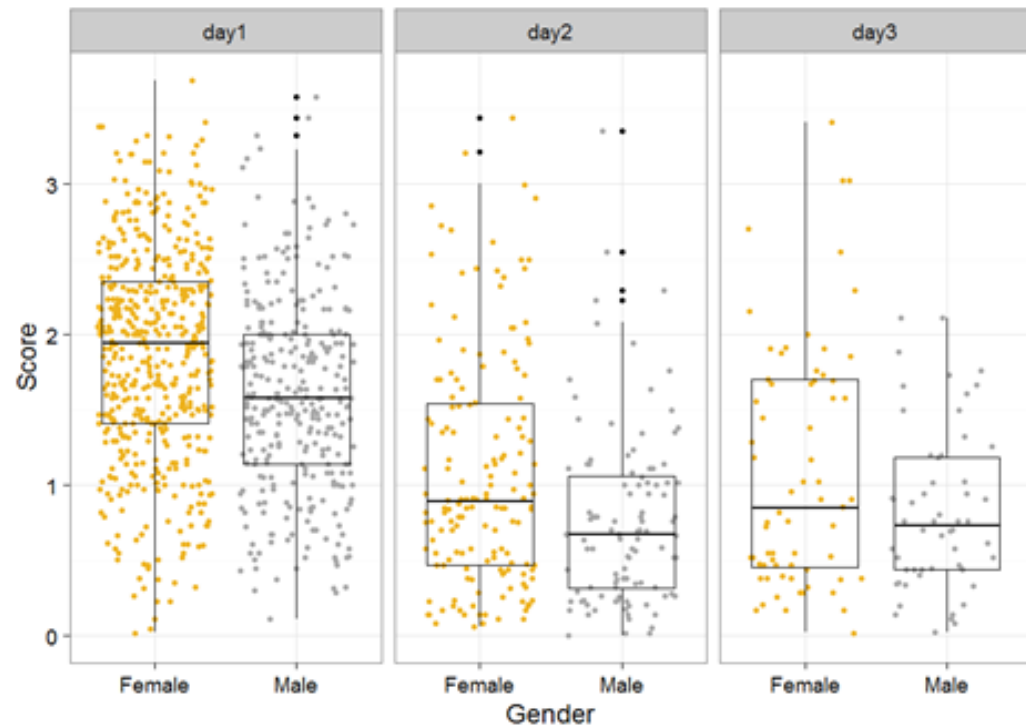
Data Exploration

Quantitative data: Boxplot or Beanplot



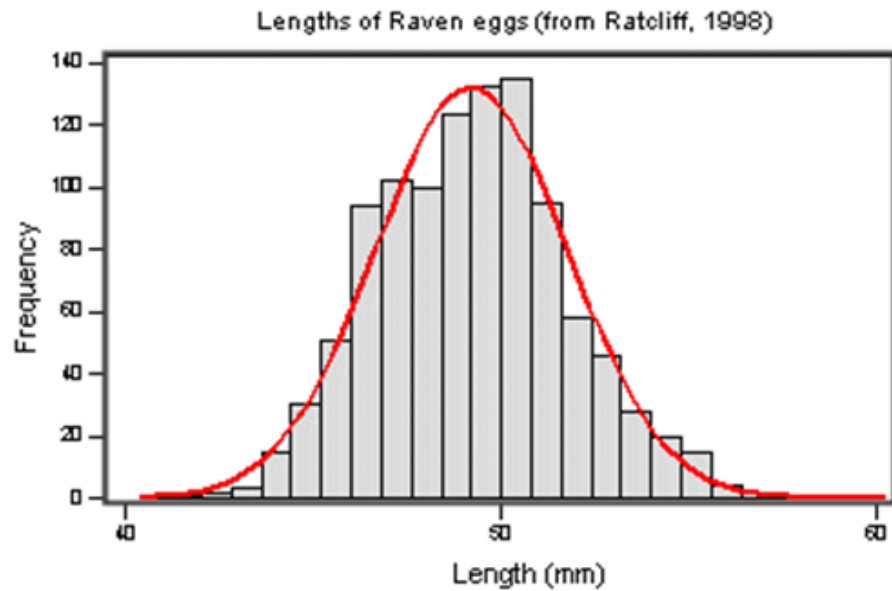
Data Exploration

Quantitative data: Boxplot and Beanplot and Scatterplot

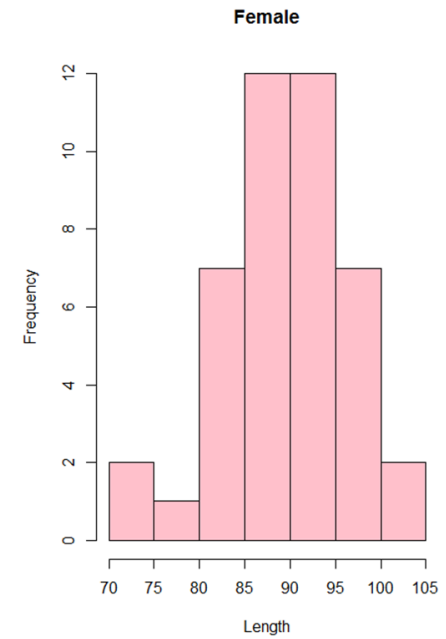
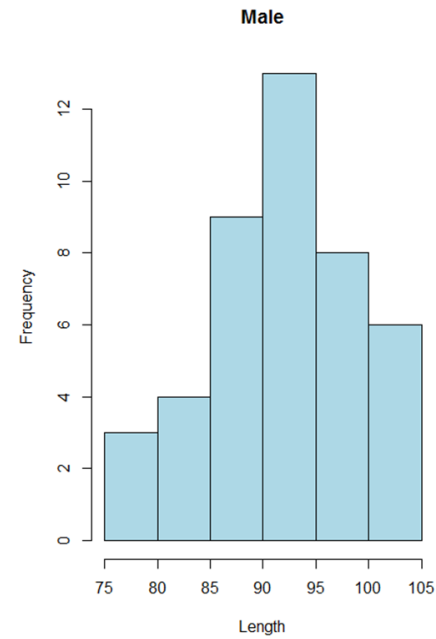


Data Exploration

Quantitative data: Histogram



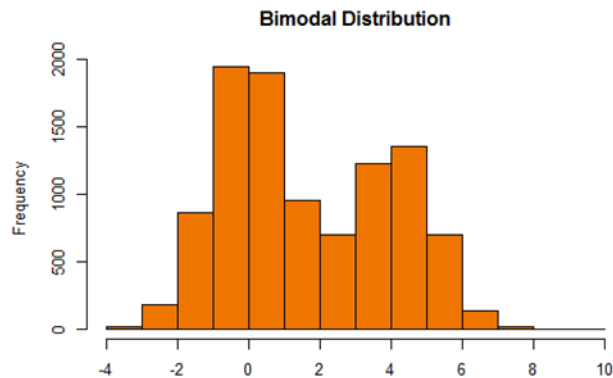
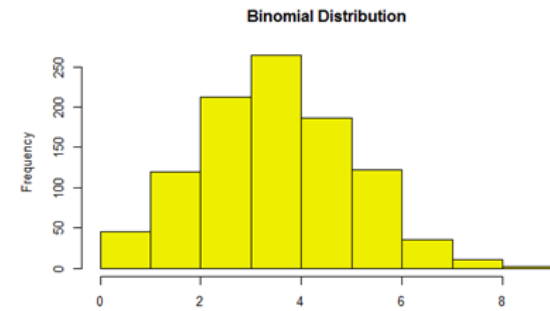
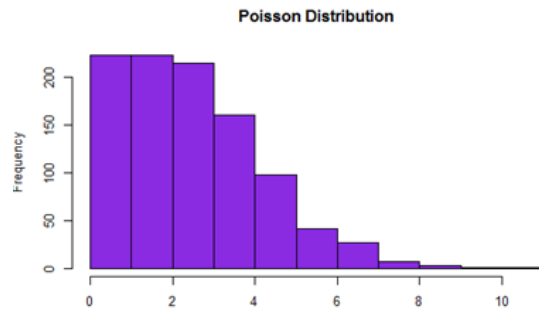
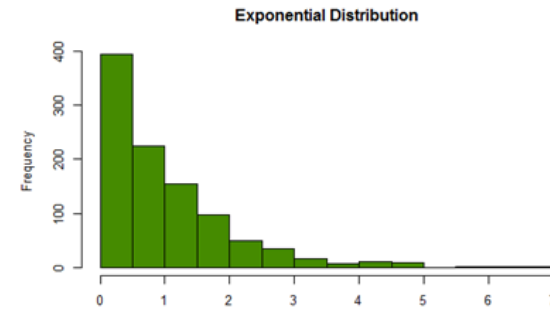
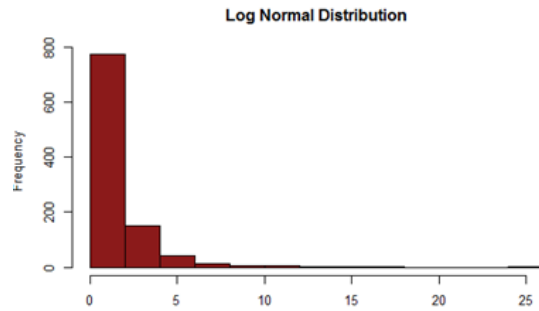
Big sample



Small sample

Data Exploration

Quantitative data: Histogram (distribution)



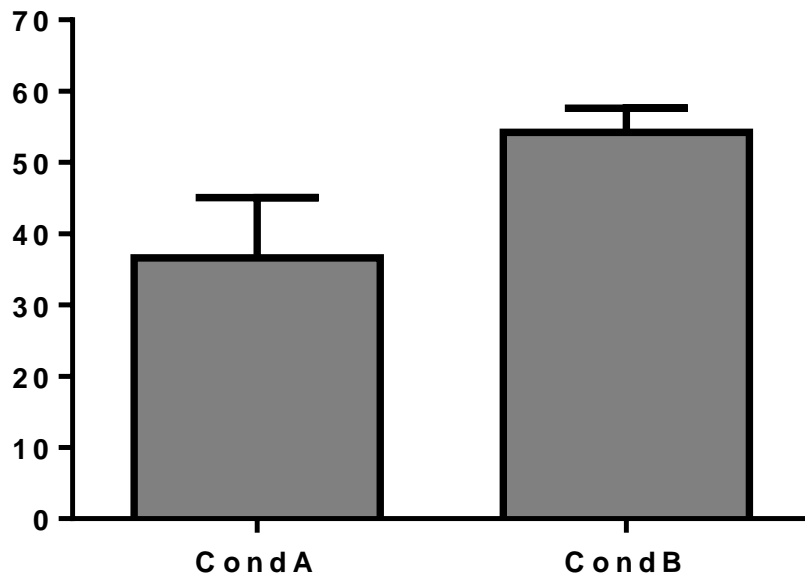
Data exploration \neq plotting data

Data Exploration

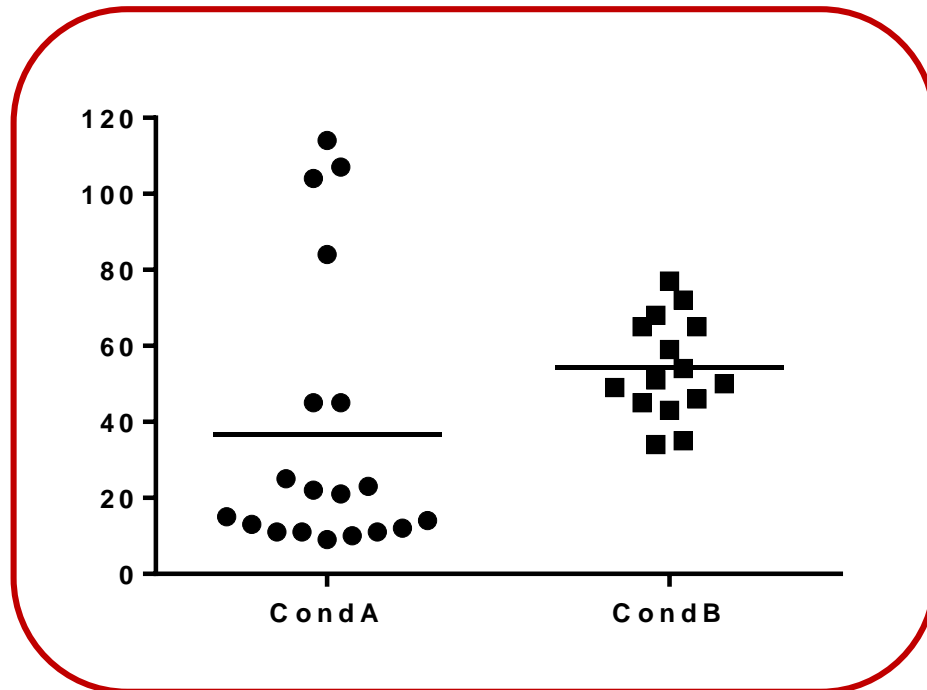
Plotting is not the same thing as exploring

- One experiment: change in the variable of interest between CondA to CondB.
 - ❖ Data plotted as a **bar chart**.

The fiction



The truth

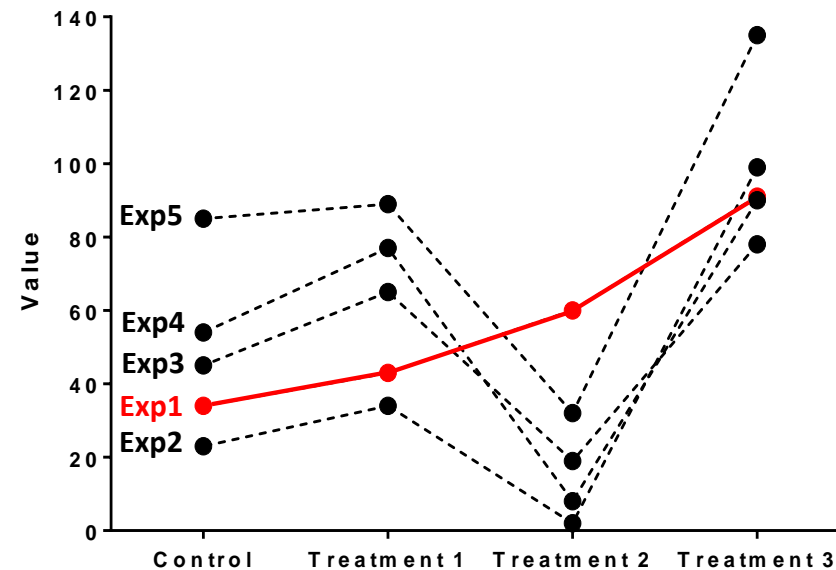
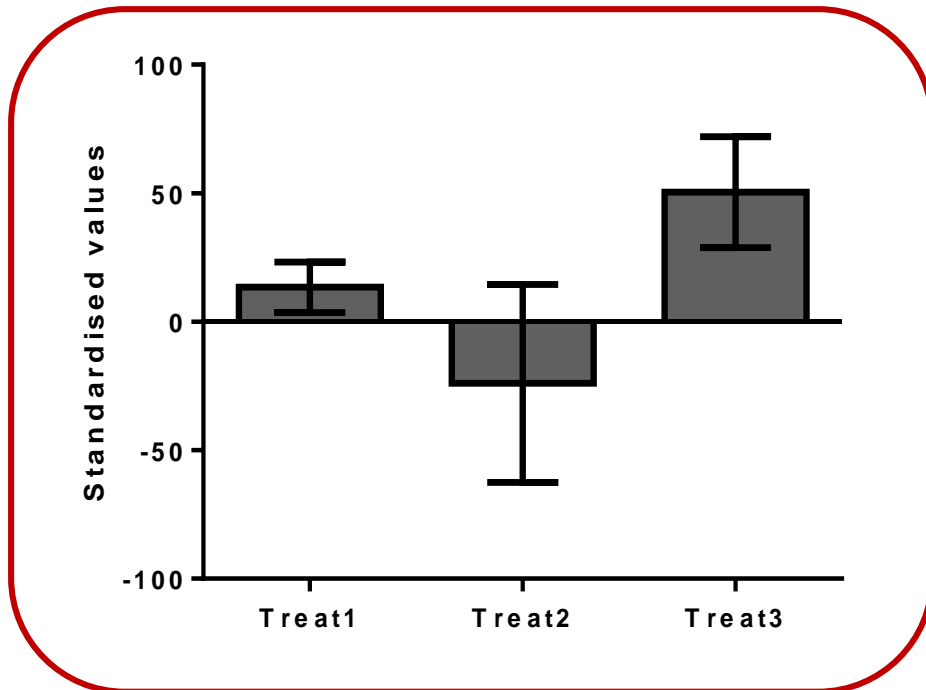


Data Exploration

Plotting (and summarising) is (so) not the same thing as exploring

- Five experiments: change in the variable of interest between 3 treatments and a control.
 - ❖ Data plotted as a **bar chart**.

The truth (if you are into bar charts)



Data Exploration

Plotting (and summarising and choosing the wrong graph) is (definitely) not the same thing as exploring

- Four experiments: Before-After treatment effect on a variable of interest.
- Hypothesis: Applying a treatment will decrease the levels of the variable of interest.
- ❖ Data plotted as a **bar chart**.

