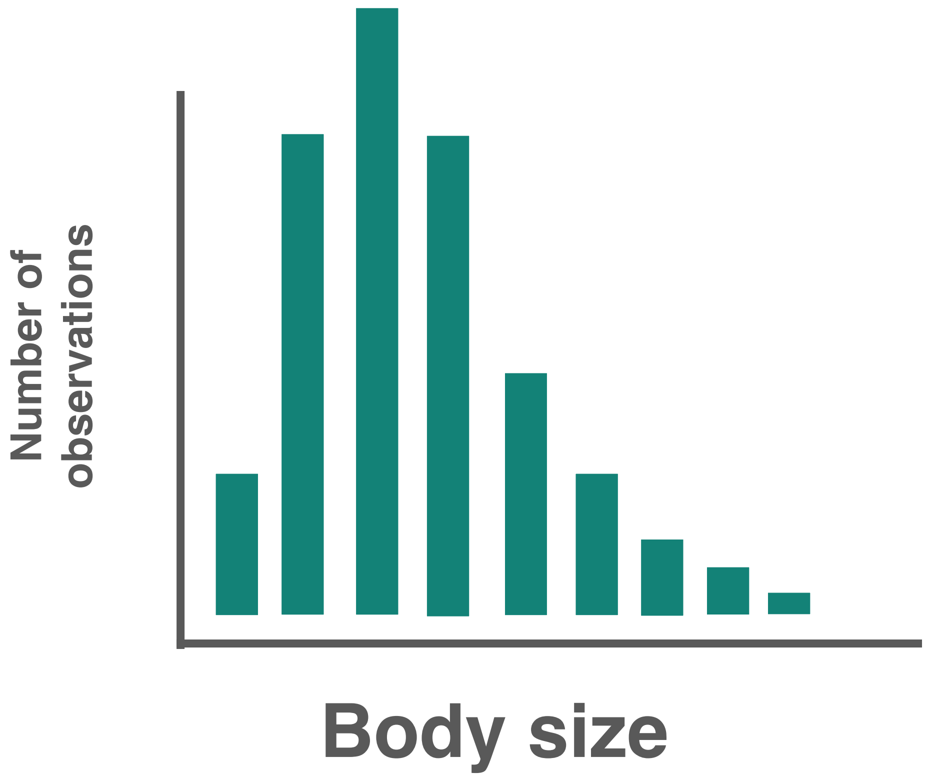
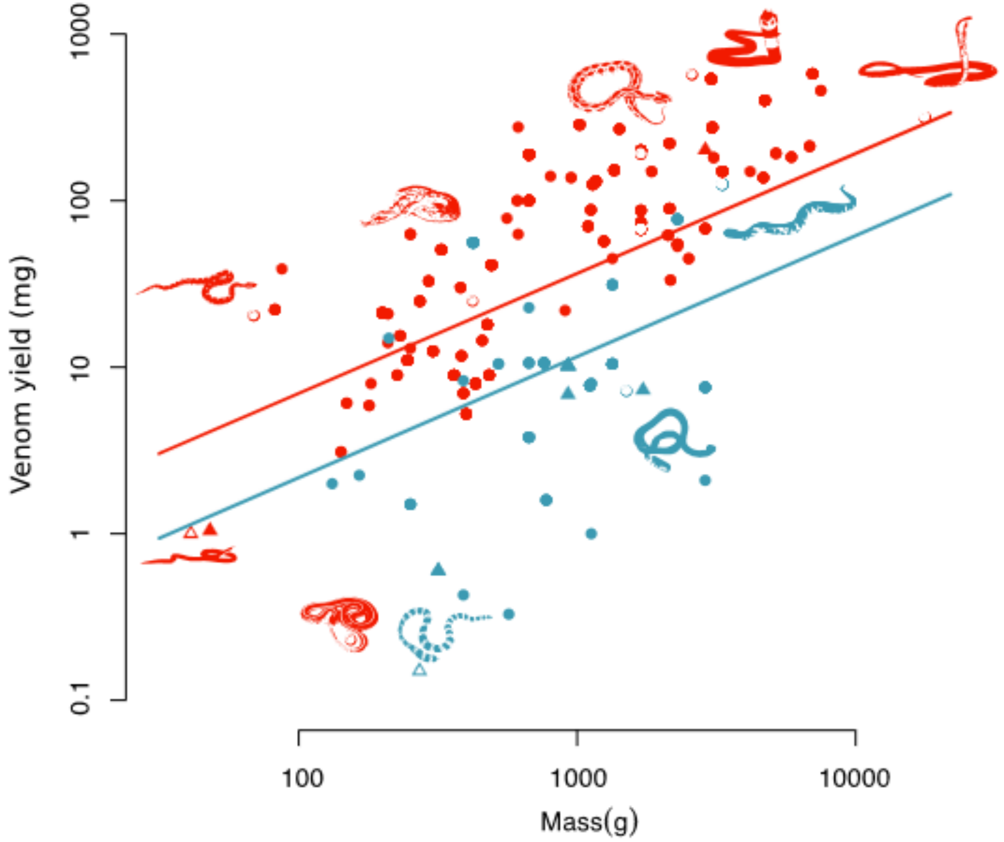
Biostatistics

Lecture 3: Data types and Figures





- Different types of data
- Summary statistics
- Histograms and distributions
- Plot types

Types of data

Qualitative

Data is in categories

Nominal

Categories with no order e.g. Hair color (red, black)

Ordinal

Categories that can have an order e.g. income (low, medium high)

Quantitative

Data can be expressed in numbers

Discrete

Count data that are whole numbers e.g. number of species (1,100, etc)

Continuous

Integers, numbers that can be fractions e.g. body mass (10, 2.5, 4.44)

Types of data

Qualitative

Data is in categories

The data can only be expressed in groups or categories

Quantitative

Data can be expressed in numbers

The data can be expressed in numbers

Types of data

Qualitative

Data is in categories

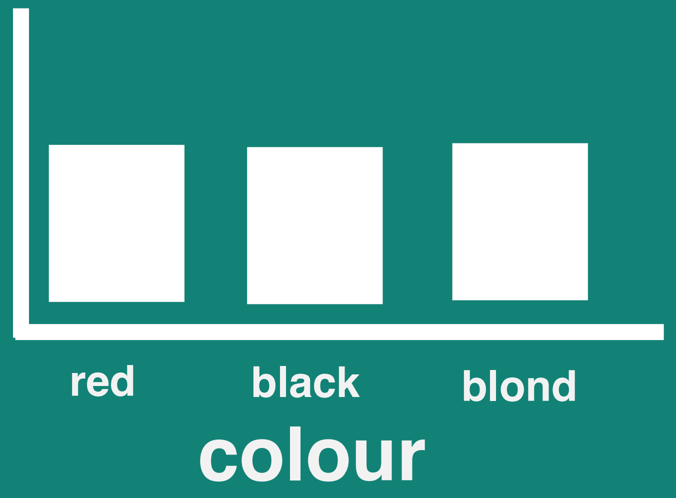
Nominal

Categories with no order
e.g. Hair color (red, black)

Nominal

Groups which cannot be ordered in a way that signifies different levels of value

For example, we cannot say that different hair colors are greater or less than others



Types of data

Qualitative

Data is in categories

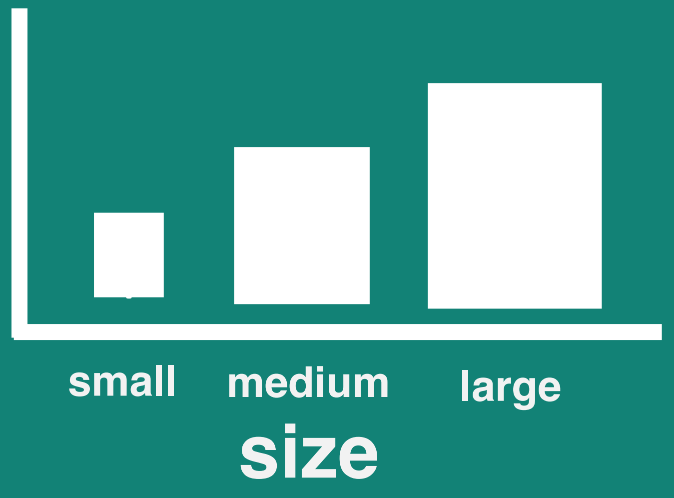
Ordinal

Categories that can have an order e.g. size (low, medium high)

Ordinal

Groups which can be ordered in way that signifies different levels of value

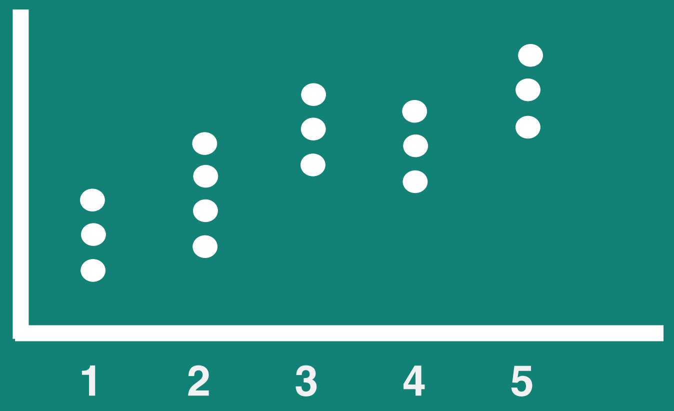
For example, while not quantified, size categories of small, medium, large indicates that small < medium < large



Types of data

Discrete

Quantitative data that can only be expressed in whole numbers, for example, count data of number of species



Number of species

Quantitative

Data can be expressed in numbers

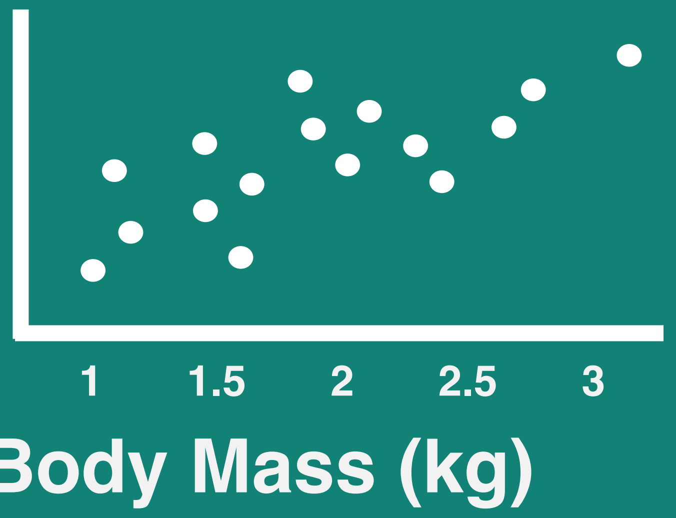
Discrete

Count data that are whole numbers e.g. number of people (1,100, etc)

Types of data

Discrete

Quantitative data that can be expressed as an integer (number with fractional values), for example, body mass



Quantitative

Data can be expressed in numbers

Continuous

Integers, numbers that can be fractions e.g. body mass (10, 2.5, 4.44)

Data distributions

What does my data look like?

How are my observations spread across the groups

They may be evenly spread across each group

Data distributions

What does my data look like?

How are my observations spread across the groups

Or distributed unevenly across groups

Data distributions

What does my data look like?

How are the observations for the different species distributed for the iris dataset.

Use the below code in R to check

```
iris_data <- iris

plot(iris_data$Species)
```

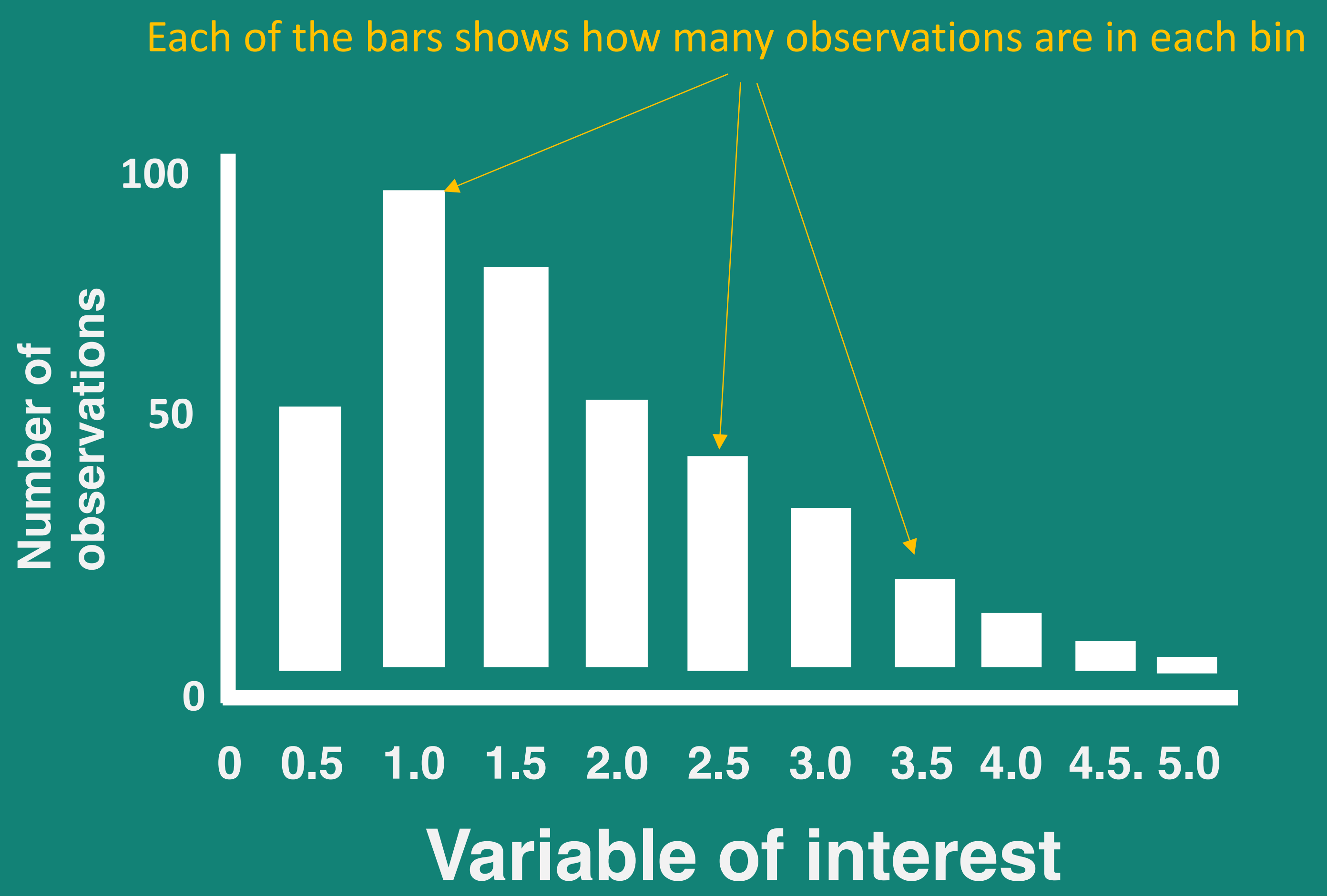
Histograms

What does my data look like?

For continuous data we can ask the same question using histograms

Histograms put continuous data into bins and plots them.

For example, we are plotting the number of observation between 0 and 0.5 in the first bin, than the number of observations between 0.5 and 1 in the second bin etc.

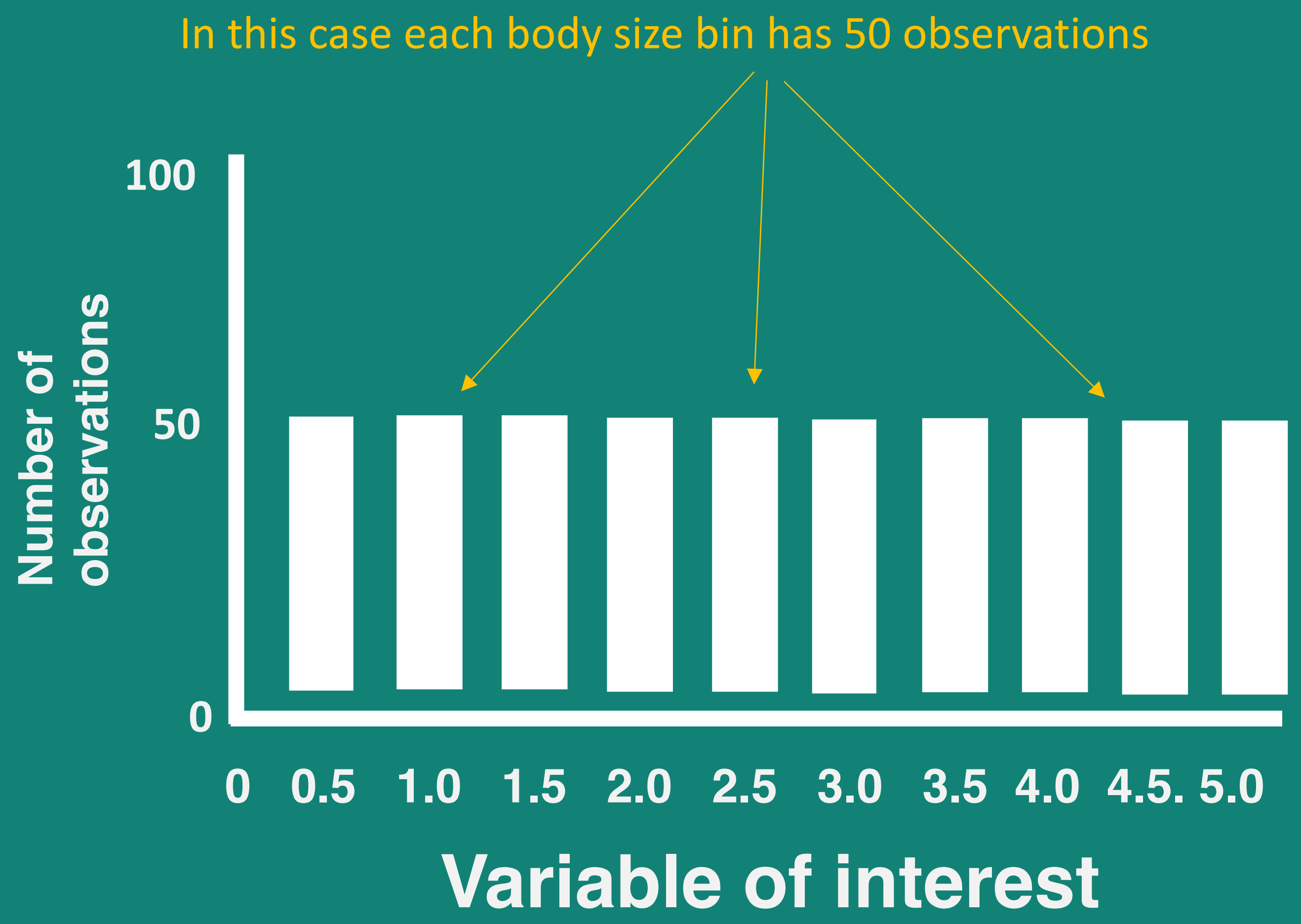


Histograms

Uniform

Data can have a uniform distribution

i.e. any value is equally likely to be observed across the range of the data.



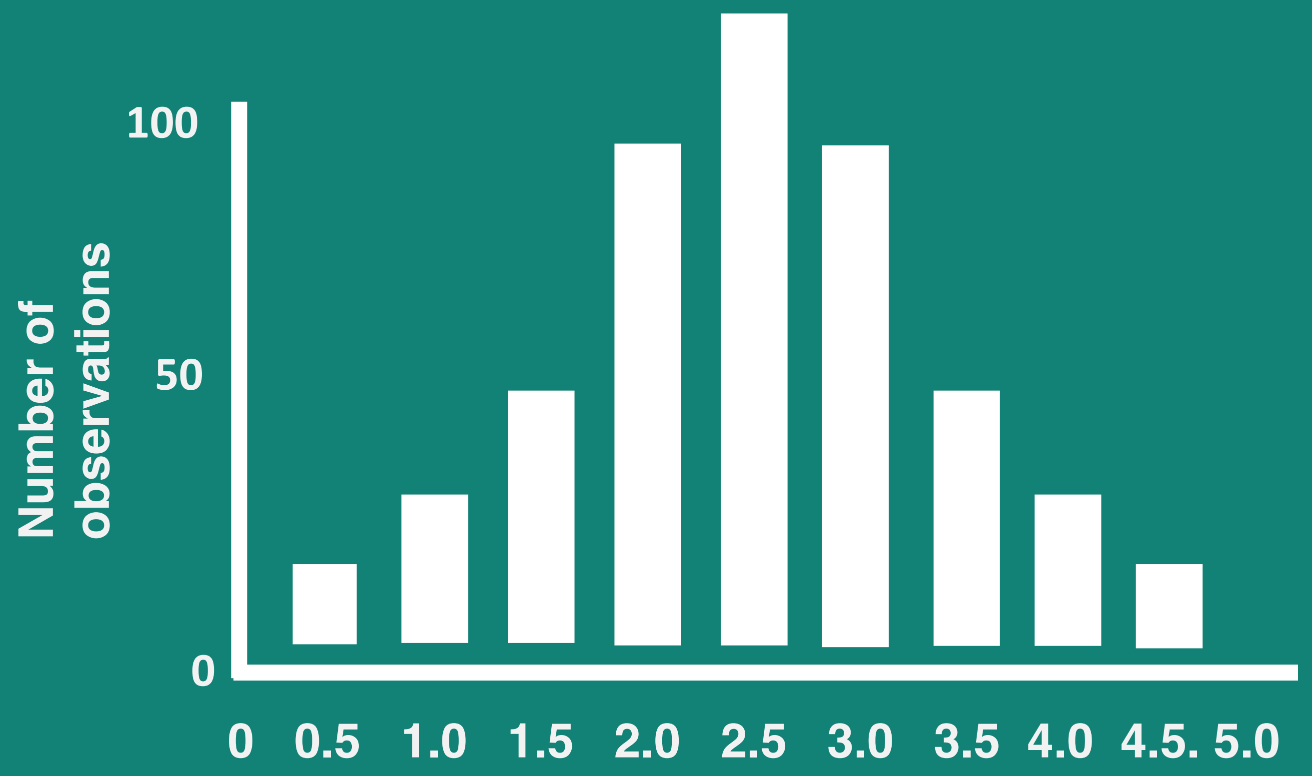
Histograms

Normal

Normal distribution

One of the common type of distribution in biological data

This is where there is a central common value with less frequent values spread evenly on both sides of the most common value.



Histograms

What does my data look like?

What is the distribution of Sepal Width in our iris dataset

Use the below code in R to check

```
iris_data <- iris

hist(iris_data$Sepal.Width)
```

Summary statistics

Using numbers to explain what my data looks like
e.g. averages, ranges, etc.

Central Tendency

Mean

The arithmetic mean is the average value, Calculated as the sum of values divided by the number of values.

The mean ($\bar{x}$) of some sequence of numbers *x*, which has *n* entries is

$$\bar{x} = \frac{\sum x}{n}$$

Central Tendency

Mean

The arithmetic mean is the average value, Calculated as the sum of values divided by the number of values.

The mean ($\bar{x}$) of the numbers

c(1, 2, 2, 3, 4, 7, 9)

$$\bar{x} = \frac{1+2+2+3+4+7+9}{7}$$

$$\bar{x} = 4$$

In R we can use the `mean()` function

`mean(c(1, 2, 2, 3, 4, 7, 9))`

Central Tendency

Median

The middle value. The value which separates the lower and upper half of the values.

To get the median of the numbers

c(20, 10, 5, 30, 5)

we reorder according to rank

c(5, 5, **10**, 20, 30)

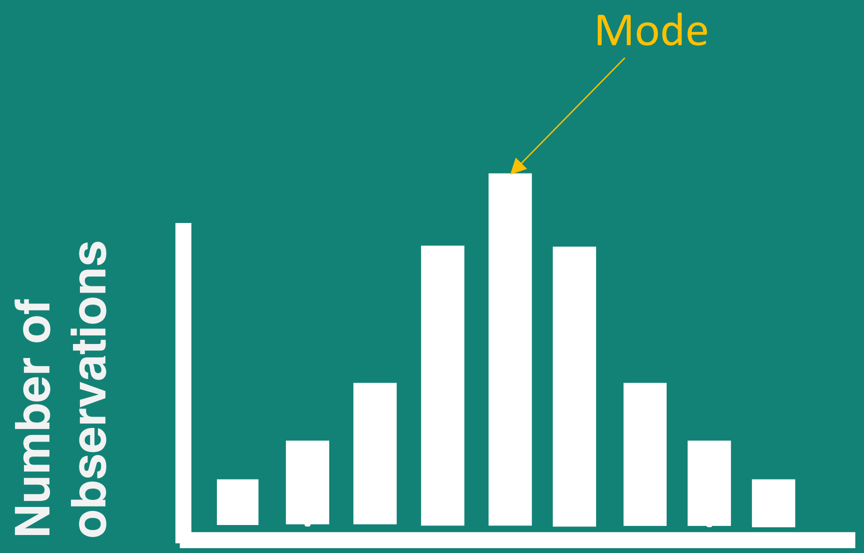
10

In R we can use the `median()` function

`median( c(1, 2, 2, 3, 4, 7, 9) )`

Central Tendency

The mode represents the peak of a distribution



Mode

The most frequent value observed.

Central Tendency

In R we can use the `hdr()` function

```
library("hdrcde")

hdr(c(1, 2, 2, 3, 4, 7, 9))
```

Mode

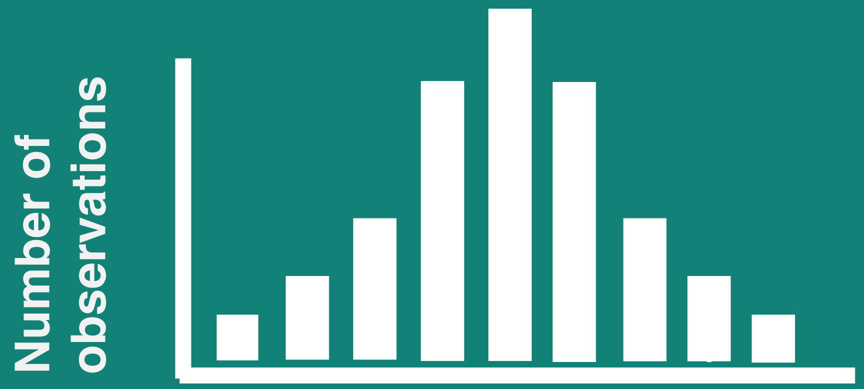
The most frequent value observed.

Measures of variance

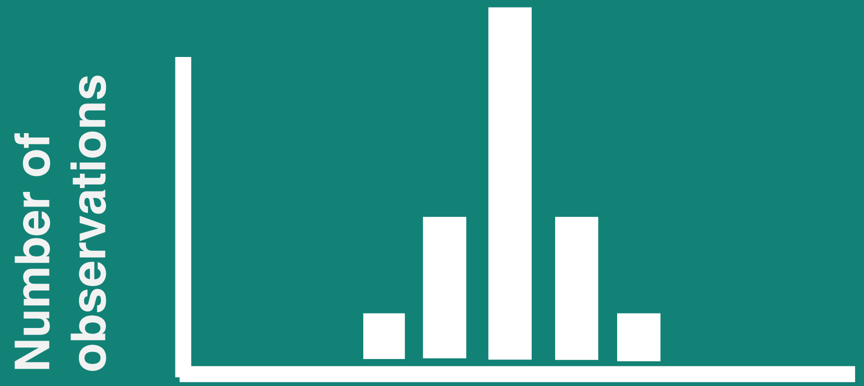
Spread of the data

How much does the data spread from some measure of central tendency.

High spread of data



Low spread of data



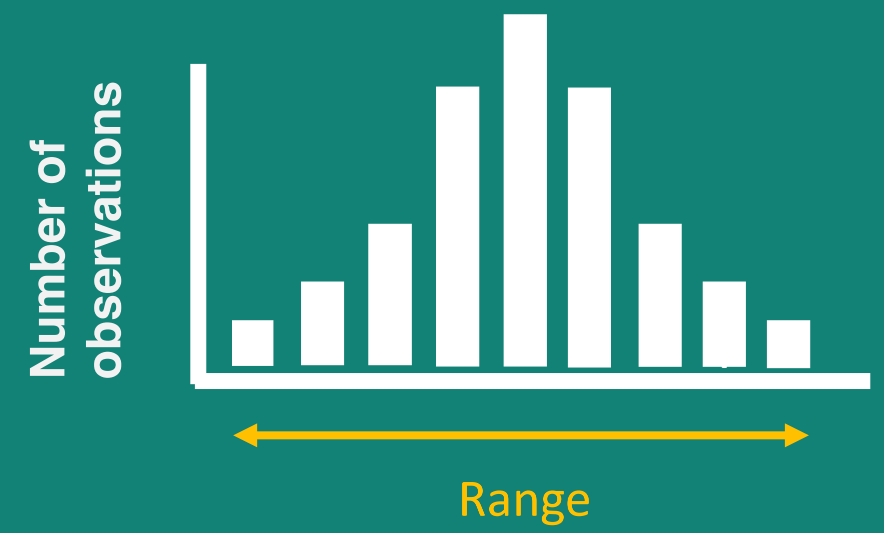
Measures of variance

Spread of the data

How much does the data spread from some measure of central tendency.

The simplest and most basic measure of this is the range of the values. This is the minimum and maximum values.

```
range(iris_data$Sepal.Width)
```



Measures of variance

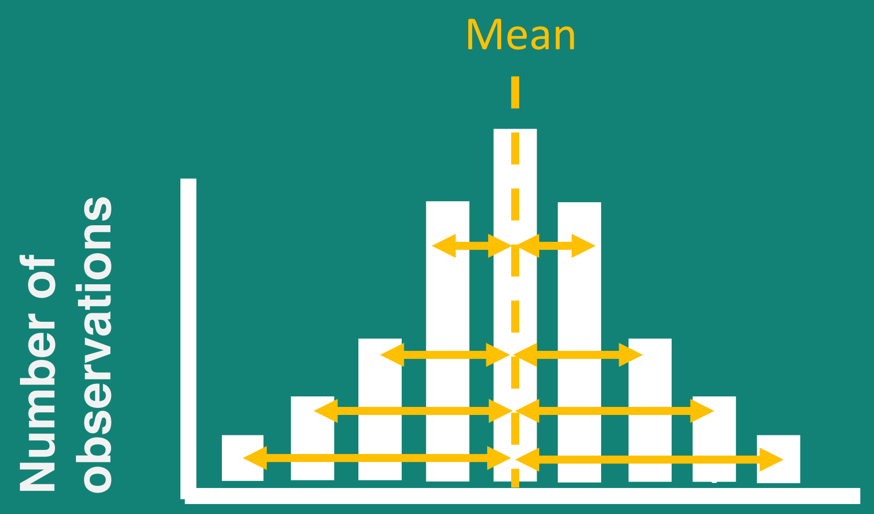
Variance

A measure of how far the data spread.

Calculated by getting the distance of each datapoint from the mean

$$S^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

S^2 = variance
$\bar{x}$ = mean
N = sample size



Measures of variance

Variance

A measure of how far the data spread.

Calculated by getting the distance of each datapoint from the mean

$$S^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

S^2 = variance
$\bar{x}$ = mean
N = sample size

The variance of our list of numbers can be calculated as

c(1, 2, 2, 3, 4, 7, 9)

$$S^2 = \frac{(1-4)^2 + (2-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (7-4)^2 + (9-4)^2}{7-1}$$

$$S^2 = \frac{9 + 4 + 4 + 1 + 0 + 9 + 25}{6} = \frac{52}{6}$$

S^2 = 8.67

Measures of variance

Variance

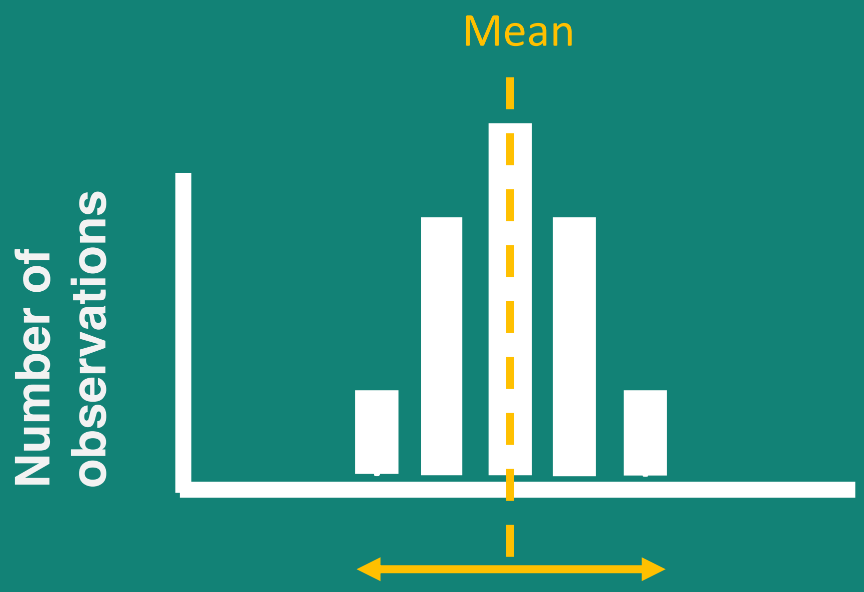
A measure of how far the data spread.

Calculated by getting the distance of each datapoint from the mean

$$S^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

S^2 = variance
$\bar{x}$ = mean
N = sample size

Data with low variance will have a tight distribution



Measures of variance

Variance

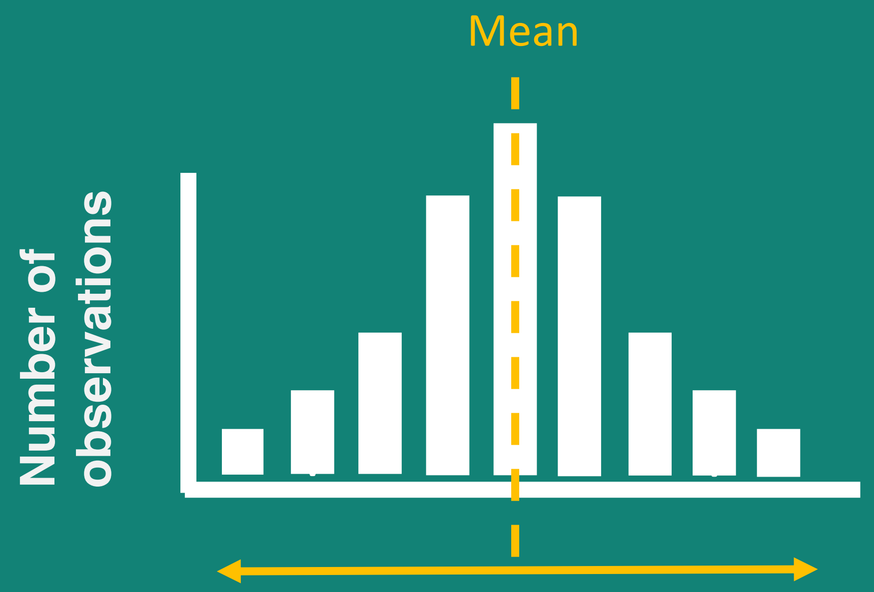
A measure of how far the data spread.

Calculated by getting the distance of each datapoint from the mean

$$S^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

S^2 = variance
$\bar{x}$ = mean
N = sample size

Data with high variance will have a wider distribution



Measures of variance

Variance

A measure of how far the data spread.

Calculated by getting the distance of each datapoint from the mean

$$S^2 = \frac{\sum(X - \bar{x})^2}{n - 1}$$

S^2 = variance
$\bar{x}$ = mean
N = sample size

In R we can use the `var()` function

```
var(c(1, 2, 2, 3, 4, 7, 9) )
```

Measures of variance

Standard deviation

A measure of how far the data spread.

$$\sigma = \sqrt{S^2}$$

$$\sigma = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$$

σ = standard deviation

$\bar{x}$ = mean

N = sample size

In R we can use the `sd()` function

```
sd(c(1, 2, 2, 3, 4, 7, 9) )
```

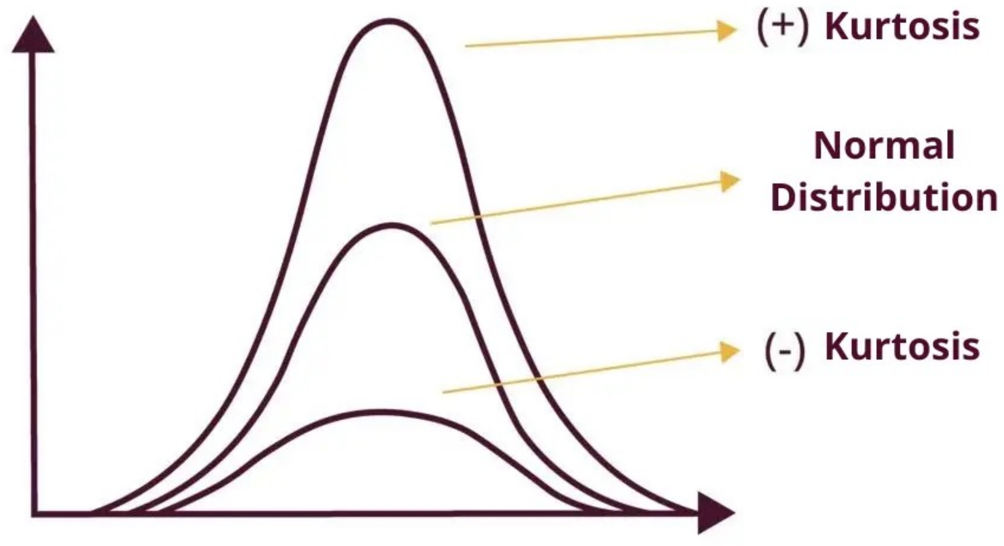
Measures of variance

Kurtosis

A measure of how flat the distribution
Is relative to a normal distribution.

If the data is very tight to the mean it has positive Kurtosis,

if the data is flat is has a negative Kurtosis



Measures of variance

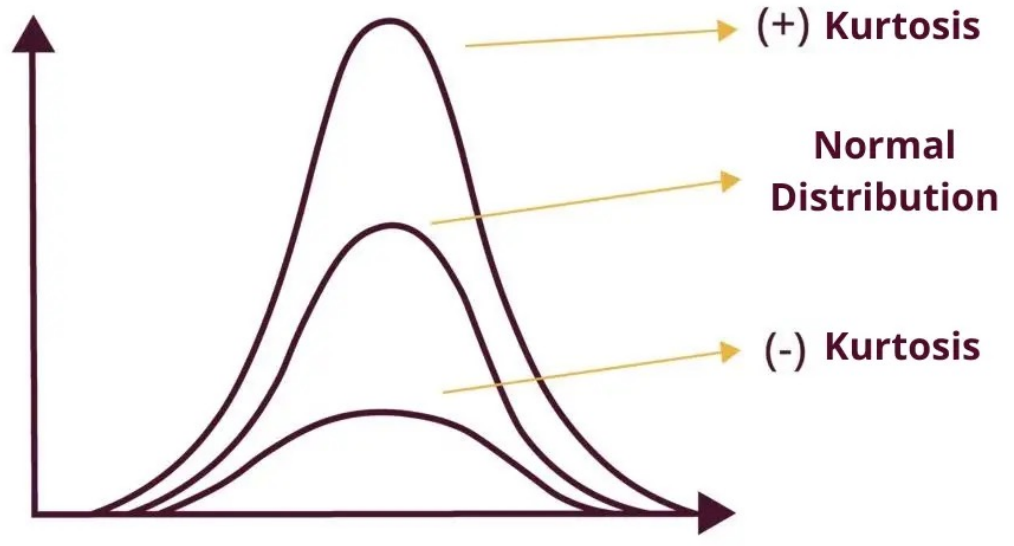
Kurtosis

This is rarely used as a summary statistic but you can calculate in R using the moments package

```
install.packages("moments")

library("moments")

kurtosis(iris_data$Sepal.Width)
```



Summary statistics

Summary statistics of Sepal.Width

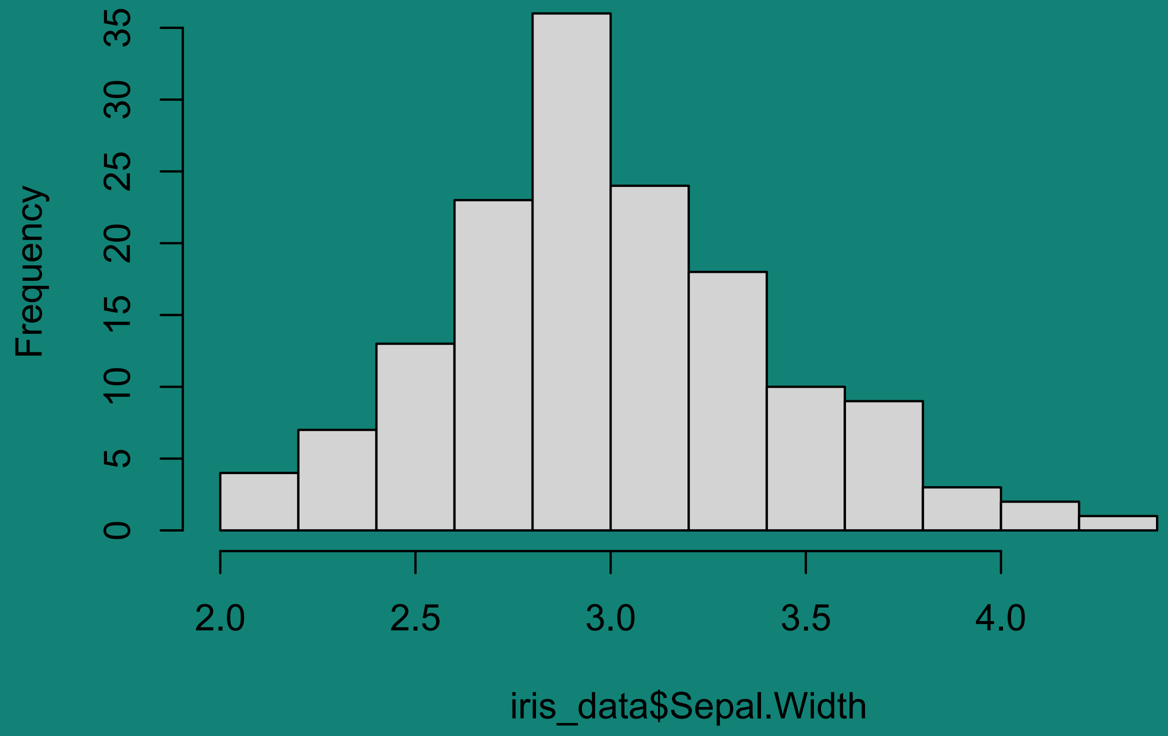
Using R calculate the mean, median, mode, range, variance and standard deviation of `Sepal.Width`

R Code examples

```
iris_data <- iris

hdr(iris_data$Sepal.Width)

var(iris_data$Sepal.Width)
```

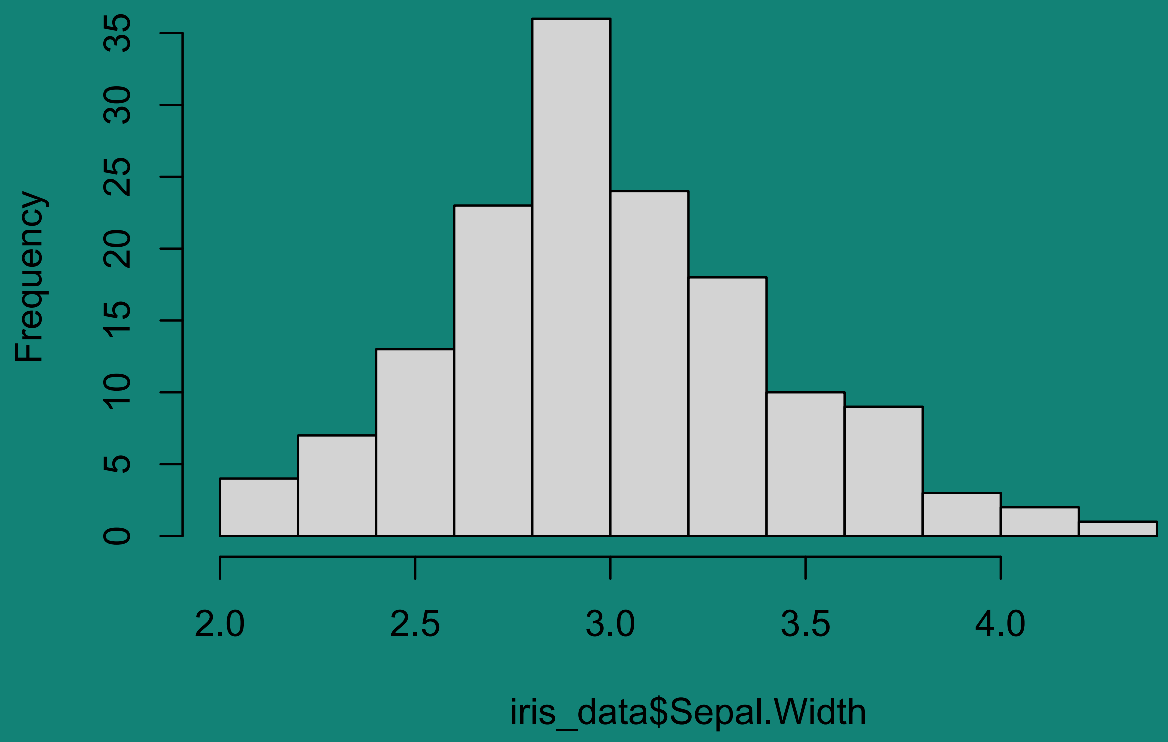


Summary statistics

Summary statistics of Sepal.Width

Because the distribution of Sepal Width is close to normal the Mode, mean and median will be very similar.

Mode, mean and median all very similar

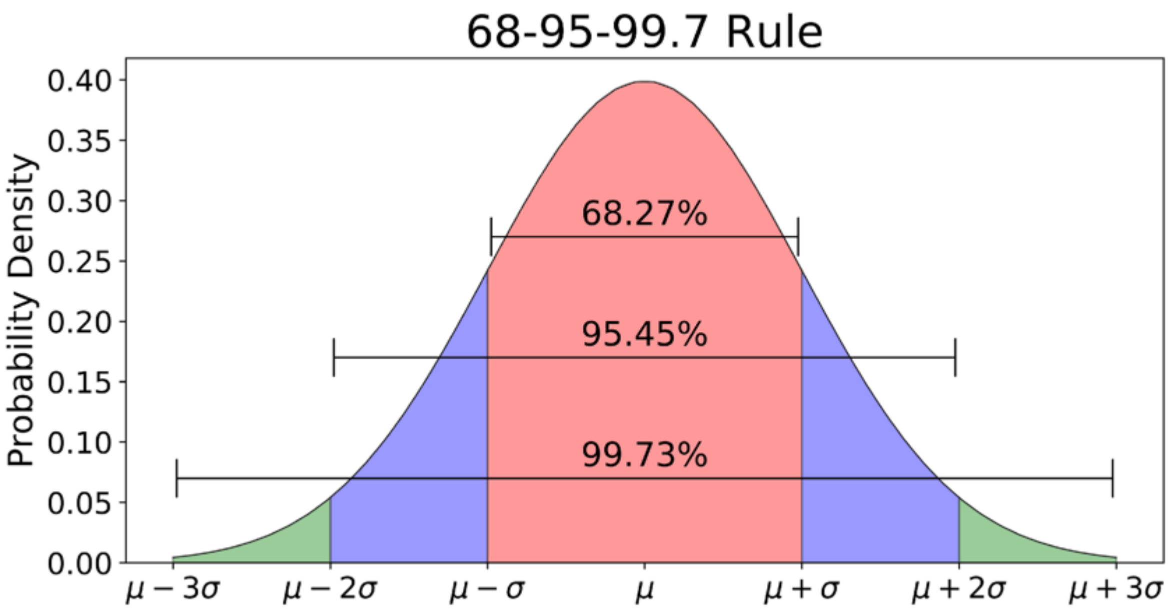


Data distributions

Normal distribution

Sepal Width is an example of something close to a normal distribution.

In a normal distribution
68.27% of the values are between the 1 standard deviation,
95.45% between 2 standard deviations and
99.7% between 3 values of standard deviation.



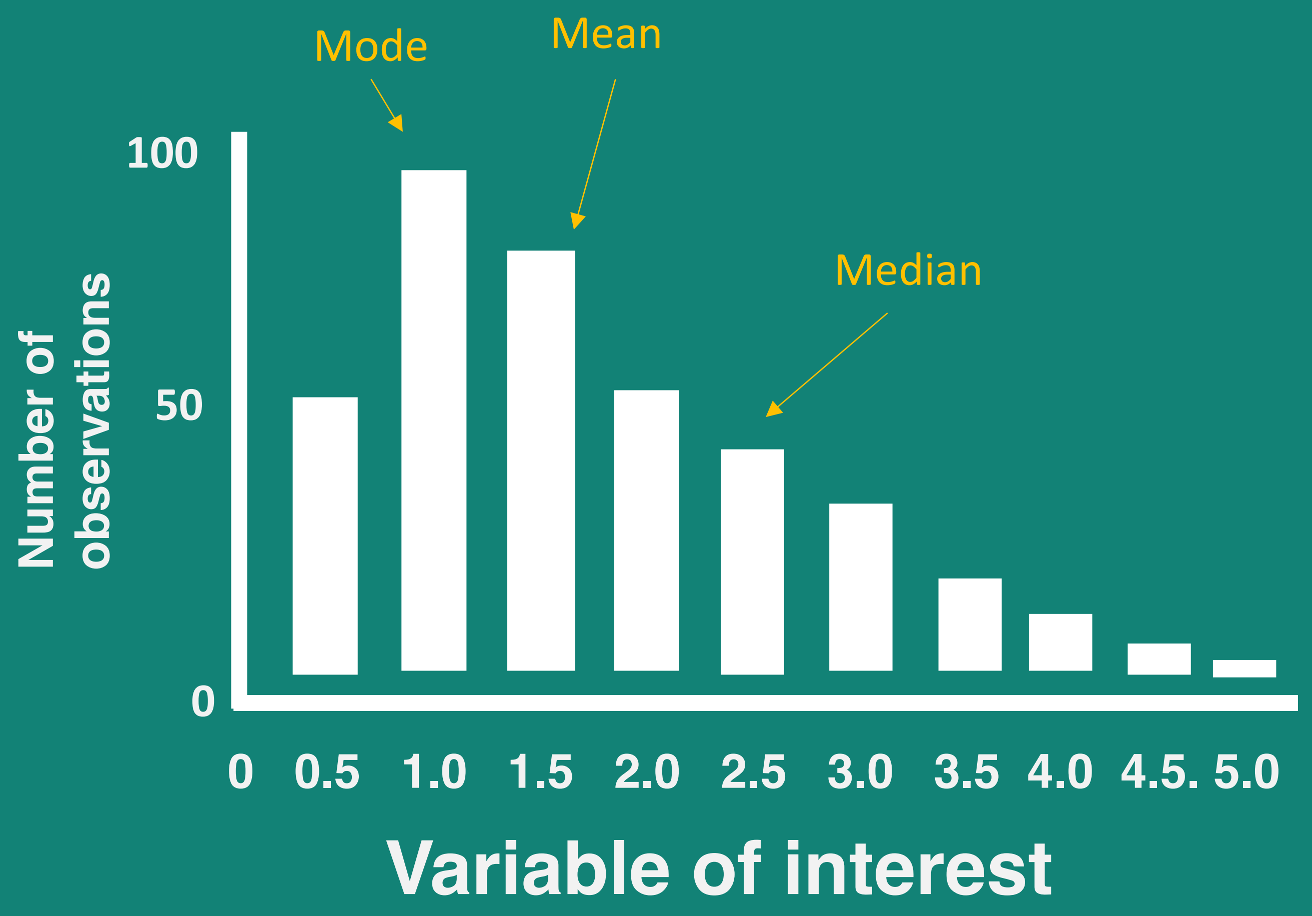
Data distributions

Non-normal

This is not true for other distributions of data

For example, in this example the mean, mode and Median are all different. This is why its useful to have several measures of central tendency.

This is an example of a skewed distribution



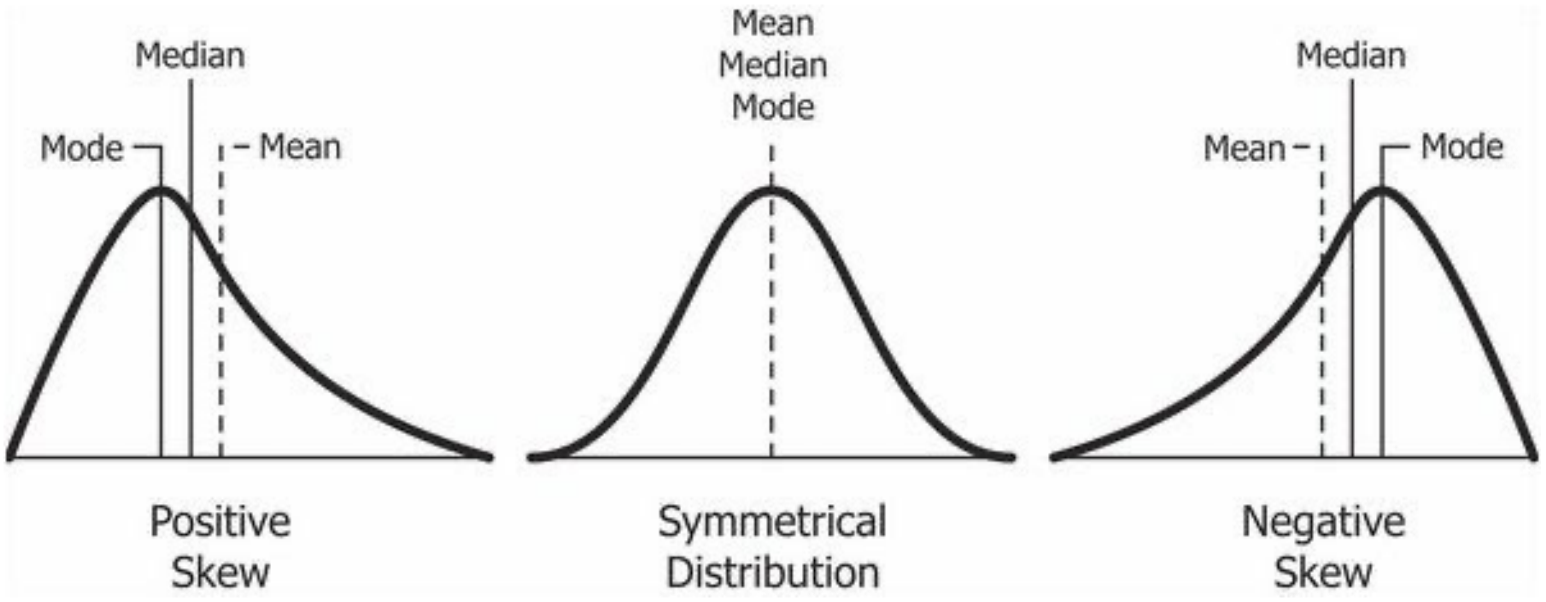
Data distributions

Skewness

A measure of the symmetry of a distribution.

If the peak (i.e. mode) of the distribution is to the left of the mean/median it is positively skewed

If the peak (i.e. mode) of the distribution is to the right of the mean/median it is negatively skewed



Data distributions

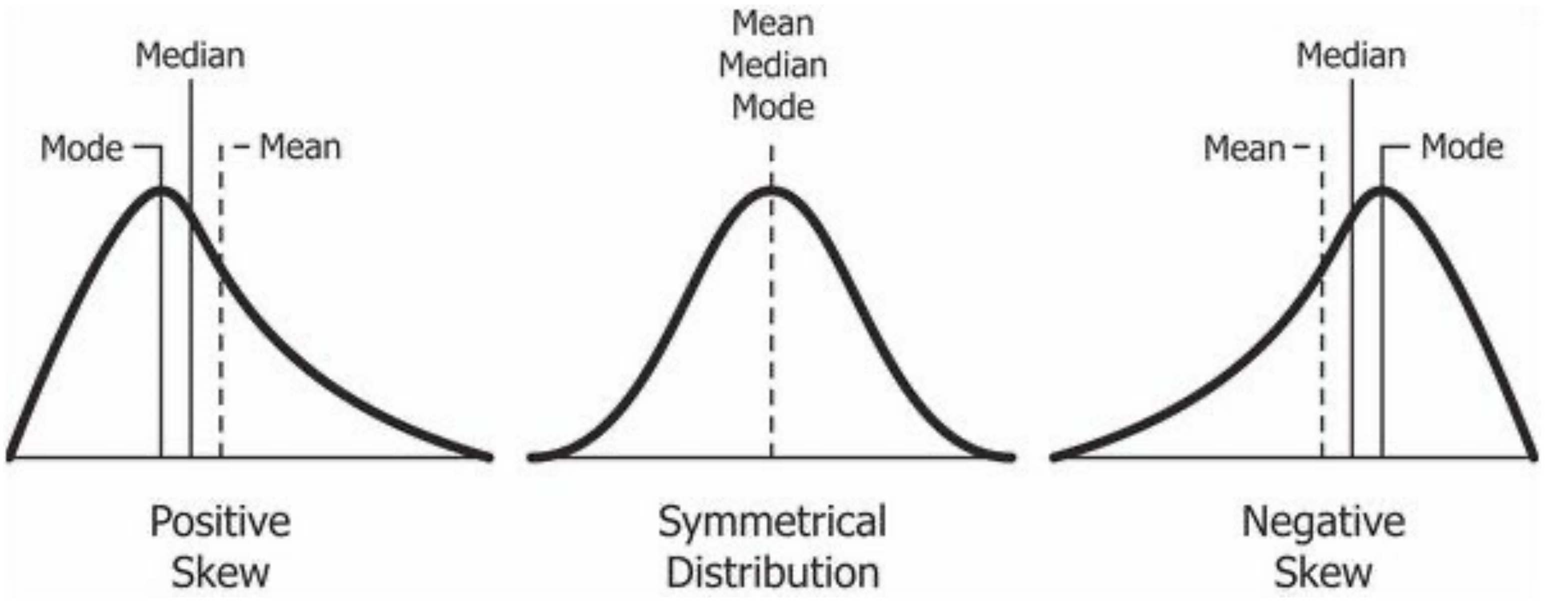
Skewness

This is rarely used as a summary statistic but you can calculate in R using the moments package

```
install.packages("moments")

library("moments")

skewness(iris_data$Sepal.Width)
```



Data distributions

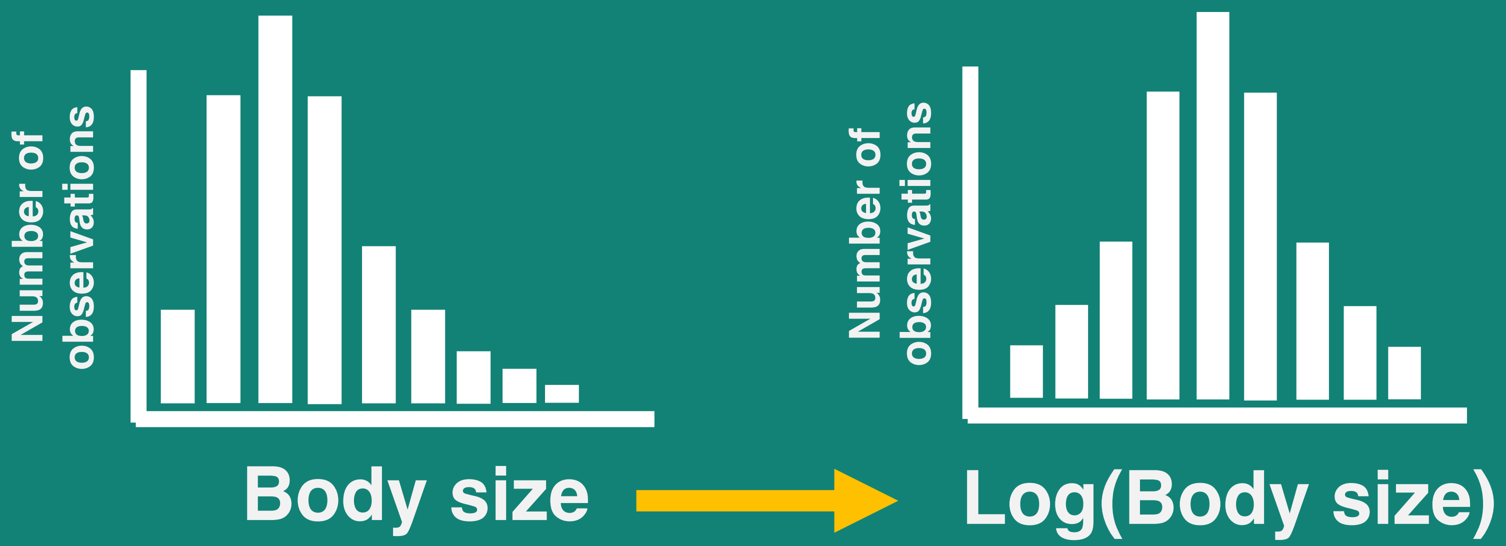
Log -Normal

Log-Normal distribution

Positively skewed

A distribution that when log transformed becomes normal.

Very common in biology especial when scales are very large such as with body size across species.



Log transformation
Make a log normal distribution
Into a normal distribution

Data distributions

Log -Normal

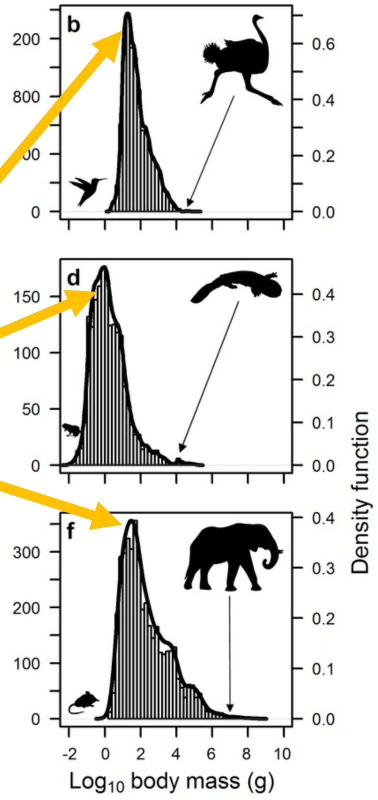
Log-Normal distribution

Positively skewed

A distribution that when log transformed becomes normal.

Very common in biology especial when scales are very large such as with body size across species.

Most animals are small leading to a log normal distribution



Data distributions

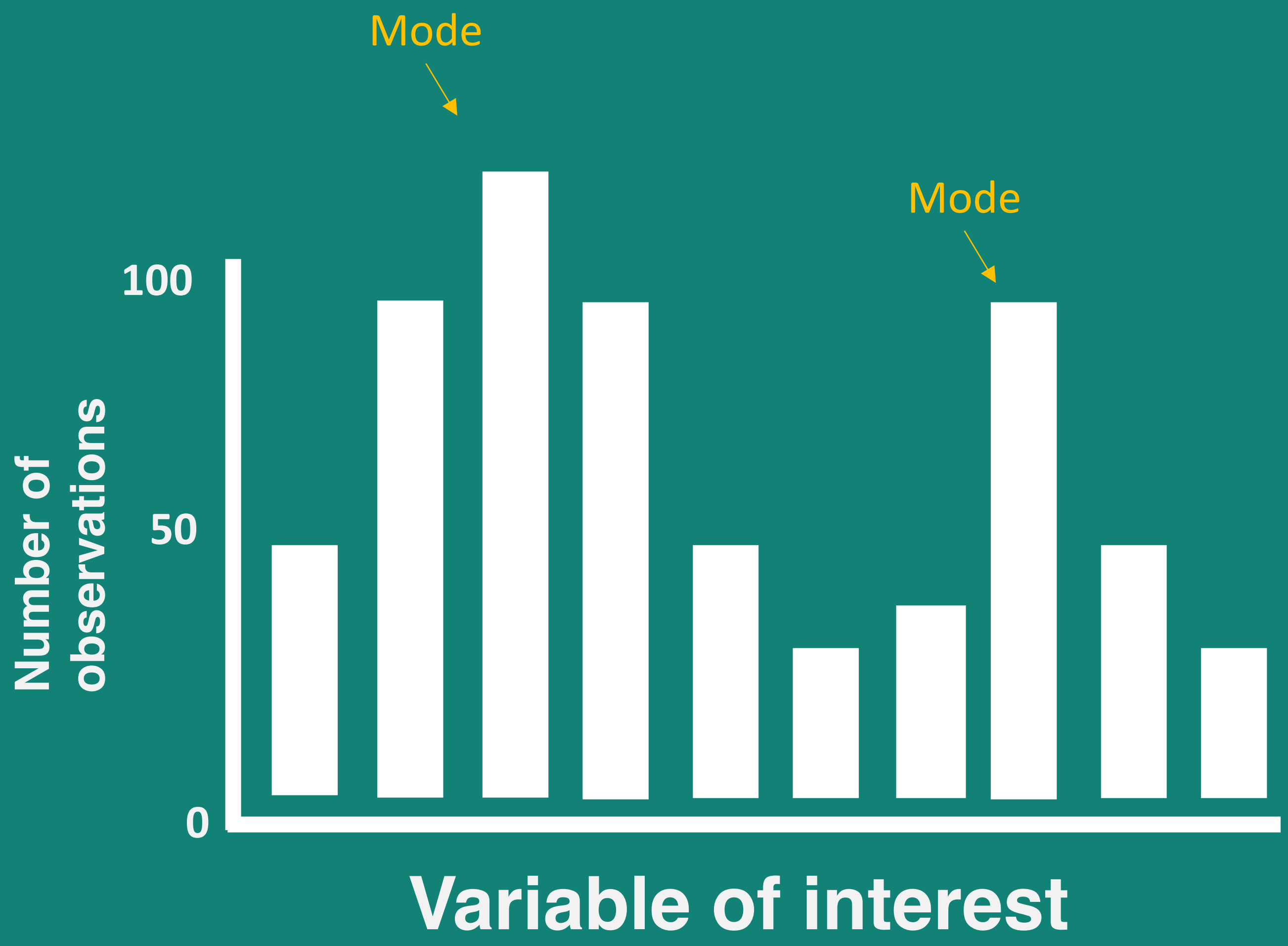
Bimodal

Bimodal distribution

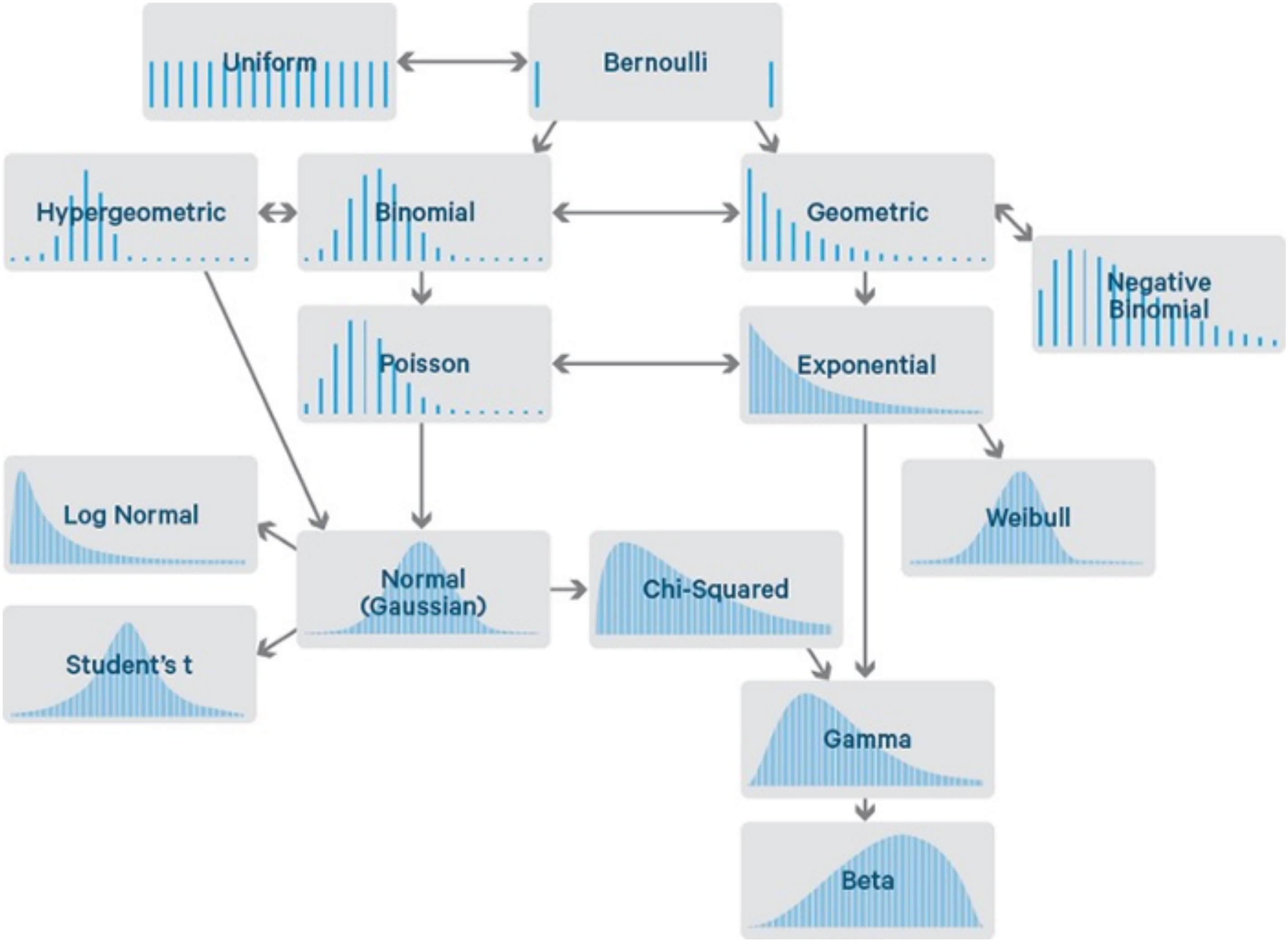
Data with 2 peaks (i.e. 2 modes)

Often can indicate distribution is made up of different groups

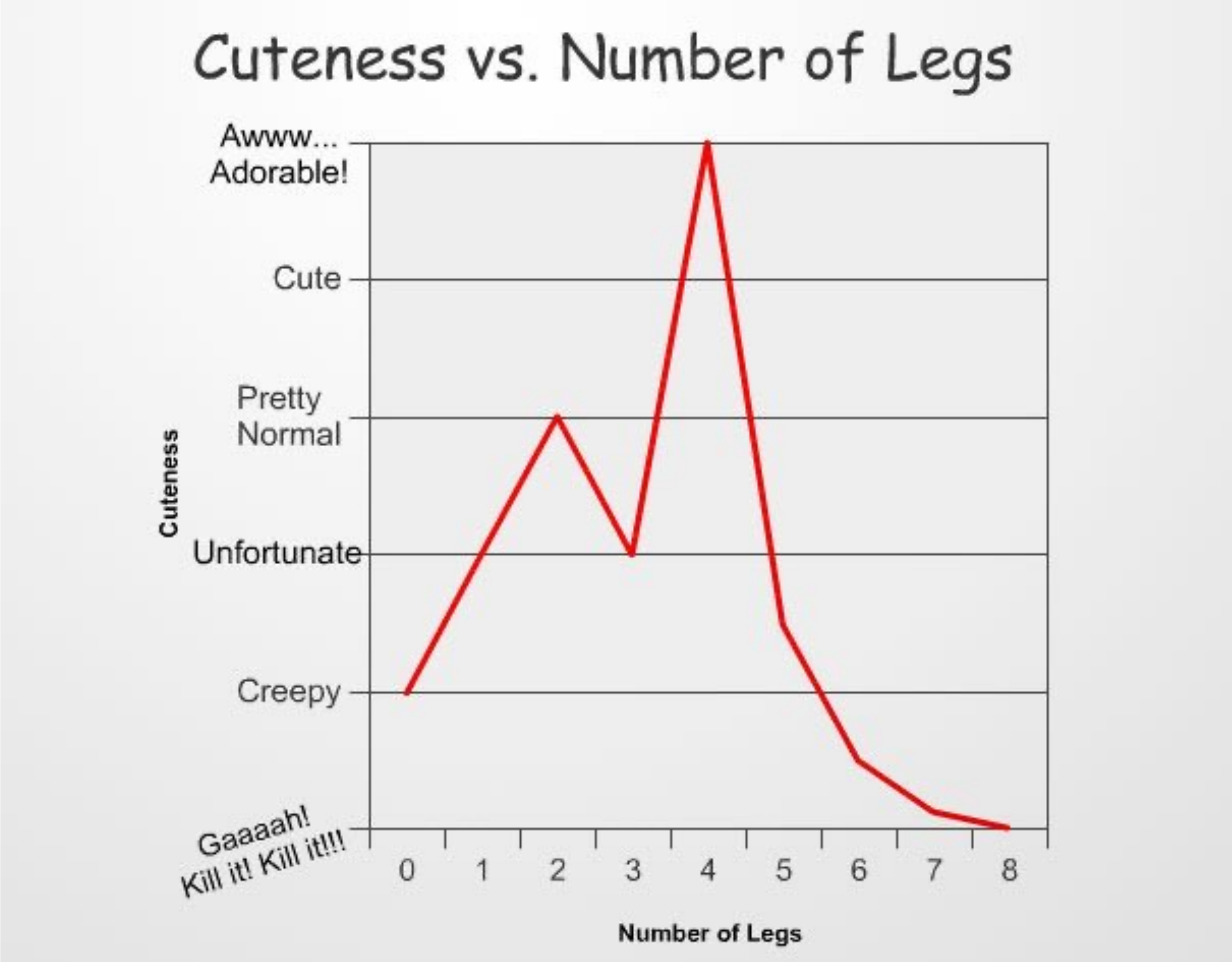
Example, plot the histogram of hist(iris_data$Petal.Width)



There are lots of distributions that describe the shape of data and are linked to the process that generate these shapes.



Plotting data



Use boxplot, violin plots etc. for Qualitative data

Qualitative

Data is in categories

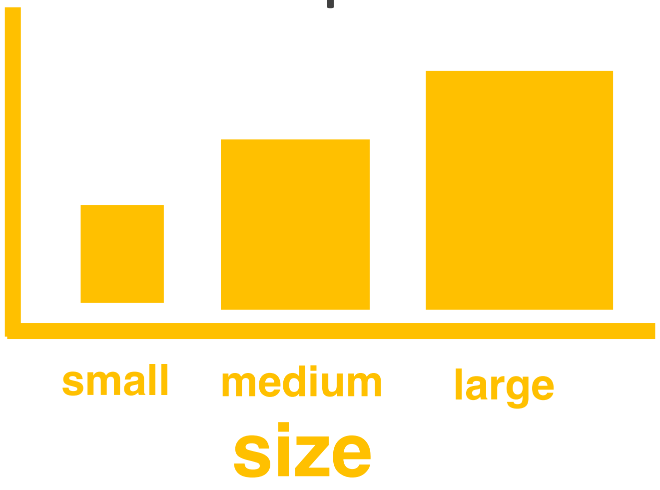
Nominal

Categories with no order e.g. Hair color (red, black)

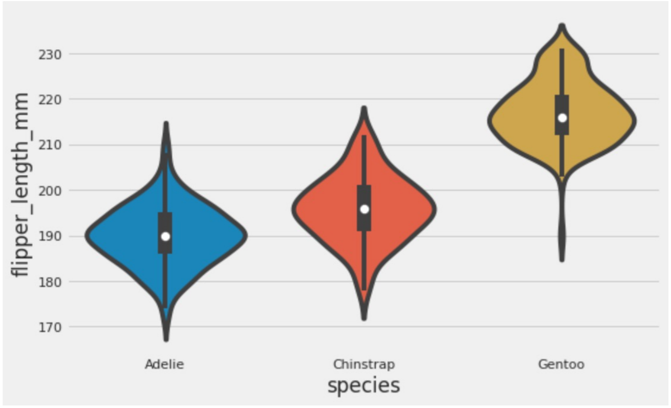
Ordinal

Categories that can have an order e.g. income (low, medium high)

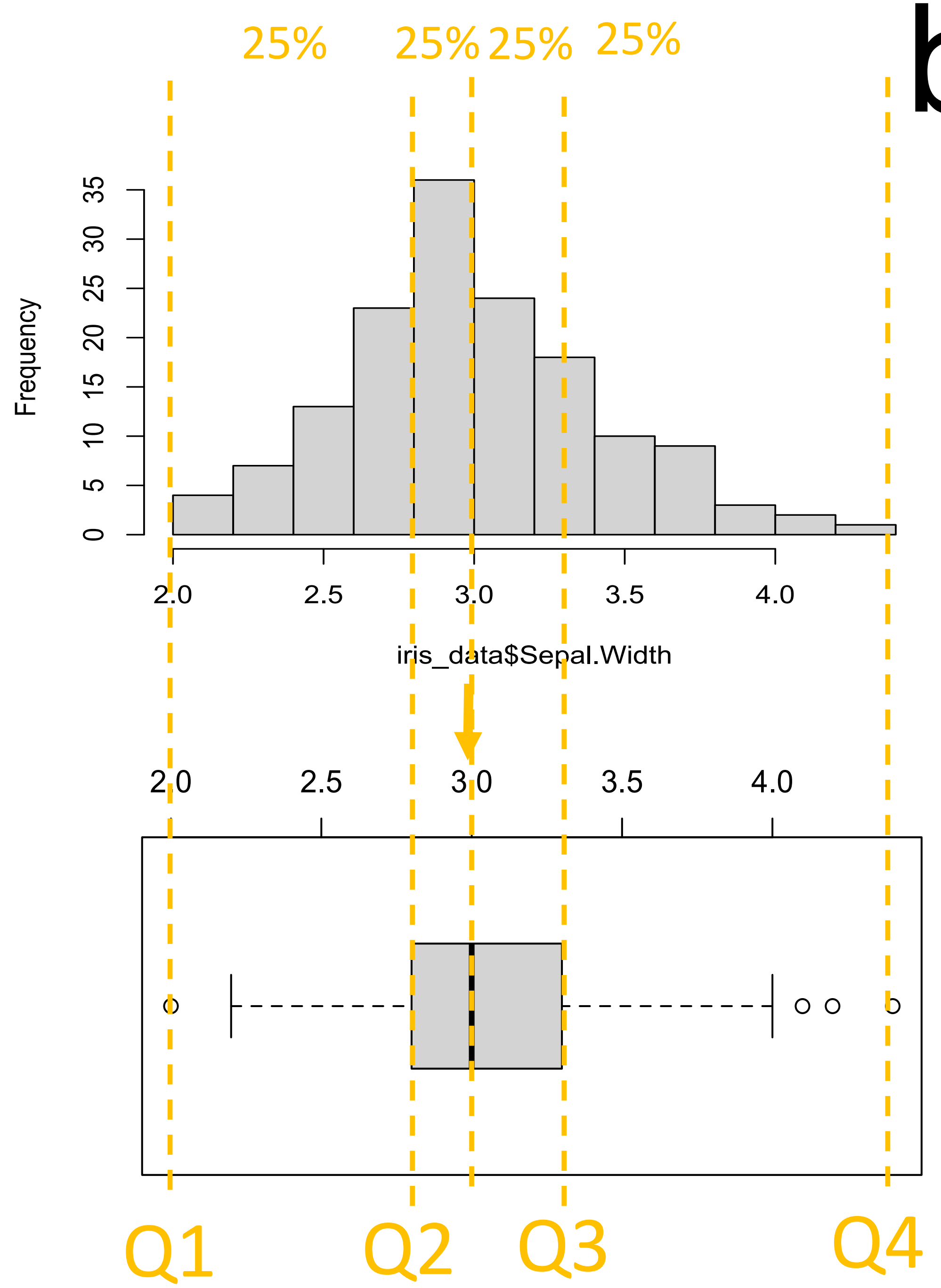
Boxplot



Violin plot



boxplot



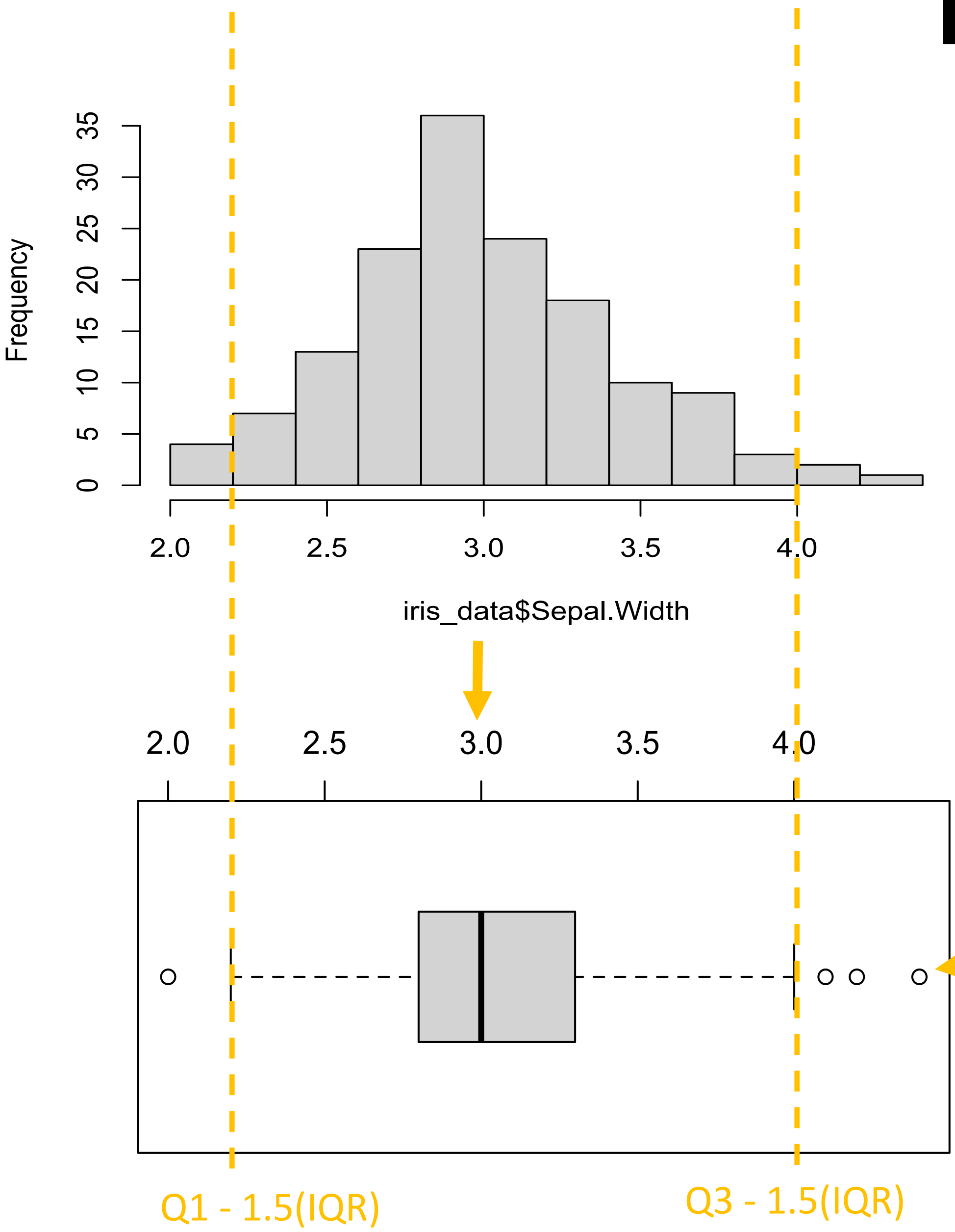
A single box plot is a version of the Distribution simplified using the median and interquartile range

The black line is the median

The distribution can be broken down into 4 sections, called quartiles each with 25% of the data.

The box extends to the 2nd and 3rd quartiles which is called the interquartile range.

boxplot



A single box plot is a version of the Distribution simplified using the median and interquartile range

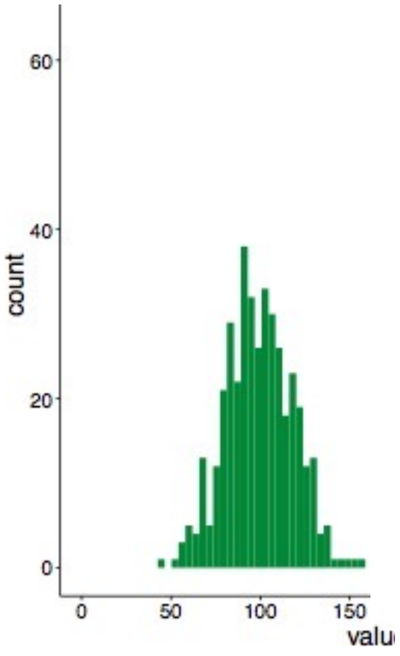
The whiskers extend to 1.5 times the interquartile range (IQR)

The dots at the end are called outliers and fall outside the whiskers

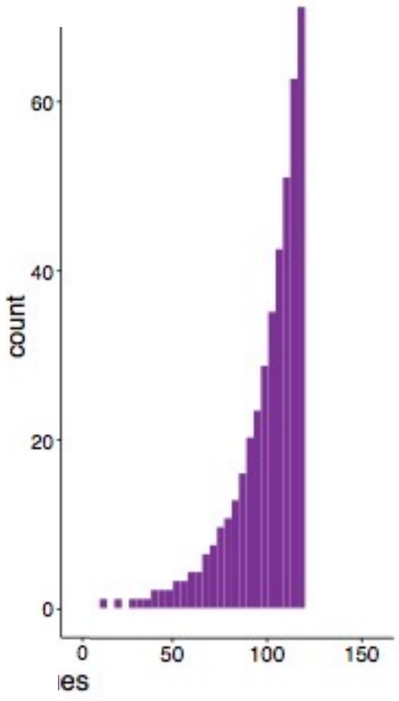
Barchart of boxplot?

Lets compare simulated body size data for 2 groups

They have the same mean and the same standard deviation

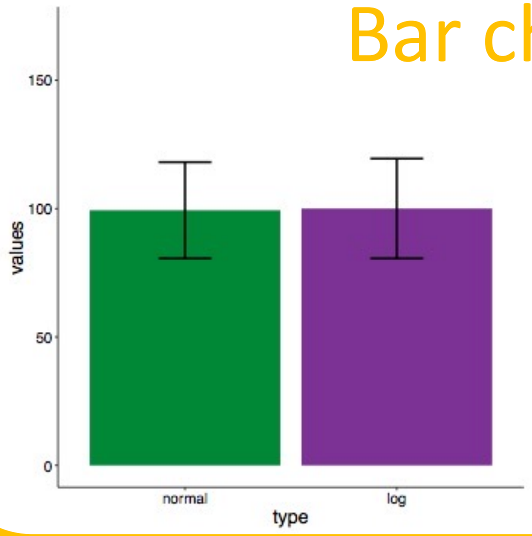


Body size



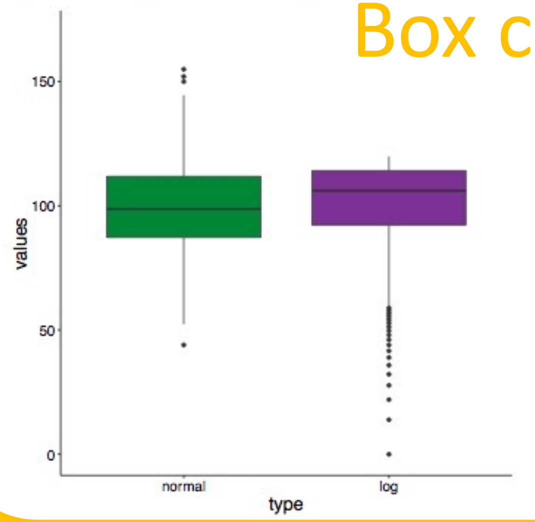
Body size

Bar chart



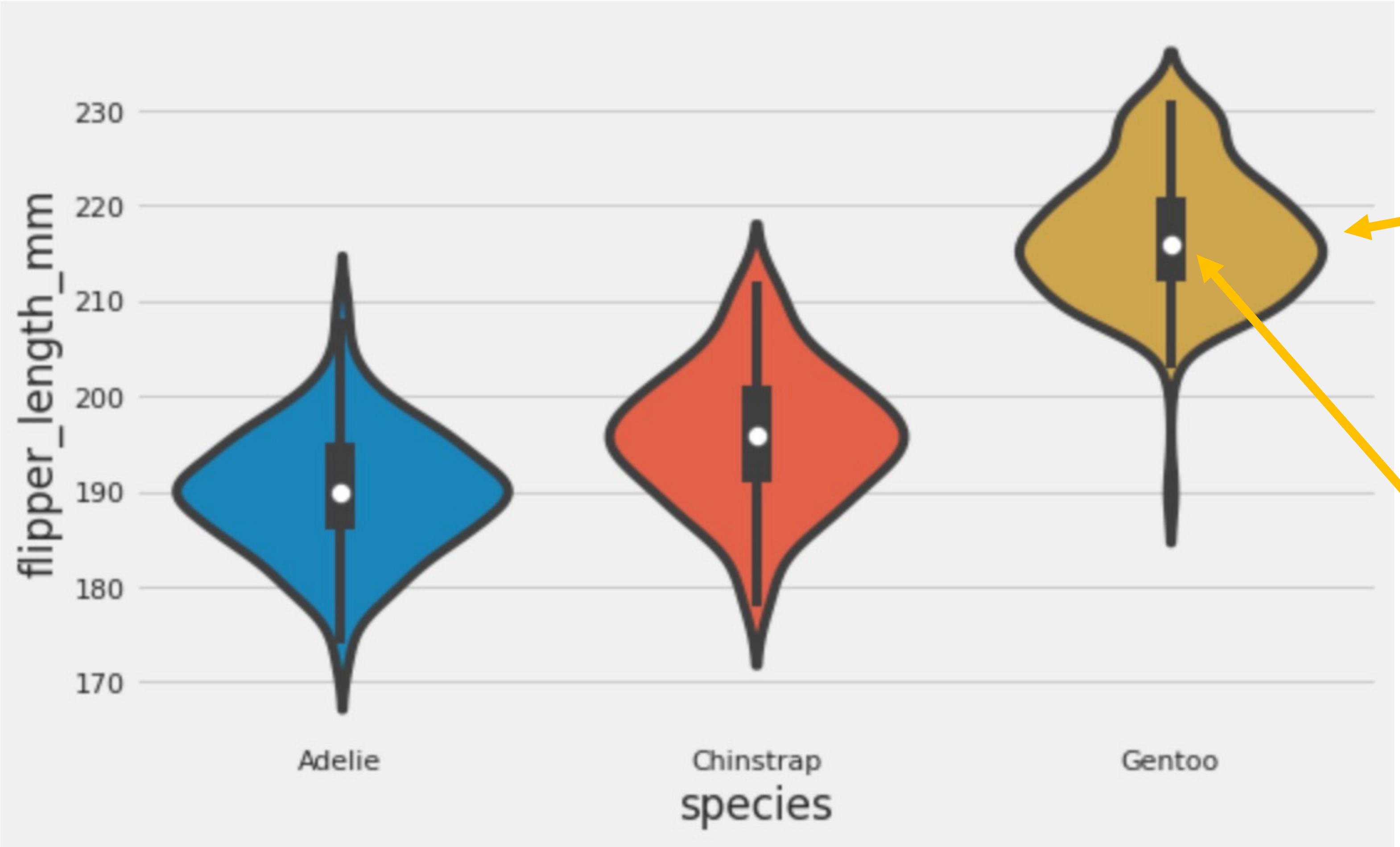
Bar chart indicates that there is no difference

Box chart



A boxplot is better at demonstrating the spread of the data in each of the groups

Violin plots add distribution



A histogram turned on its side to show data distribution

Simplified boxplot for summary statistics

If comparing two quantitative variables

Quantitative

Data can be expressed in numbers

Discrete

Count data that are whole numbers e.g. number of species (1,100, etc)

Continuous

Integers, numbers that can be fractions e.g. body mass (10, 2.5, 4.44)

Scatter plots

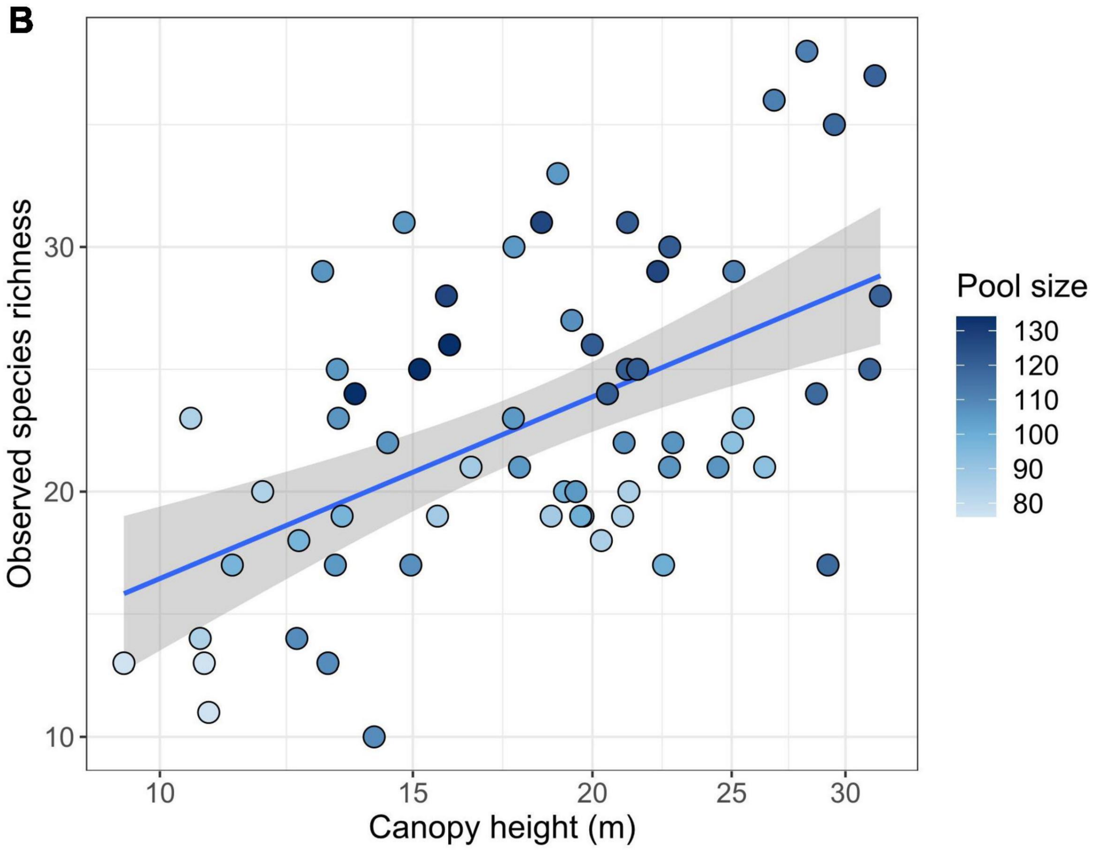
Response variable changes in response to changes to the explanatory variable.

This always goes on the y-axis

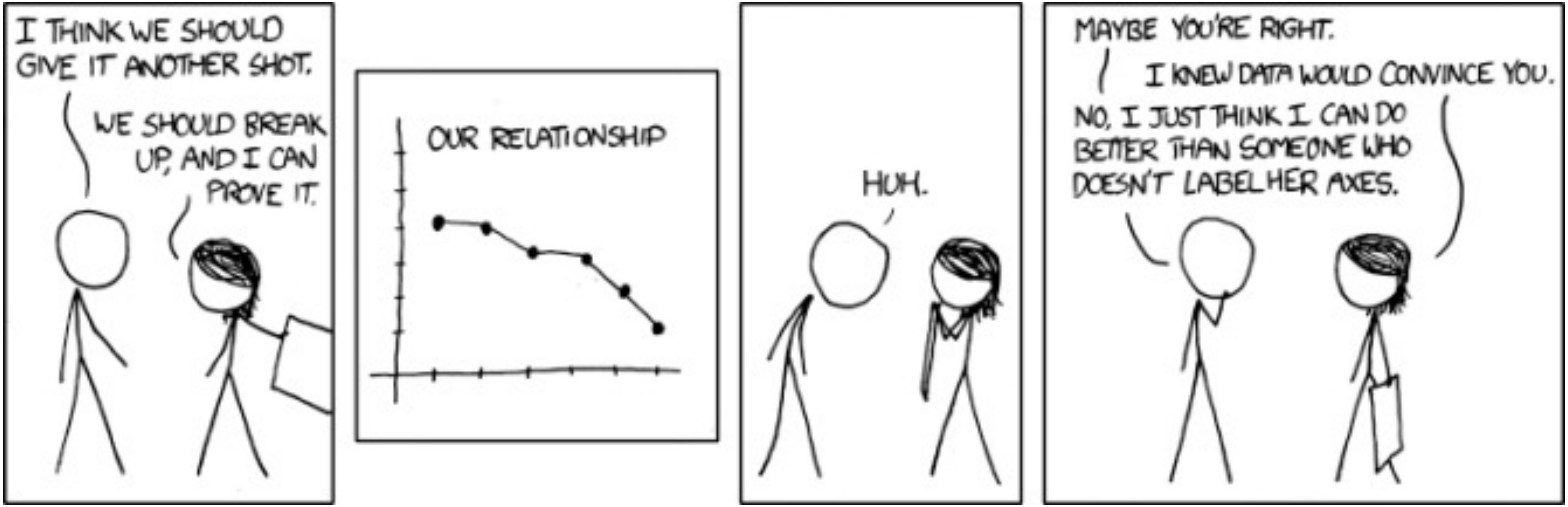
Explanatory variable
This is the variable
That causes a change in the other variable

This always goes on the x-axis

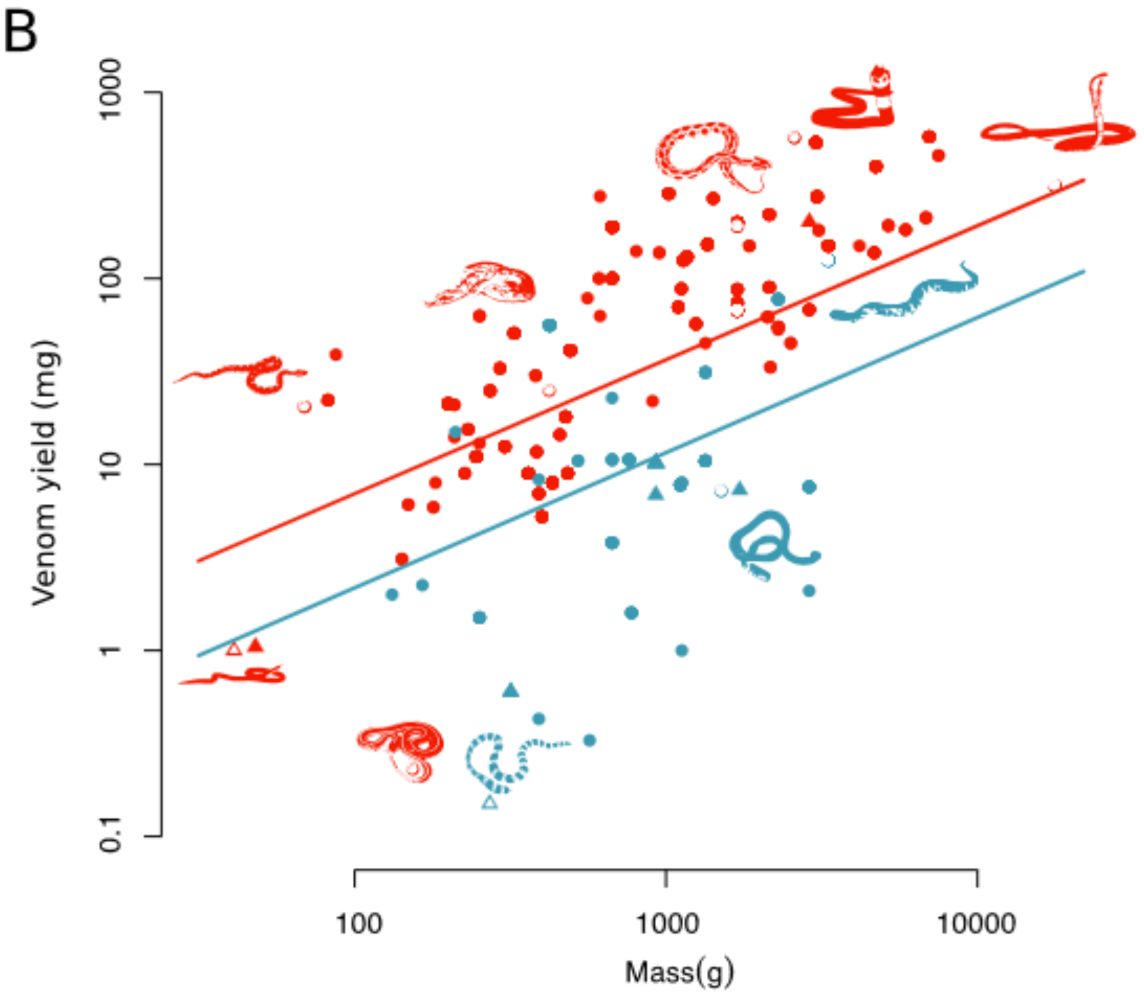
Changes in Canopy height here are predicted to cause changes in species richness. Hence, Canopy height is on the x-axis as the explanatory variable and Species richness on the y-axis as the response variable



Do’s and don'ts of graphs

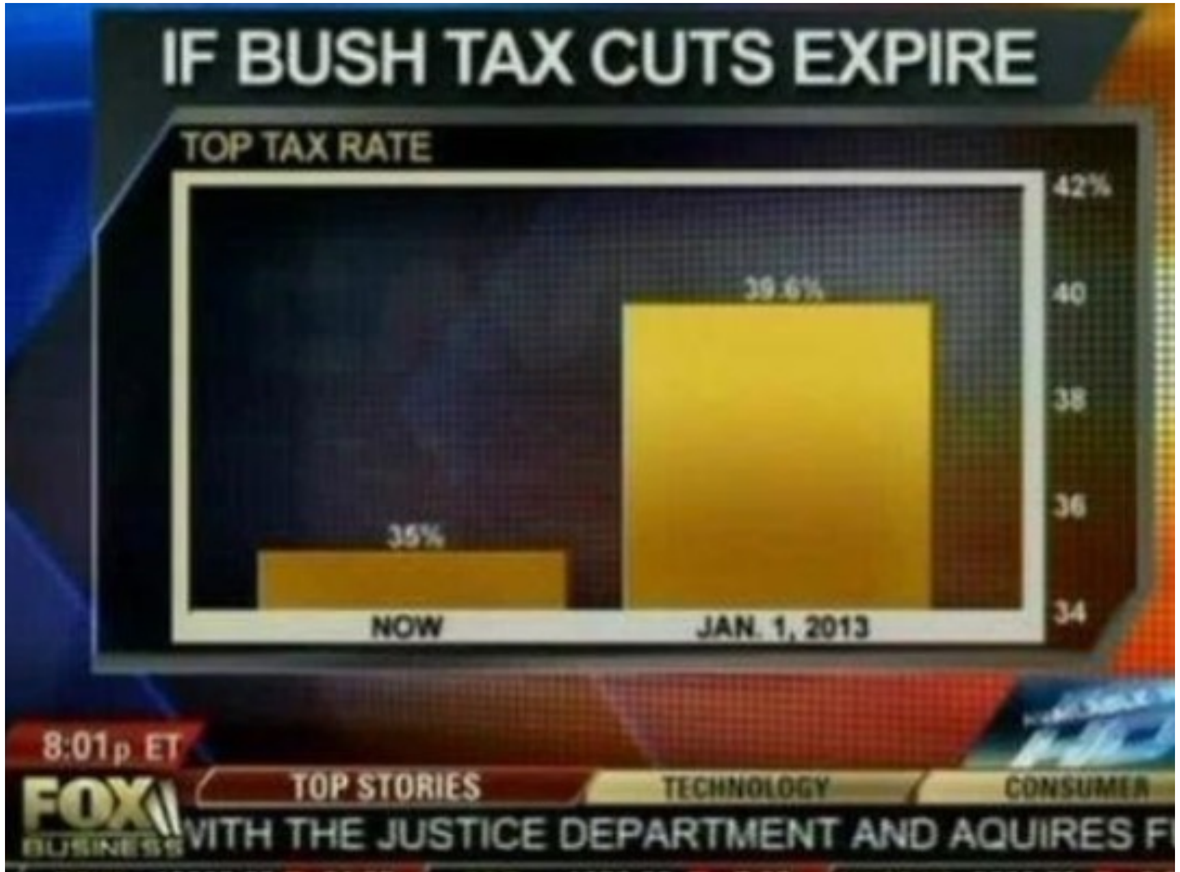


Label the axis and give the units

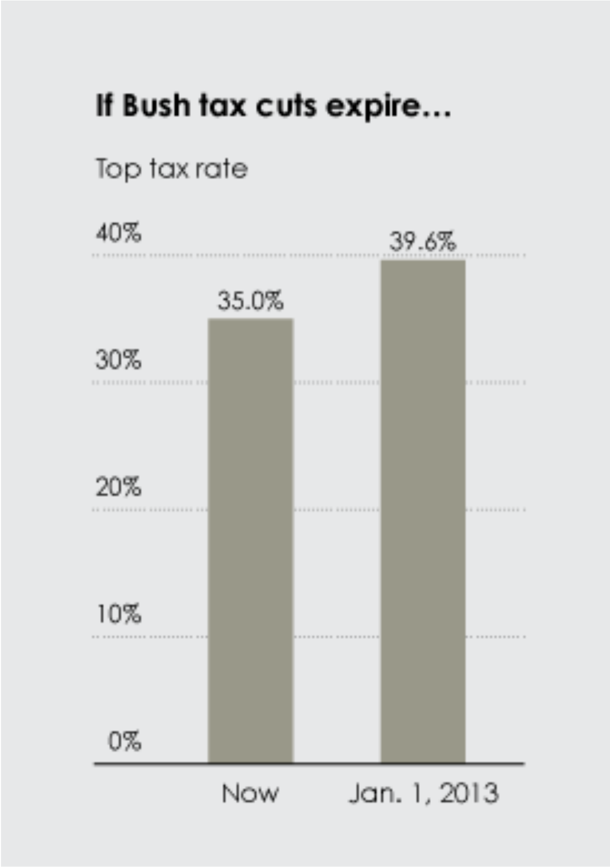


Do's and don'ts of graphs

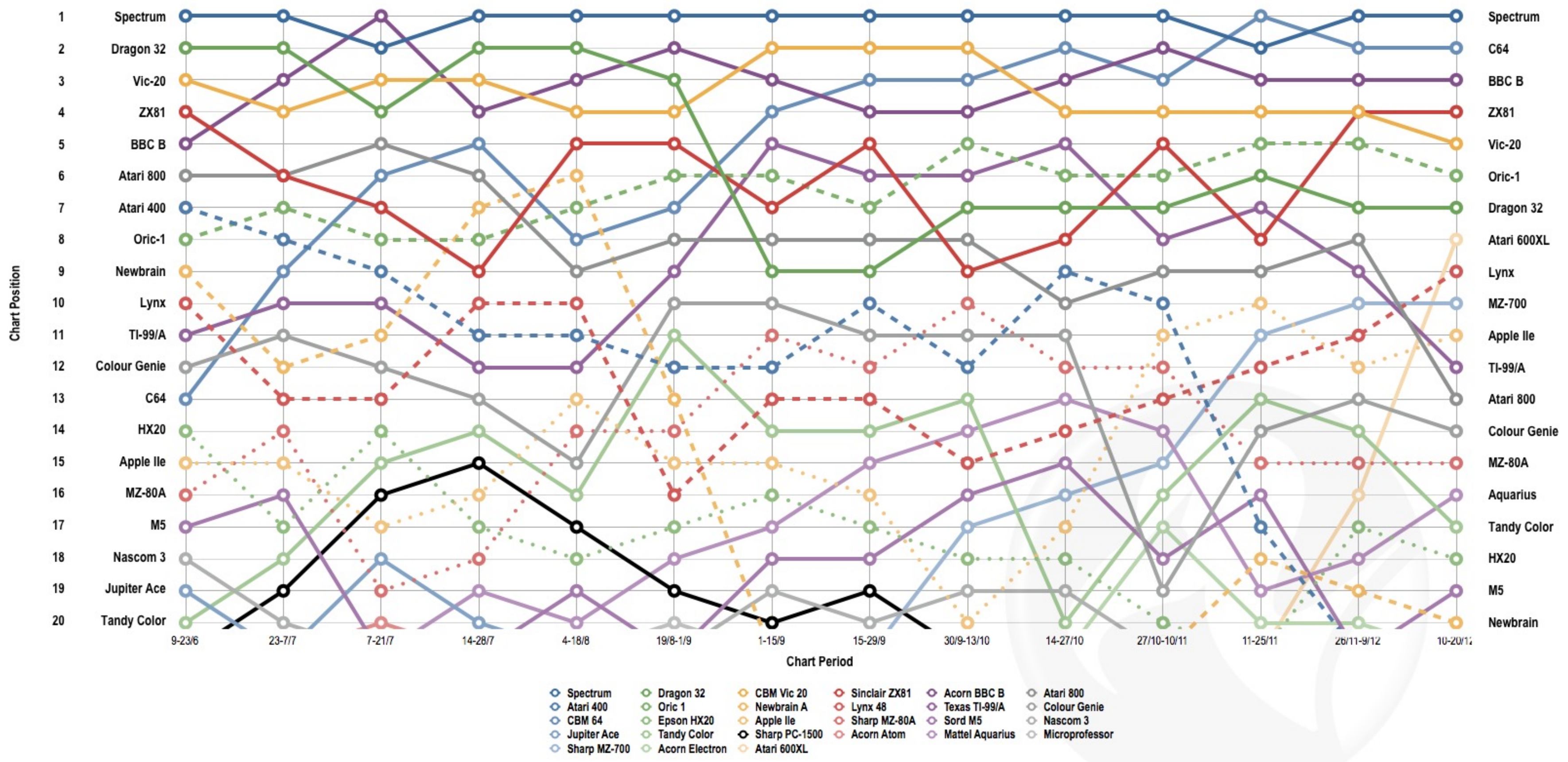
Truncated graph giving false impression of large difference



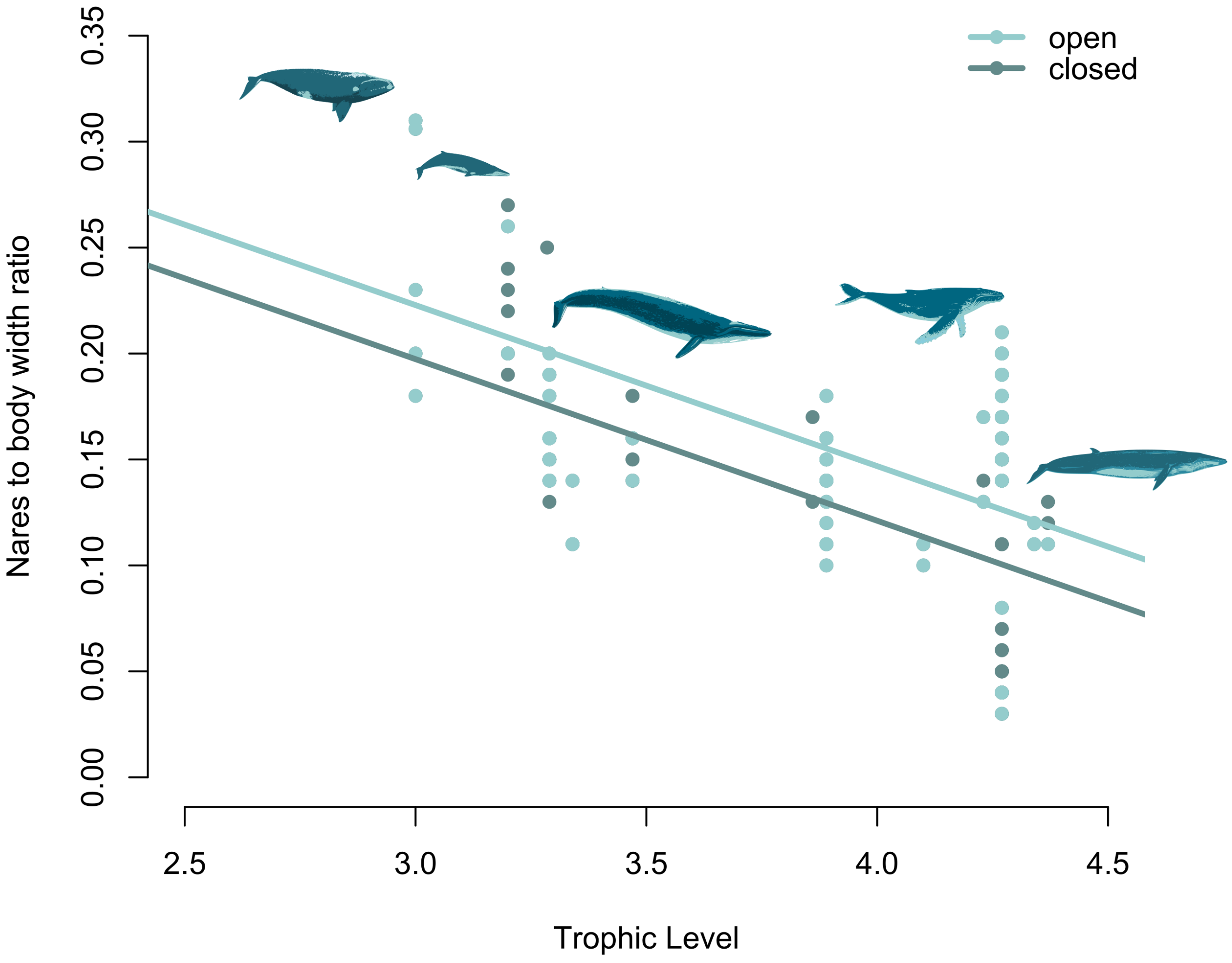
Non-truncated graph show only small differences



Keep it simple
Make sure the main point is clear



Colors

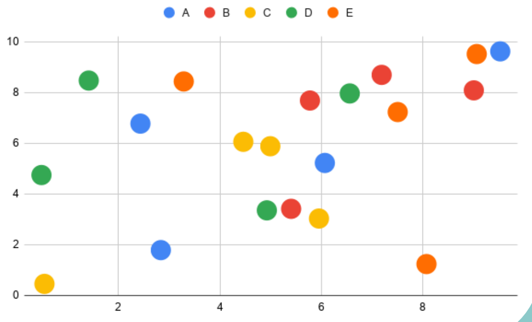


Use color pallets or color gradients

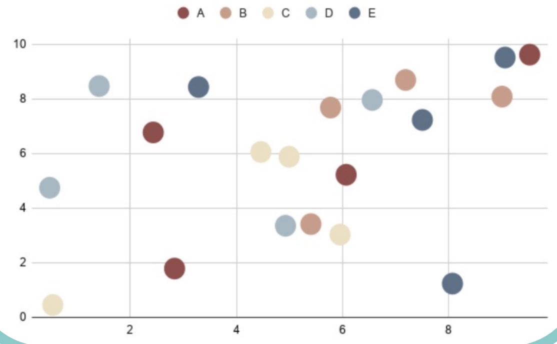
Here I used a color pallet of different shades of blue to based on the whales to distinguee between the two groups.

Colors

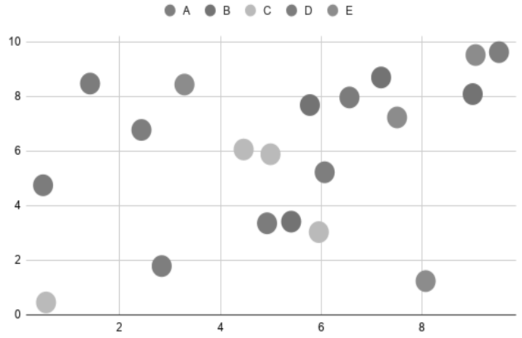
Not color blind friendly



Color blind friendly



Printer friendly too



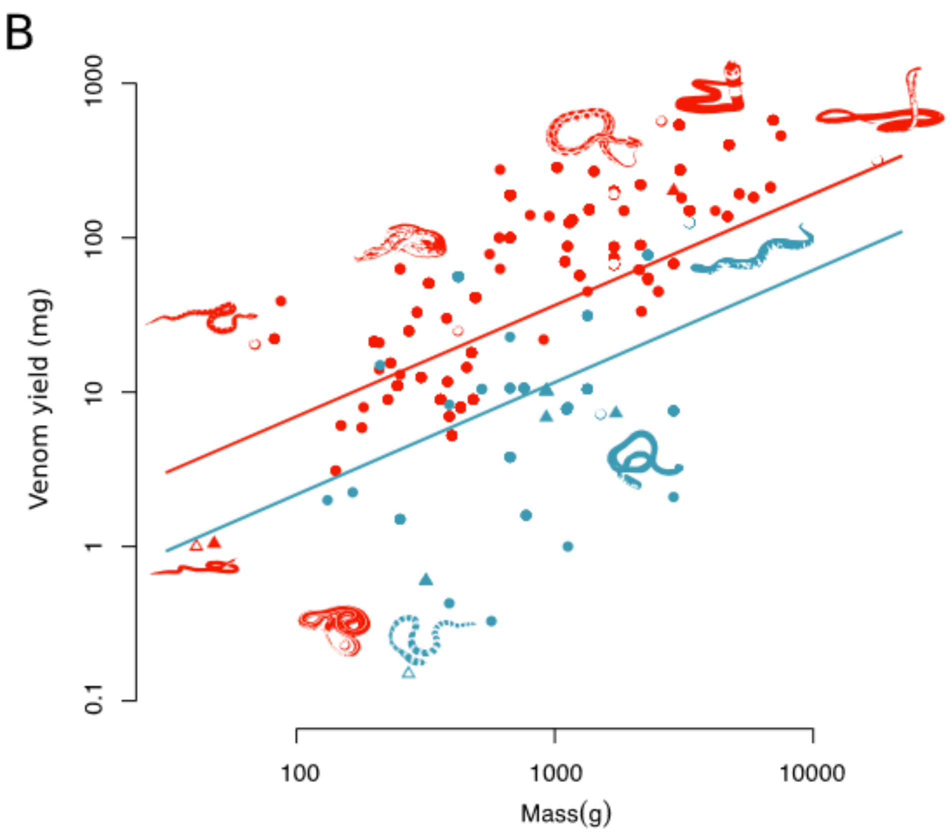
Think of color blindness

Avoid using red with green

Use light and dark shades of colours (if printed in black and white would your graph still work?)

Use colour blind friendly pallets

Colors



Think of color blindness

Avoid using red with green

Use light and dark shades of colours (if printed in black and white would your graph still work?)

Use colour blind friendly pallets