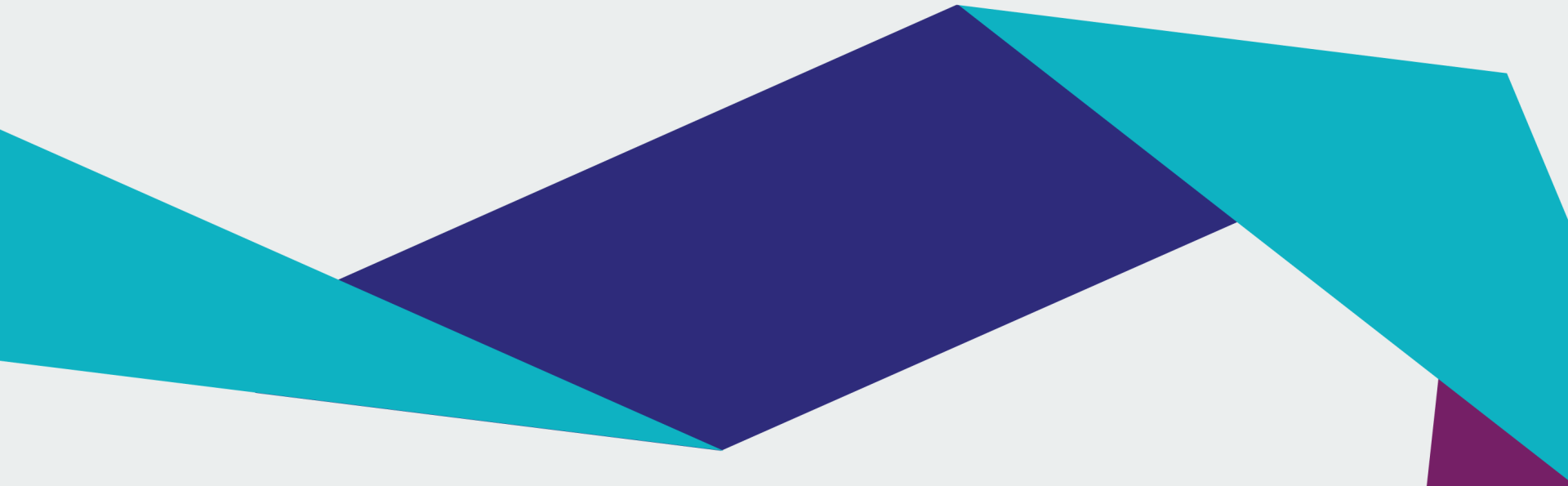


# Basic Statistics

John Pinney



# Course Aims

By the end of this course, you will be better able to:

- **Understand** the need for statistics, their uses and limitations.
- **Interpret** graphical representations of sampled data
- **Calculate** descriptive statistics for sampled data
- **Apply** the normal distribution to answer questions about a population
- **Identify** potential sources of bias in a scientific study

- 1. What is statistics?

# The biggest question(s)...



What are statistics and why should we use them?



Can you name some common statistics we all encounter?

by Jef Mallett

May 08, 2006



# Statistics rule the world...

- Weather forecasting
- All medicine (sample data + stats required to prove a medicine works!)
- Election polls
- Economic / financial forecasting (e.g. your wages, rent, council tax, utility bills, road tax)
- Spotify, Facebook, Google, Instagram, Snapchat, news outlets, Amazon
- University league tables
- Prison sentences
- University grades
- London underground timetabling
- “All” scientific discoveries
- H&S rules
- The internet





# Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett<sup>1</sup>, Abigail A. Baird<sup>2</sup>, Michael B. Miller<sup>1</sup>, and George L. Wolford<sup>3</sup>

<sup>1</sup> Psychology Department, University of California Santa Barbara, Santa Barbara, CA; <sup>2</sup> Department of Psychology, Vassar College, Poughkeepsie, NY;

<sup>3</sup> Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

## INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

## METHODS

**Subject.** One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

**Task.** The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

**Design.** Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

**Preprocessing.** Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a T<sub>1</sub>-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

**Analysis.** Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was included to account for low frequency drift. No autocorrelation correction was applied.

**Voxel Selection.** Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

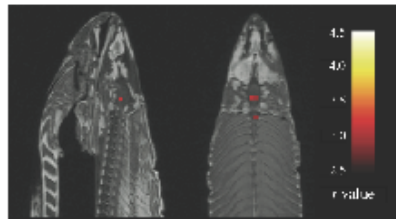
## DISCUSSION

Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ( $p < 0.001$ ) and low minimum cluster sizes ( $k > 8$ ) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

## REFERENCES

- Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.
- Friston KJ, Worsley KJ, Frith CD, Frith CD, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

## GLM RESULTS

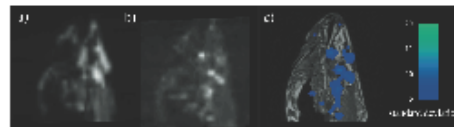


A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were  $s(131) > 3.15$ ,  $p(\text{uncorrected}) < 0.001$ , 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm<sup>3</sup> with a cluster-level significance of  $p = 0.001$ . Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ( $p = 0.25$ ).

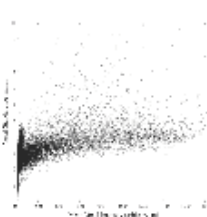
## VOXELWISE VARIABILITY



To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T<sub>1</sub>-weighted image.

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ( $r = 0.54$ ,  $p < 0.001$ ). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.



...but statistics handled incorrectly can lead to false conclusions!

# The biggest answer(s)...

Statistics means *handling data effectively*:

- collection, organisation, analysis, interpretation & presentation



# The biggest answer(s)...

We use statistics to:

- **Analyse** a full set of data to produce simple summary metrics (**descriptive statistics**)
- **Calculate** something we can't directly measure (**inferential statistics**)

# Population vs sample


- Population – the entire set you wish to learn something about (e.g. every child in the UK)
- Sample – the set for which you have data for (e.g. the people who answer your questionnaire)
- The population includes the sample... but a sample usually doesn't cover the whole population!

# Descriptive Statistics

- Tools for analysing & interpreting sampled data and presentation of results
- e.g. what is the average weight of a child who drinks 25 cans of fizzy drinks a day?
- Can identify (and communicate):
  - Typical (central) values;
  - Shape and spread of the distribution of values;
  - Interesting patterns and relationships in the data.

# Inferential Statistics

Use information on a **sample** of individuals to make **inference** about the wider **population** of like individuals.

- You are investigating the Body Mass Index for children in the UK. You cannot study exhaustively the population of every child, due to money, time, effort etc...  you restrict your investigation to a **well chosen** sample
- With statistics, we can infer something about the population using only the data from the sample

# From Sample to Population

- **Population:** The totality of subjects of interest
- **Sample:** The subset of the population we actually studied.
- **Key idea:** take results obtained in the sample and use them as our best estimate of the true population values of interest.

e.g. we know the sample mean of BMI,  $\bar{\mathbf{x}} = 23.9$ , use this to estimate the population mean.

# Parameters vs. Statistics

- A **parameter** is a number that describes the population (theoretical quantity, cannot be observed)
  - It is a fixed number but we do not know its value
  - e.g. Theoretical mean and variance,  $\mu$  and  $\sigma^2$ , of BMI of UK population
- A **statistic** is a number calculated from the data used to describe/summarize the data
  - e.g. Sample mean and variance (  $\bar{x}$  and  $s^2$ ) of BMI of students in this class
- We can use statistics such as  $\bar{x}$  and  $s^2$ , to estimate unknown parameters of interest. This is **Statistical Inference**.

# Types of Data: Categorical

## No numerical relationship between categories

- **Nominal Data:** no obvious ordering of categories.
  - Favourite colour: green, blue, orange
  - When there are only 2 possible categories, data is called **dichotomous or binary** (i.e. 0 = have been to Disneyland, 1 = haven't been).
- **Ordinal Data:** there is natural order of the categories.
  - Height: short, medium, tall
  - Physical Activity: Low/Moderate/High

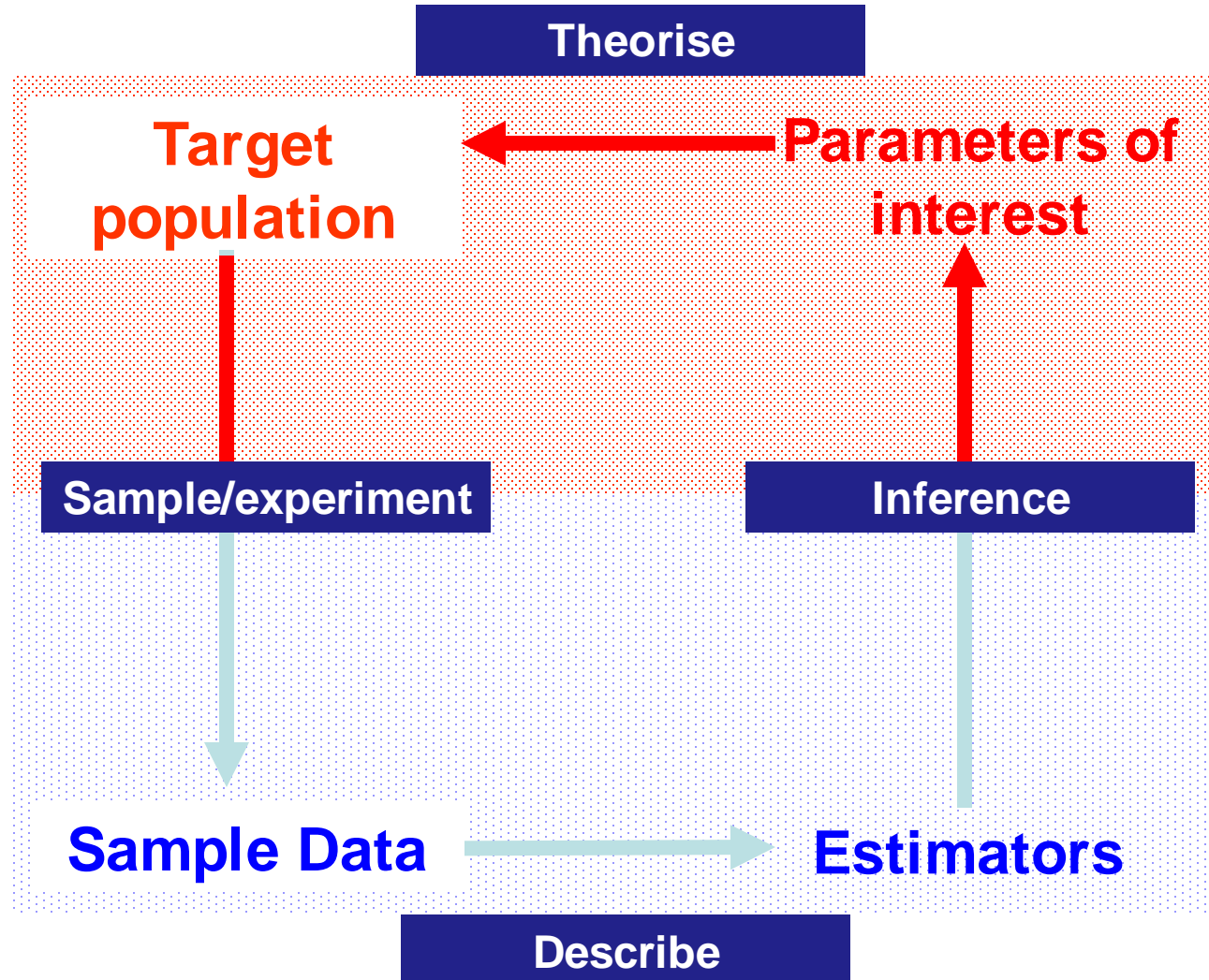
# Types of Data: Quantitative

**Data are numerical, arising from counts or measurements**

- **Discrete data:** can only take specified values, e.g. integers.
  - number of children,
  - number of beds available in a hospital.
  - Shoe size
- **Continuous data:** measurable quantities with possible values in an interval/range
  - BMI
  - Blood pressure
  - Length of foot



# Statistical analysis



# Steps in a research project

1. Formulation of hypotheses (e.g. people who do not eat fruit and vegetables have a higher BMI)
2. Planning and design (what data should be collected?)
3. Data collection
4. Data processing and exploratory displays
5. Data analysis
6. Interpretation
7. Presentation of results (e.g. paper)



**Statistical  
thinking  
involves  
all these  
steps**

In this course, our focus is on *descriptive statistics*. We want to be able to:

- **Quantify the effects of interest, so they can be compared between different groups**
- **Quantify variability, so we can assess whether differences between groups are real**

- 2. Displaying Data

# Displaying Data

Several possible methods

➤ **Tables**

- Frequency Tables

➤ **Figures**

- Bar Charts
- Histograms
- Scatter plots
- .....

# Frequency Tables

List all possible categories with numerical counts corresponding to each one.

**Prevalence of overweight and chronic energy deficiency (CED) among women in rural and urban areas in Bangladesh 2000-2004**

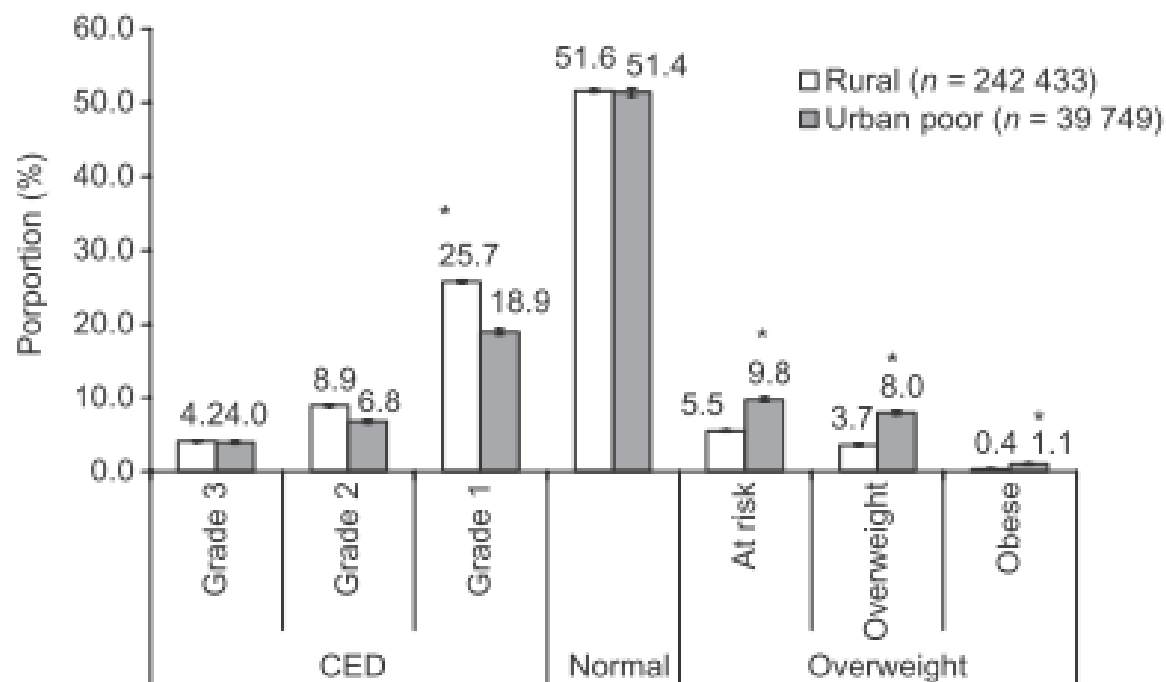
	Rural	Urban
CED Grade 3	10182 (4.2%)	1590 (4%)
CED Grade 2	21577 (8.9%)	2703 (6.8%)
CED Grade 1	62305 (25.7%)	7512 (18.9%)
Normal	125095 (51.6%)	20431 (51.4%)
At risk of overweight	13334 (5.5%)	3895 (9.8%)
Overweight	8970 (3.7%)	3180 (8.0%)
Obese	970 (0.4%)	437 (1.1%)

n = 242433

n = 39749

Shafique S. et al., Trends of under- and overweight among rural and urban poor women indicate the double burden of malnutrition in Bangladesh, *Int.J.Epi.*, 2007; 36:449-457

# Bar Charts



Categories on horizontal axis

Vertical bar for each category, height of bar represents frequency

Shafique S. et al., Trends of under- and overweight among rural and urban poor women indicate the double burden of malnutrition in Bangladesh, *Int.J.Epi.*, 2007; 36:449-457

# Histograms

- One of most common graphs of numerical data – unlike bar graph can be used for discrete and continuous data.
- A histogram is like a bar graph, with values grouped in intervals (classes for discrete, bins for continuous) on x-axis
- Effectively, the histogram has both x and y-axes as numerical – therefore the area, not height, is the crucial metric here!



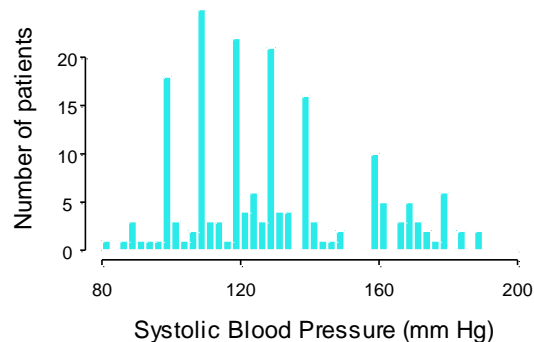
# Histograms: warning

- The number of classes / the bin size can distort the data – we must be careful when selecting a bin size!

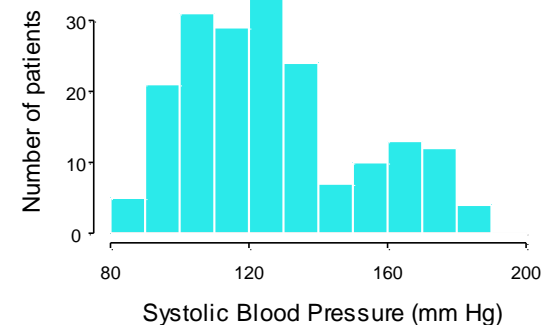
**Example: Systolic  
Blood Pressure  
190 persons, age 40-59**

**Different numbers of  
classes**

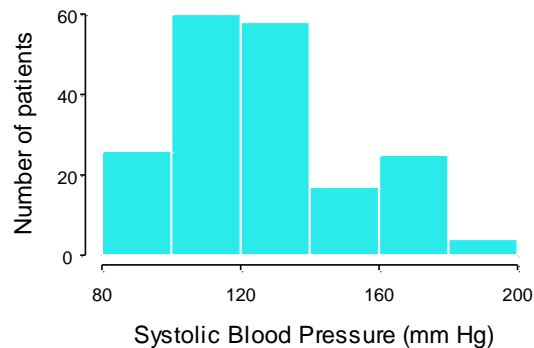
48 classes



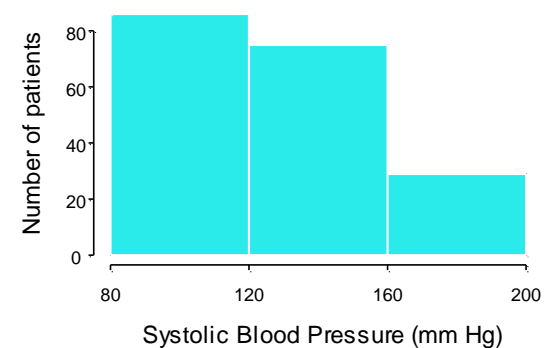
12 classes



6 classes



3 classes

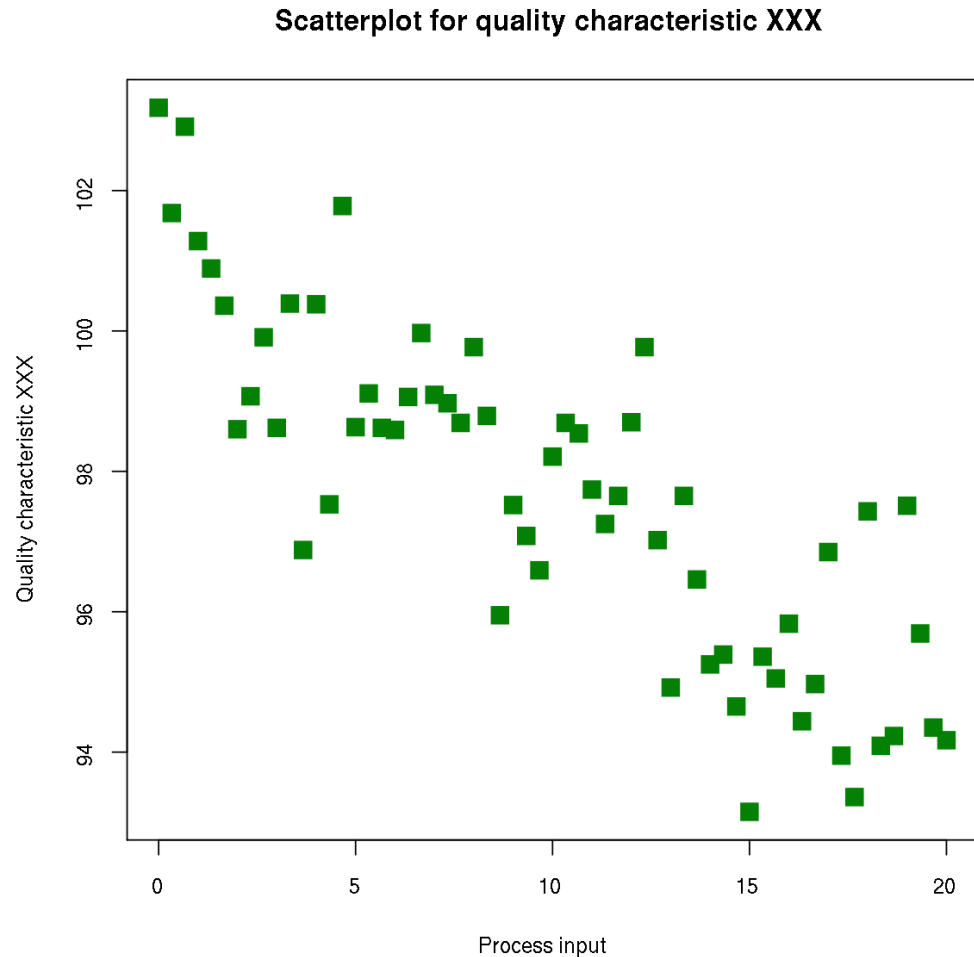


# Selecting Bin Size

- Look for **natural divisions in the data** – e.g. if we're dividing exam marks, can we have the bins as grade boundaries?
- Try and keep **all bins the same size & include all the data** – outliers can skew, so these must be treated carefully.
- Use whole numbers and numbers divisible by the same size if possible (e.g. for 20 samples, use a bin size of 5 – not 6 or 7!)
- If comparing data sets, use **same binning parameters** – this makes the two sets directly comparable.

# Scatter plot

We can display two variables simultaneously using a **scatter plot** (e.g.  $x$  vs  $y$ )



- 3. Descriptive statistics

# Example (contd.)

Need to **quantify** the effects of interest:

- BMI
- Soft drink consumption
- We also need to quantify the **dispersion** e.g. are the variations between individuals too great to draw a definitive conclusion?

# Arithmetic mean

- Denote data by  $x_1, x_2, \dots, x_n$
- **Arithmetic mean** given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Can be used for both continuous and discrete data – not for categorical data

# Mean: Example

You are investigating the Body Mass Index for people with different diets

Diet 1: low consumption of  
fizzy drinks, n=10

24.1	23.5	18.5	16.7	26.3	28.5
25.2	23.4	22.5	29.9		

$$\bar{x} = (24.1+23.5+18.5+16.7+26.3+28.5+25.2+23.4+22.5+29.9)/10 = 23.9$$

Diet 2: high consumption of  
fizzy drinks, n=12

22.1	20.5	18.5	16.9	24.1	26.0
22.2	28.4	21.5	31.9	23.0	17.2

$$\bar{x} = (22.1+20.5+18.5+16.9+24.1+26+22.2+28.4+21.5+31.9+23+17.2)/12 = 22.7$$

# The Median

- **Median** is the middle observation - such that 50% of data lies below its value. Order the data and select the middle value (if  $n$  is even, take the midpoint of the middle two values)
- Main advantage of median over arithmetic mean is insensitivity to outliers (extreme points).
- This is also the main disadvantage...



# Median: Example

Same example as before, let's concentrate on Diet 1

Diet 1: low consumption  
of fruit/veg, n=10

24.1	23.5	18.5	16.7	26.3	28.5
	25.2	23.4	22.5	29.9	

Order the data **16.7, 18.5, 22.5, 23.4, 23.5, 24.1, 25.2, 26.3, 28.5, 29.9**

**n** is even, so we need to  
take the 5<sup>th</sup> and the 6<sup>th</sup>  
observations

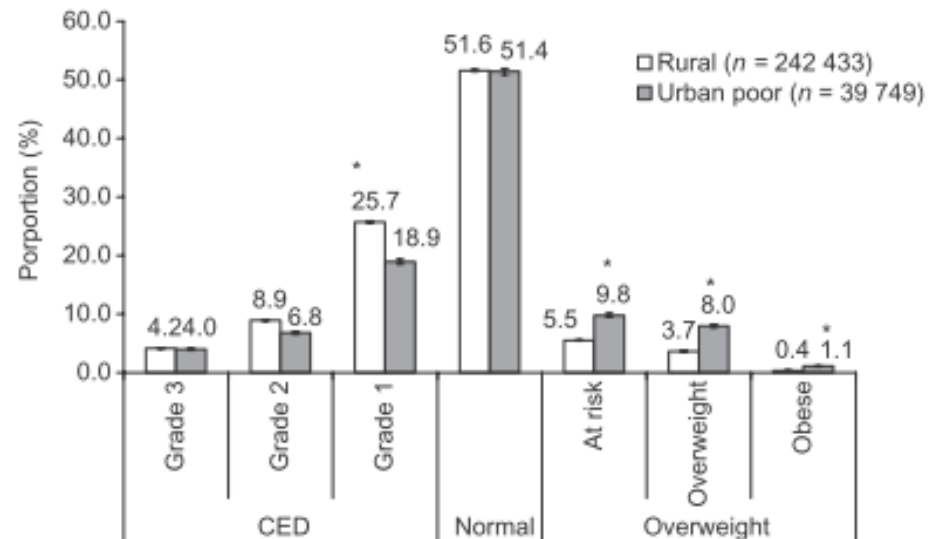
5<sup>th</sup> obs = 23.5

6<sup>th</sup> obs = 24.1

$$\text{Med} = (23.5 + 24.1) / 2 = 23.8$$

# The Mode

- **Mode** is data value or category occurring most frequently.



- Insensitivity to outliers
- Particularly useful for qualitative data and for discrete data – when fractional values don't make sense  
e.g. favourite football team
- There may be more than 1 mode.

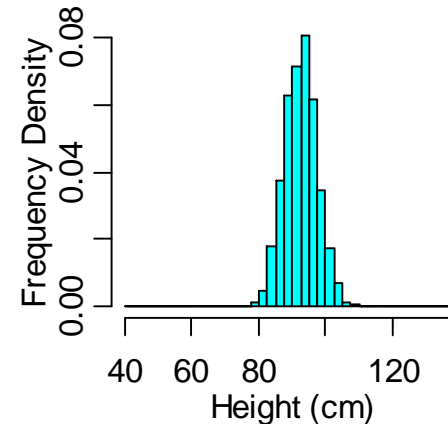
# Variability

- People differ in their characteristics (physical activity, socio-economic status, genetic predisposition, etc...)
  - ➔ BMI will **vary** between people with the same diet
- We want to quantify whether the difference in BMI is due to fizzy drinks, other factors, or a combination of the two?
- Sources “not of interest” are treated as **random variability / dispersion**
  - ➔ Must take random variability into account when drawing conclusions.

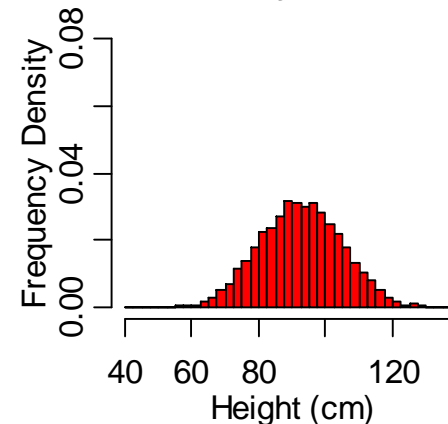
# Measures of Dispersion

- Entirely possible for two data sets to have same mean, median and mode but look very different...
- Measures of dispersion** aim to characterise degree of spread or variability within data set.

**Boys Height  
2-3 years**



**Boys Height  
1-5 years**



# Quantifying variability in data

**Quartiles:** divide the distribution in 4 equal parts.

❖ 1<sup>st</sup> quartile = 25th percentile, 2<sup>nd</sup> = 50th percentile (median), 3<sup>rd</sup> = 75th percentile

Diet 2: high consumption  
of fruit/veg, n=12

22.1	20.5	18.5	16.9	24.1	26.0
22.2	28.4	21.5	31.9	23.0	17.2

Order the data    16.9 17.2 18.5 20.5 21.5 22.1 22.2 23.0 24.1 26.0 28.4 31.9

1<sup>st</sup> quartile: 3<sup>rd</sup> obs = 18.5

2<sup>nd</sup> quartile (Med): 6<sup>th</sup> and 7<sup>th</sup> obs =  $(22.1 + 22.2) / 2 = 22.15$

3<sup>rd</sup> quartile: 9<sup>th</sup> obs = 24.1

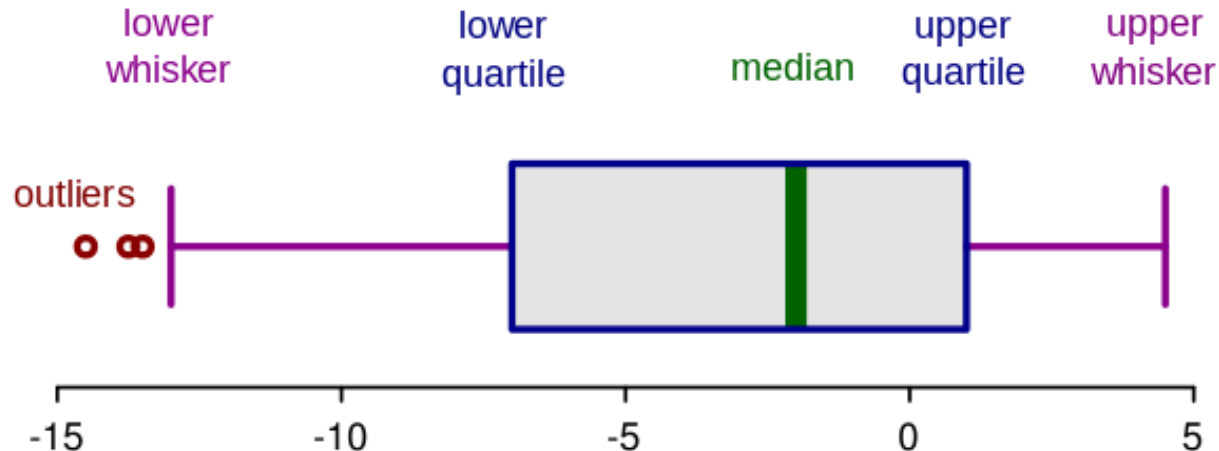
# Inter Quartile Range - IQR

- **Inter-quartile range:** (3rd – 1st quartile) 25th to 75th percentile.

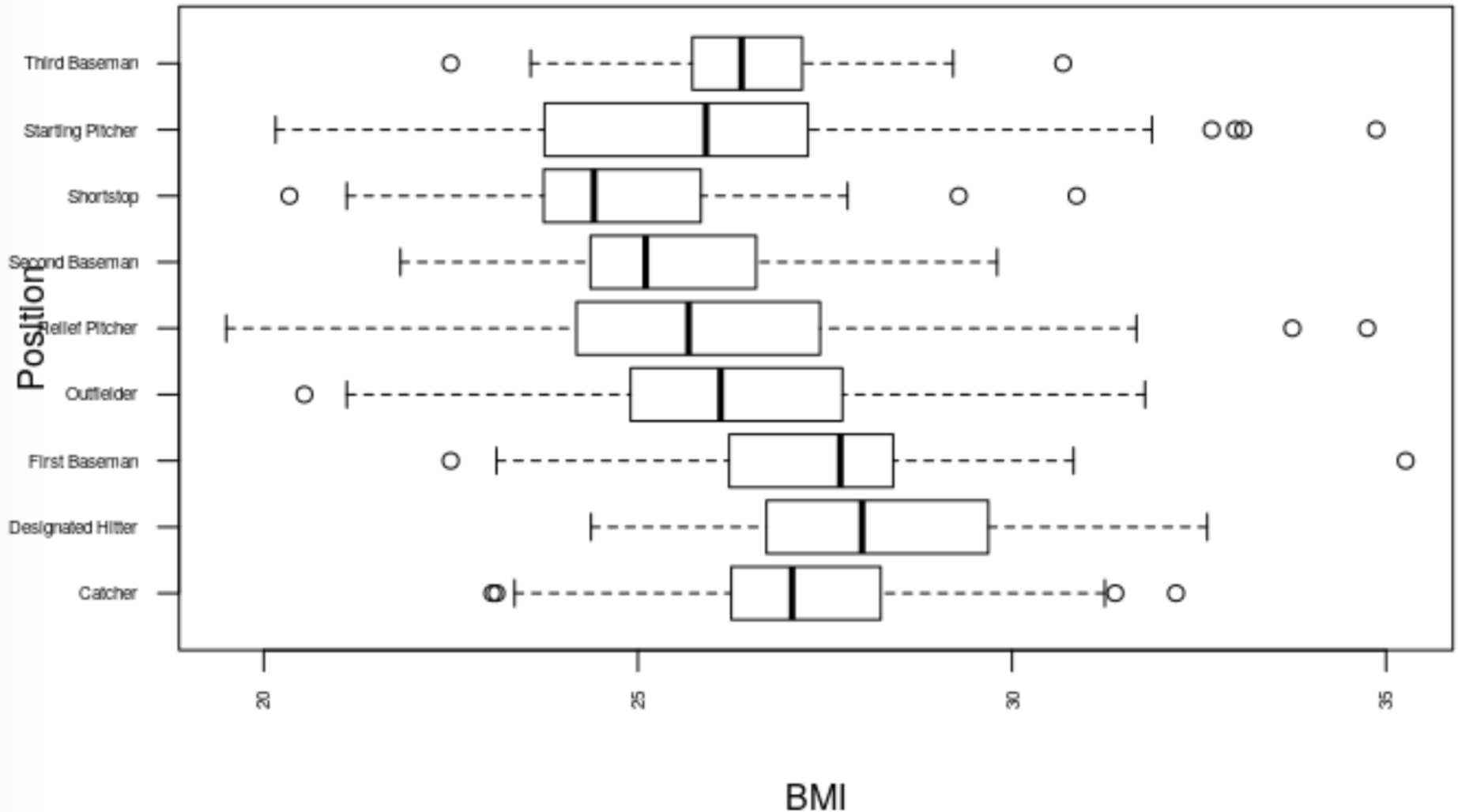
Not sensitive to outliers, but only uses central 50% of observations

# Box (and Whisker) plots

- Graphical summary of the distribution of a **quantitative** variable
- Median** (50th percentile) is marked as central line
- Box represents the **quartiles** (25th and 75th percentiles)
- Whiskers mark the max. and min. values  $< 1.5 \times \text{IQR}$  from the box
- Outliers** ( $> 1.5 \times \text{IQR}$  from the box) marked as separate lines/points



# Box (and Whisker) plots





# The Variance

- **Variance** quantifies amount of variability or spread around the sample mean.
- The variance of  $n$  observations,  $x_1, \dots, x_n$ , is defined as

$$sd^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

# Example

You are investigating the Body Mass Index for people with different diet

Diet 1: low consumption of fruit/veg, n=10

24.1	23.5	18.5	16.7	26.3	28.5
25.2	23.4	22.5	29.9		

$$\bar{x} = 23.9$$

$$sd^2 = \frac{\left[ (24.1-23.9)^2 + (23.5-23.9)^2 + (18.5-23.9)^2 + (16.7-23.9)^2 + (26.3-23.9)^2 + (28.5-23.9)^2 + (25.2-23.9)^2 + (23.4-23.9)^2 + (22.5-23.9)^2 + (29.9-23.9)^2 \right]}{(10-1)} = 16.45$$

# Standard Deviation (SD)

- Rather than use variance, more common to report its square root or standard deviation as this is in the **same units as observations**, i.e.

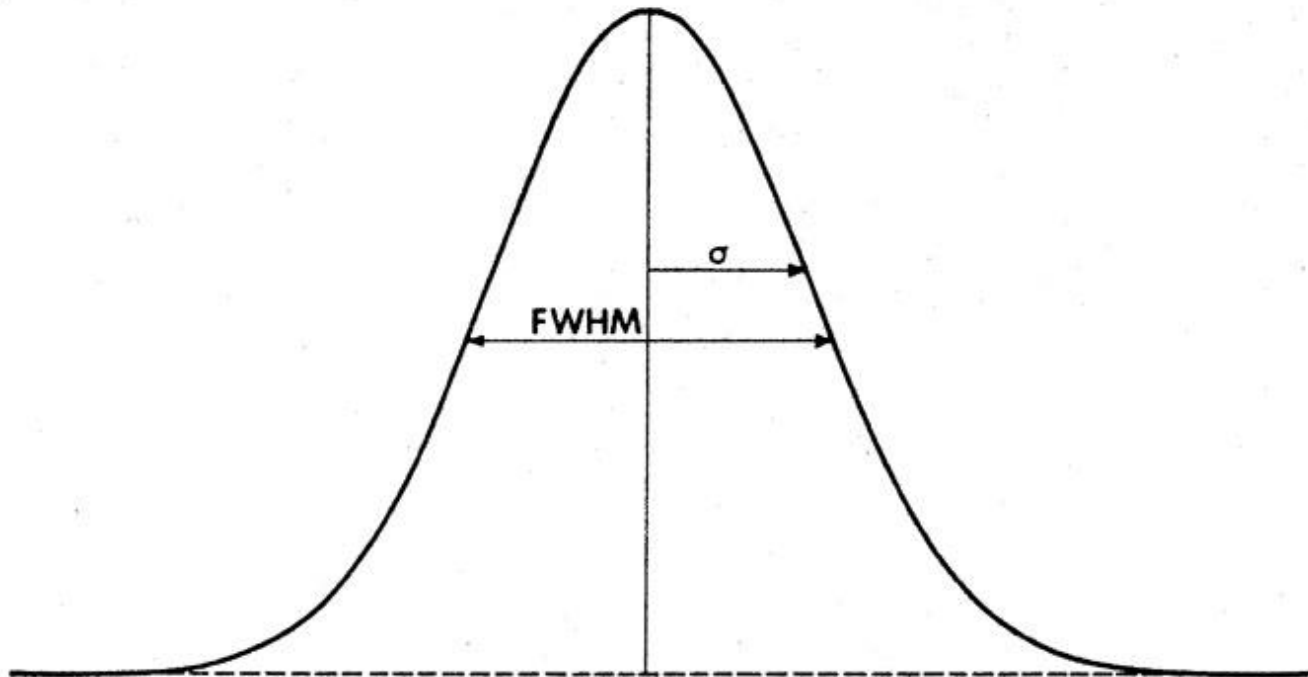
$$sd = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- From the previous example  $sd = \sqrt{16.45} = 4.05$
- Standard deviation is very sensitive to outliers

# Other Measures Exist...

Full-Width at Half Maximum – the width of a “normal” distribution (more on normal distributions in very soon!)

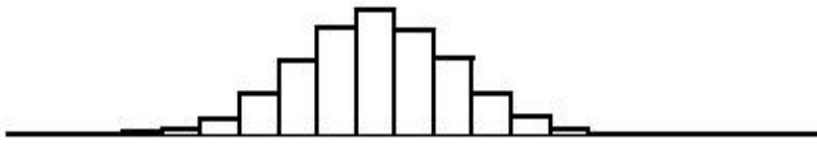
**2.4  $\sigma = \text{FWHM}$**  (for a Normal distribution)



# Shapes of a distribution

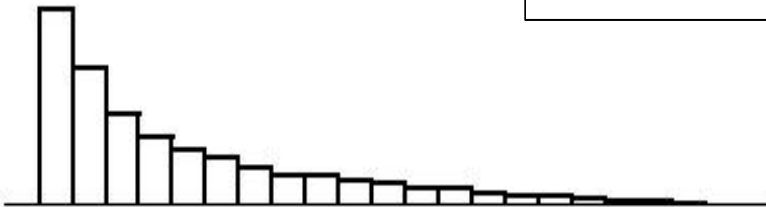
Symmetric

Mean  $\approx$  Median  $\approx$  Mode



Positive Skew

Mode < Median < Mean



Negative Skew

Mean < Median < Mode



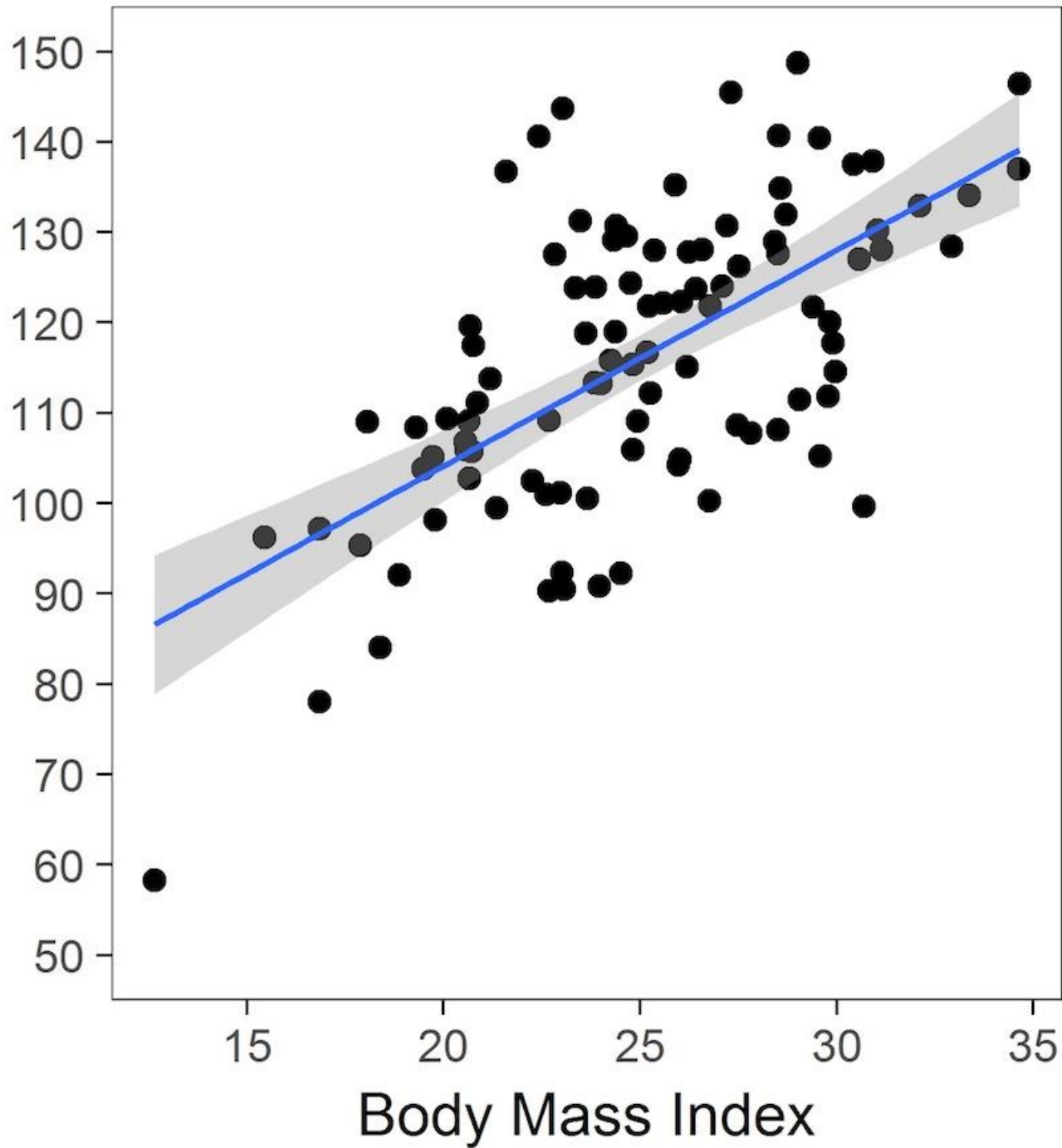
- **Skewness** measures symmetry around the mean
- Positively skewed  $\rightarrow$  long right tail.
- Negatively skewed  $\rightarrow$  long left tail.
- Symmetric  $\rightarrow$  equal tails
- If skewed, median and IQR are preferable

# Correlation of two variables

If a relationship exists between two variables (e.g. BMI and blood pressure), we say that they have a **correlation, or dependence.**

We are often interested in establishing whether a linear relationship exists between two things.

Systolic Blood Pressure



# Pearson Correlation Coefficient

- One method of calculating the correlation is the **Pearson Correlation Coefficient**

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n - 1)s_x s_y}$$

- Here,  $s_x$  and  $s_y$  are the standard deviations of  $x$  and  $y$  respectively
- Any value greatly above 0.5 is considered “correlated”.



# Example

We start with a set of two variables, (x , y)

$$\bar{x} = 7.25$$

(10, 4) (8,8) (2,7) (9,5) (9,4) (7,8)

$$\bar{y} = 6.0$$

$$r = \frac{1}{4 - 1} \left[ \frac{(10-7.25) + (8-7.25) + (2-7.25) + (9-7.25) + (9-7.25) + (7-7.25)}{(9.688) \times (2.5)} \right] = -0.5488$$

This would be considered “weak negative correlation”.  
Therefore, there is likely a weak relationship between the variables

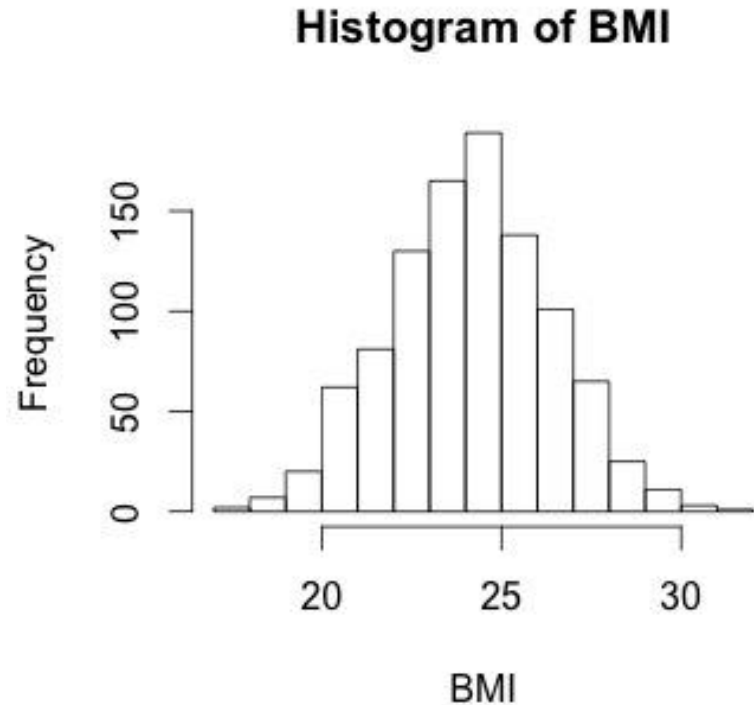
# Descriptive stats - summary

- **Describe the data (tables and plots)**
- **Summarise the data (Mean, Median, Mode)**
- **Measure the variability in the data (inter-quartile range, variance, SD, FWHM)**
- **Quantify the strength of associations between variables (Pearson correlation coefficient)**

- 4. The Normal distribution

# Distributions

- The larger the sample size, the smoother we expect the histogram to become.
- We could use a smooth curve through a histogram to represent the **probability distribution** of this variable for the population studied.
- This is a kind of statistical model.



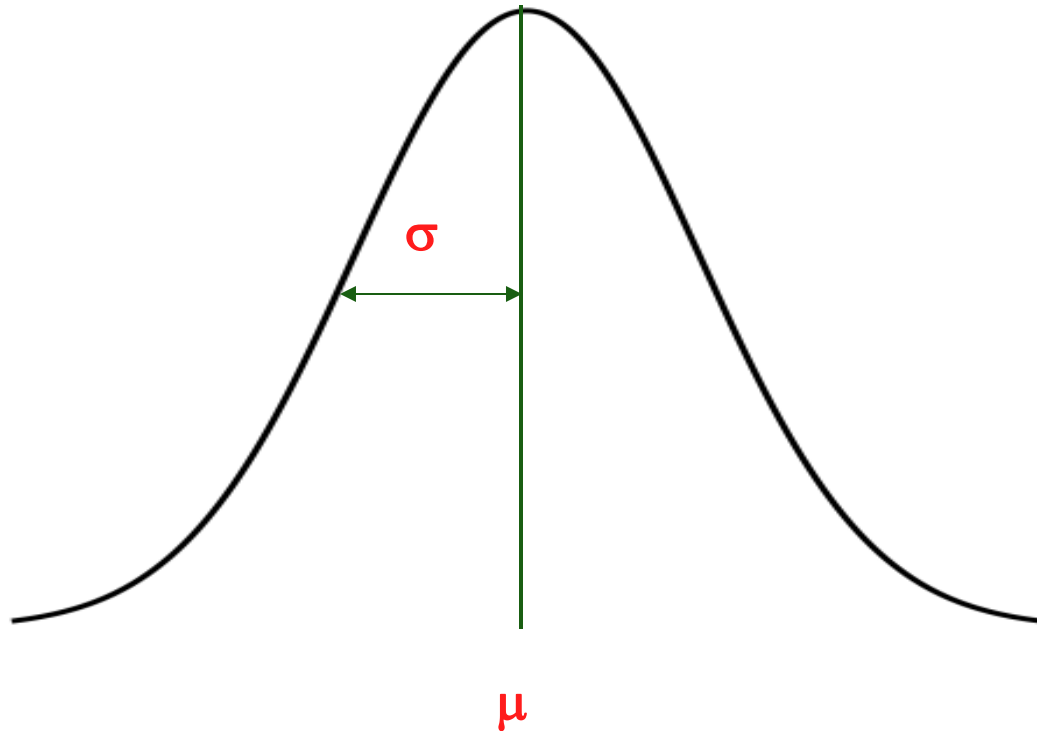
# Distributions

- The equation of the smooth curve representing the histogram of BMI would be a convenient mathematical shorthand for representing probability distribution of BMI.
- In general for continuous variables, the distribution can be a complex equation.
- But, it turns out that the distribution of many continuous variables can be approximated by a standard distribution with known properties:

**Normal (Gaussian) Distribution**

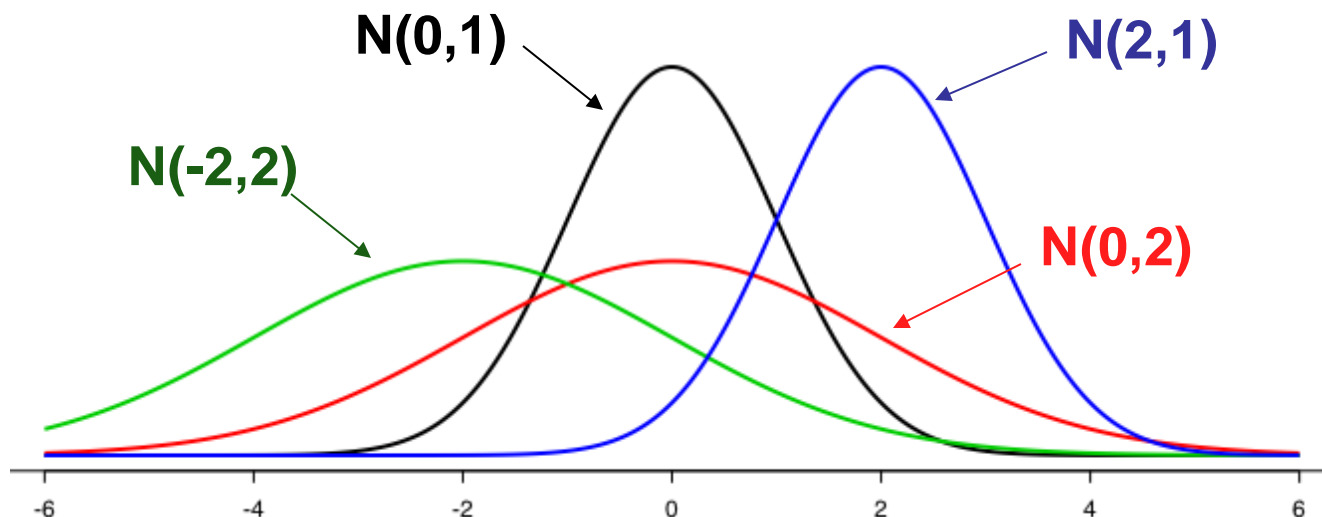
# The Normal Distribution

- The Normal distribution has the shape of a “bell curve” with parameters  $\mu$  and  $\sigma^2$  that determine the center and spread:



# Different Normal Distributions

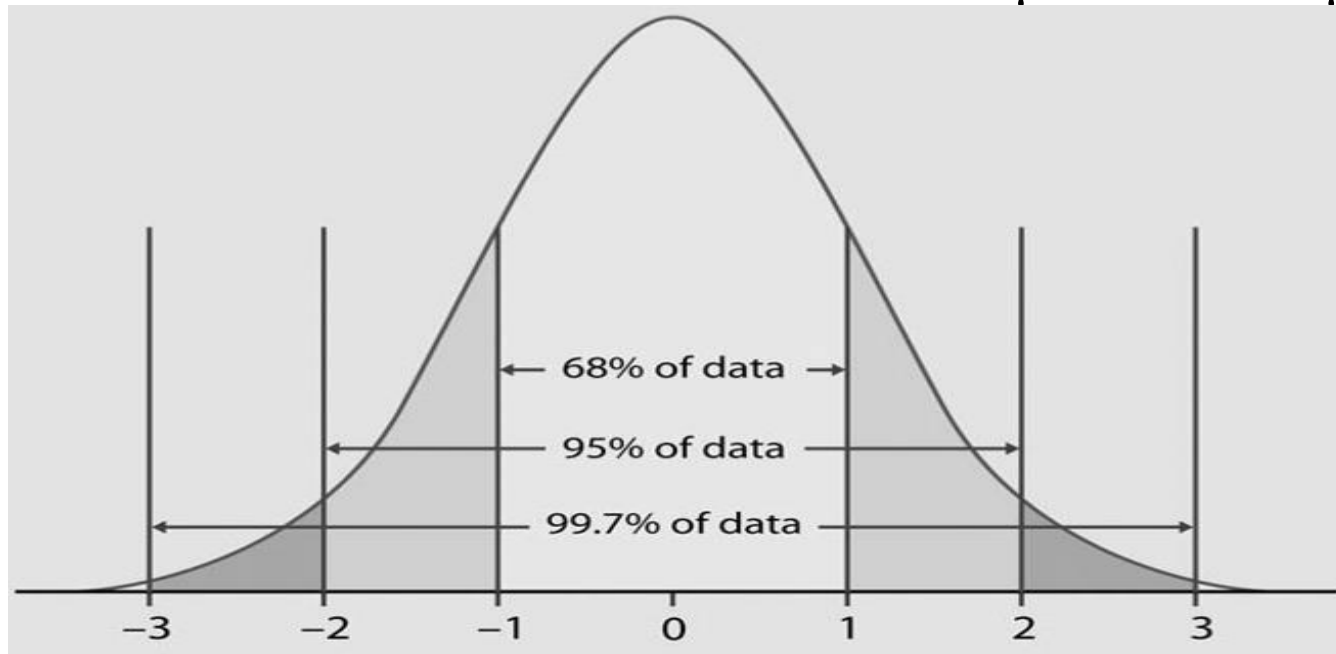
- Each different value of  $\mu$  and  $\sigma^2$  gives a different Normal distribution, denoted  $N(\mu, \sigma^2)$



- If  $\mu = 0$  and  $\sigma^2 = 1$ , we have the **Standard Normal** distribution

# Properties of Normal Distributions

- Normal distribution follows the **68-95-99.7 rule**:
  - 68.27% of observations are between  $\mu - \sigma$  and  $\mu + \sigma$
  - 95.45% of observations are between  $\mu - 2\sigma$  and  $\mu + 2\sigma$
  - 99.73% of observations are between  $\mu - 3\sigma$  and  $\mu + 3\sigma$



$\sigma$

$2\sigma$



# Uses of the Normal Distribution

- Using the Normal distribution as a model for the population behaviour, we can use it to answer questions like
  - Does this population have a larger/smaller mean than another population? (e.g. cases vs controls)
  - Does this population have a larger/smaller variance than another population?
- We use *hypothesis testing* to answer these types of questions and potentially show a statistically significant difference between two data sets.

- 5. Your turn

# Abalone dataset



# Abalone dataset

- Use appropriate plots to present the distributions.
- Summarise the data using whatever statistics you think appropriate.
- Investigate possible dependencies between the variables.

- 5. Statistics in life and research

“Most people use statistics like a drunk man uses a lamppost; more for support than illumination”

— **Andrew Lang**

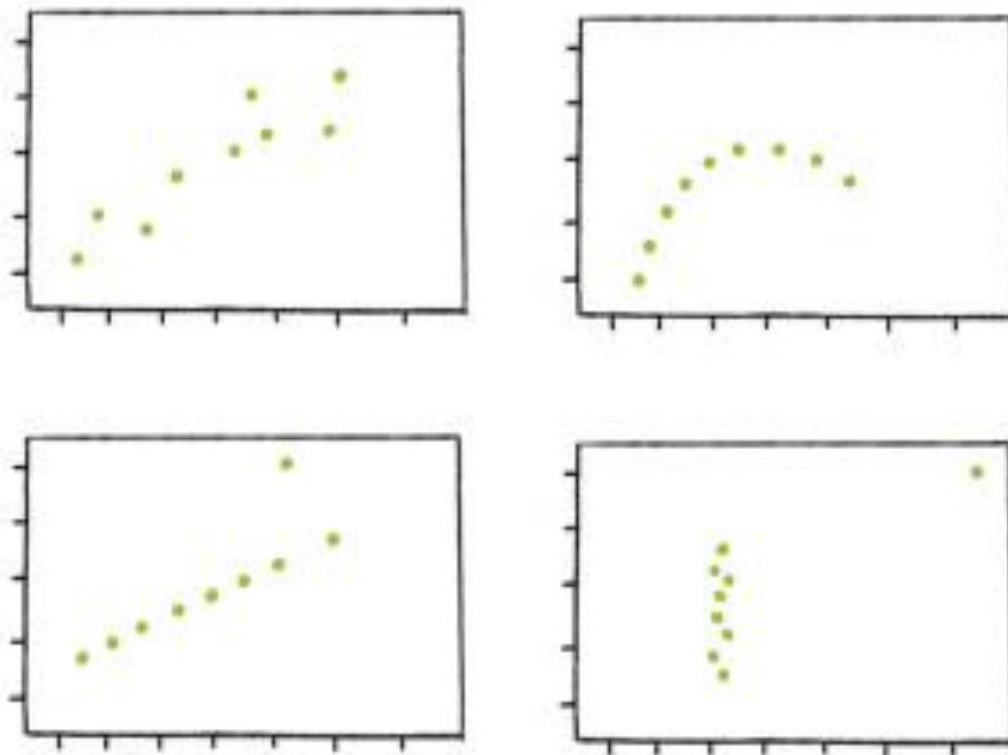
# Data fallacies

Can you think of some examples of how statistics are

misinterpreted in daily life?

deliberately used to mislead people?

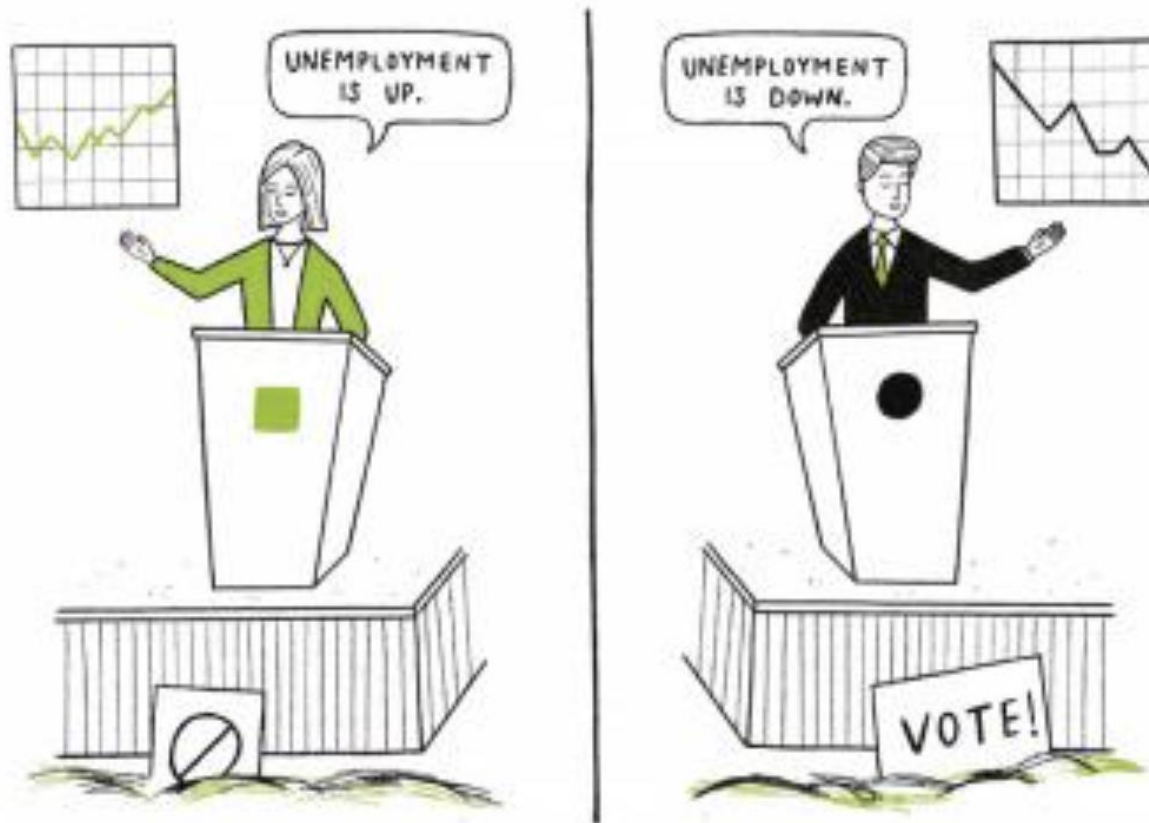
misapplied in scientific research?



## **DANGER OF SUMMARY METRICS**

Only looking at summary metrics and missing big differences in the raw data.





## CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.



## DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.



## **SAMPLING BIAS**

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

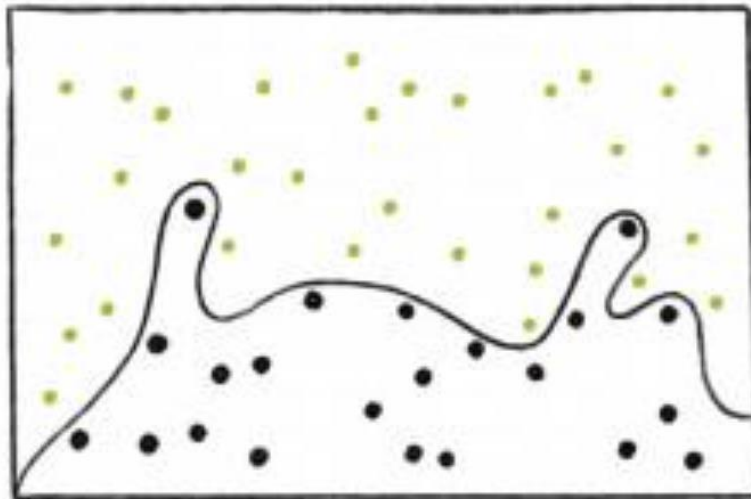


## **SURVIVORSHIP BIAS**

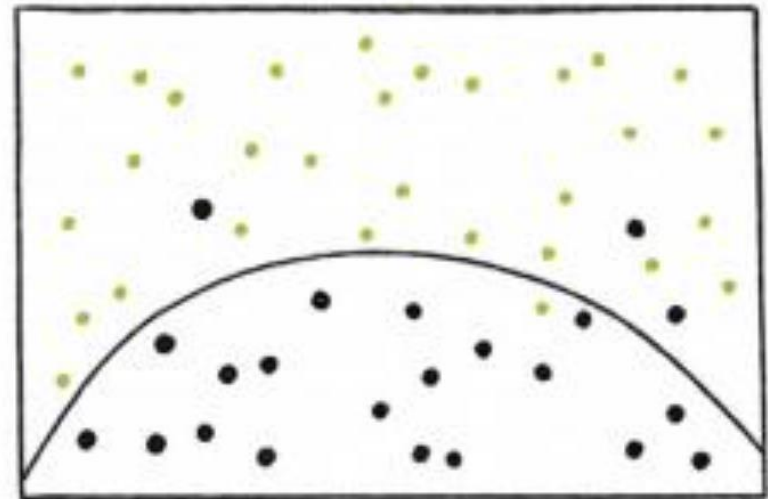
Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.



OVERFITTING



JUST RIGHT



## OVERFITTING

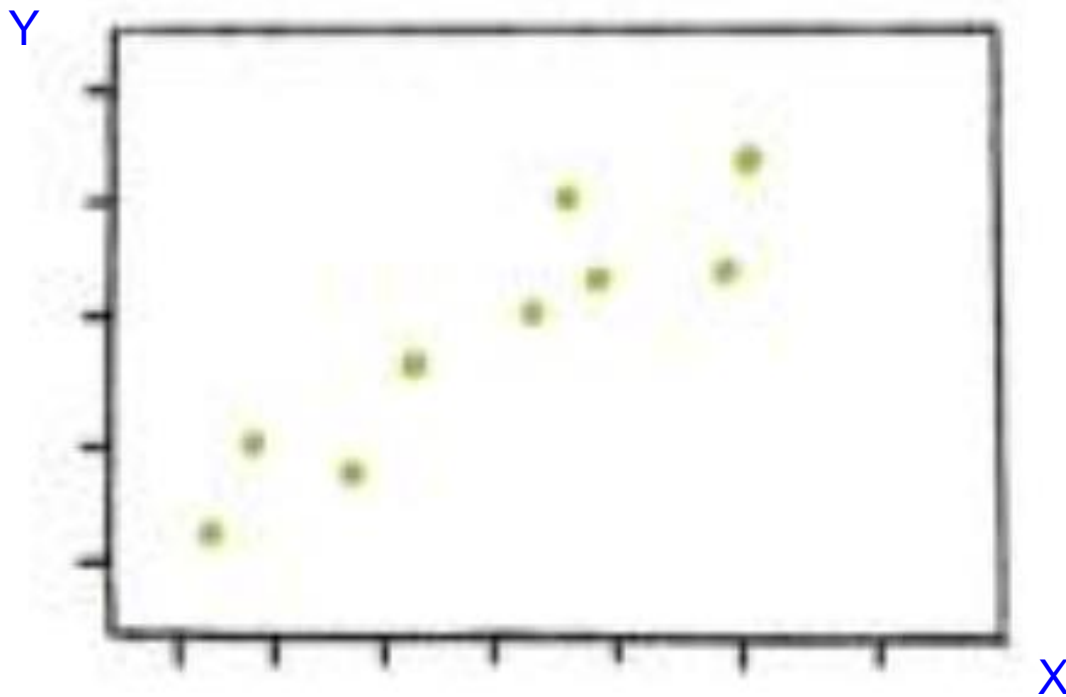
Creating a model that's overly tailored to the data you have and not representative of the general trend.



## **PUBLICATION BIAS**

Interesting research findings are more likely to be published, distorting our impression of reality.

# Correlation vs Causation



- What's the difference between correlation and causation?

# Correlation vs Causation

**Causation** means one event is the direct result of another e.g. when I flick the light switch, it **causes** the light bulb to turn on.

**Correlation** means two events or variables are associated e.g. people who cycle more often are more **likely** to be vegetarian.

- Does correlation imply causation?
- Does causation imply correlation?
- How can we test whether X causes Y?



# Experimental design

- How can we test whether X causes Y?

With a well-designed experiment, it is possible to *control* for causative factors other than those being tested. Positive results then allow us to *infer* causal relationships.

This may require

- Random sampling
- Double blinding (to counter the placebo effect)
- Appropriate choice of sample size

# When does inference go wrong?

- Several factors influence our ability to make inferences from a sample to a population:
  1. The sample may be **biased**, i.e. not representative of the population as a whole.
  2. By **chance** alone the sample may differ, to a greater or lesser extent, from the population.
  3. Using the **wrong** statistical methodology

# 1. Avoiding Bias

- **Kinds of samples:**
  - sample of convenience (possibly biased)
  - random sample (eliminates bias): every person in the population has an equal chance of being in the sample
  - stratified random (eliminates bias across strata)
- **Numerous types of bias to be aware of –**  
selection, dropout, recall ,etc.

## 2. Chance

- We cannot do anything about the second factor (chance), and we must always remember that chance alone could have given our observed finding.

(But if our sample is large enough, we can be confident that our sample mean is a good estimate of the population mean!)

- We must therefore quantify the uncertainty in our estimates of the population parameters.

# 3. Wrong method

- Even if the sample is representative of the target population, using the wrong statistical method to analyse the data leads to invalid statistical inference.
- **Example:** when comparing the mean between two groups, a hypothesis test may give incorrect conclusions if its assumptions don't hold true for the data analysed.

# Moving forward with statistics

Some other Graduate School courses you may find helpful:

- *Data exploration and visualisation*
- *Introduction to sampling and hypothesis testing*
- *Regression modelling*
- *R Programming*

# Further reading

“Points of Significance” series of articles  
from Nature Methods

[http://blogs.nature.com/methagora/2013/08/  
giving\\_statistics\\_the\\_attention\\_it\\_deserves.  
html](http://blogs.nature.com/methagora/2013/08/giving_statistics_the_attention_it_deserves.html)

# Papers used for the examples

1. Shafique S. et al., Trends of under- and overweight among rural and urban poor women indicate the double burden of malnutrition in Bangladesh, *Int.J.Epi.*, 2007; 36:449-457
2. Yadav and Krishnam, Changing patterns of diet, physical activity and obesity among urban, rural and slum populations in north India, 2008; *Obesity Review*, 9:400-408.



# Acknowledgements

These slides are based on an earlier version of the course, developed by Brett Thomas.

*Data Fallacies* examples taken from [data-literacy.geckoboard.com](https://data-literacy.geckoboard.com)