

DSAW6/7075

Presenting and Analysing data

Dr Christine Davies

Aims of Day 3

- Re-cap on: research methods; philosophical background to research
- Looking at different types of data;
- intro to inferential statistics
- Practice in use of Excel for inferential statistics

Module Assessment

Assessment Component 1 – 50%

- Written Assignment – 4,000 words or equivalent
- Critically examine a range of research methods, indicating the ontology, epistemology, and methodology associated with each, and evaluate within a range of contexts.

Assessment Component 2 – 50%

- Written Assignment – 4,000 words or equivalent
- Outline and evaluate the possible method(s) used to generate a given data set, analysing any ethical considerations involved, and any possible issues of insider research. Analyse the data provided, discussing the likely conclusions resulting from your analysis.

Re-cap

Key terms in research (and two parallel 'strands')

Ontology	objectivism	subjectivism
Epistemology	positivism	Interpretivism
Research methodology	quantitative	qualitative

Some Research methods (collecting data)

Qualitative

- Interview
- Dialogue /Narrative
- Focus group
- Observation
- Video, audio recording
- Case study
- Action research
- Diary/ log

Quantitative

- Questionnaire/ survey
- Structured interview
- Experiment
- Content analysis

Data types

Data (information) from Research

- **Primary data:** data newly-generated by research
- **Secondary data:** data generated by past research

Key types of data

Data type	description
'Scale', or 'Interval/ratio' data	Numerical data with clear, fixed intervals (eg. 1,2,3).
Ordinal data	Data which can be rank ordered, but with differences that are not clear/fixed (eg. once a day; agree strongly)
Nominal data	Data which describes categories (categorical/ qualitative data) (eg. red, green, blue).

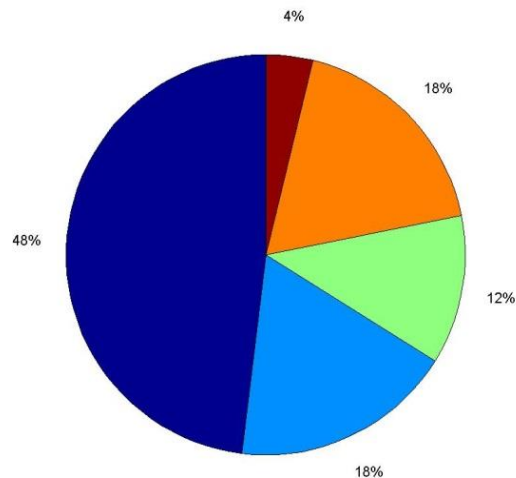
Presenting and analysing scale data: **descriptive statistics**

Recording data: Frequency tables

Job	Frequency	Cumulative Frequency
admin	546	546
engineer	647	1193
production	269	1462
management	135	1597

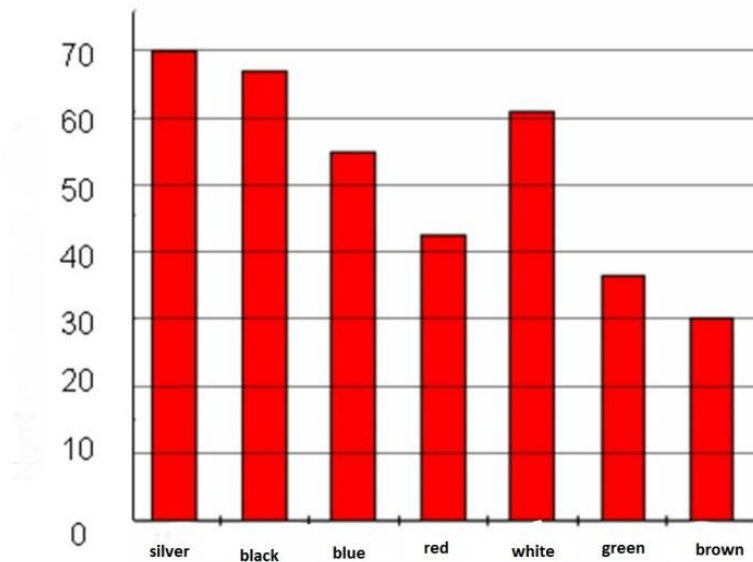
These are '**semi-quantitative**' data because one **variable** is *quantitative* (frequency), and the other (jobs) is *qualitative* (this can also be described as '**nominal**' or '**categorical**')

Representing semi-quantitative data



This is a **pie-chart**. Data is organised into categories (eg. different jobs: qualitative), with each assigned a 'piece of pie' whose size is proportional to its frequency. The whole pie (1, or 100%) corresponds to the total frequency (1597 in the previous example)

Representing semi-quantitative data (2)



- This is a vertical **bar-chart**, with *qualitative* information on the **horizontal (x) axis**, and *quantitative* information on the **vertical (y) axis**

averages, or 'central' tendency

This means a measure of middle/average values

Mean: sum of all values divided by frequency

Mode: most common value

Median: value at the 50% frequency point

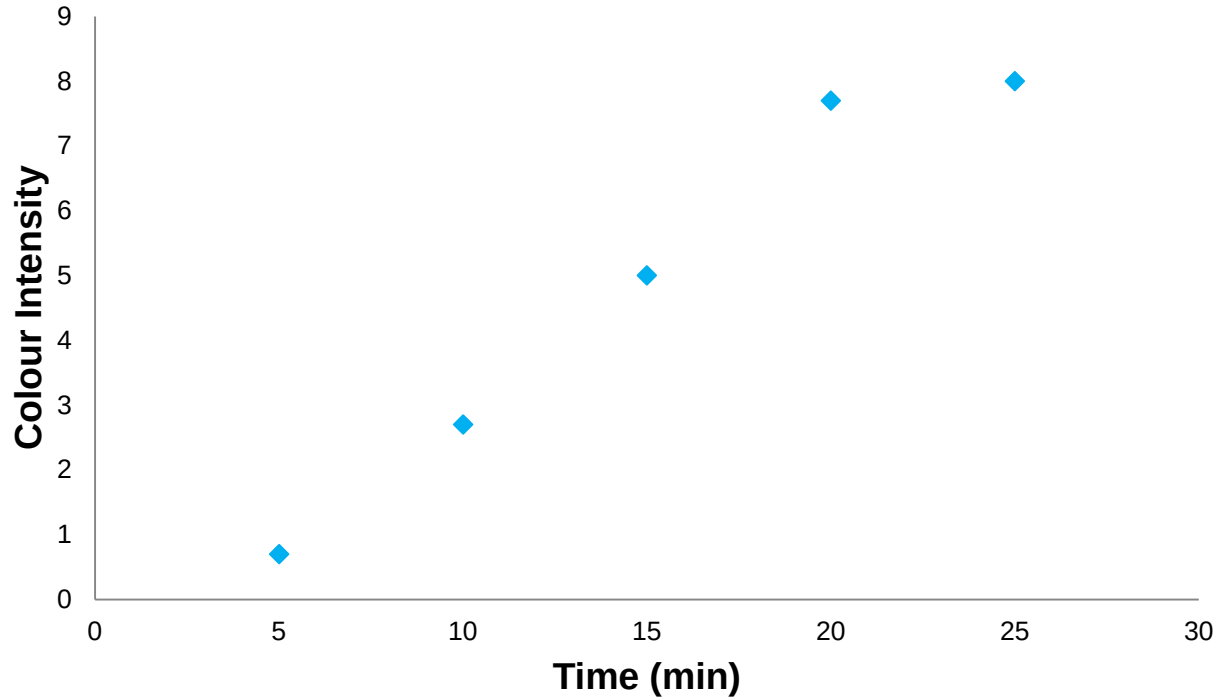
Results Table

Time (min)	Colour intensity (0-10 scale)			
	Trial 1	Trial 2	Trial 3	Average
5	1	0	1	
10	3	3	2	
15	4	5	6	
20	7	8	8	
25	8	8	8	

Summary Table showing mean values

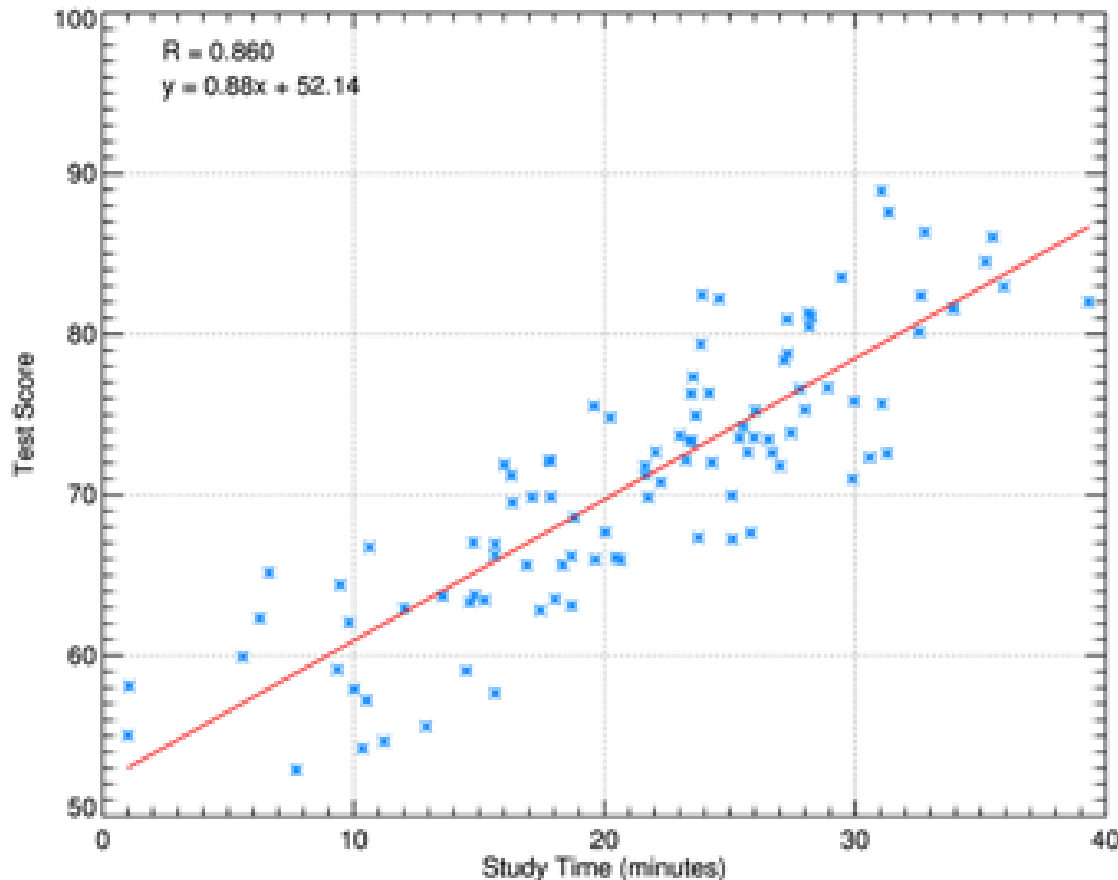
Time (min)	Mean Colour intensity (0-10 scale)
5	0.7
10	2.7
15	5.0
20	7.7
25	8.0

Scatter Plot to show relationship between process time and colour intensity

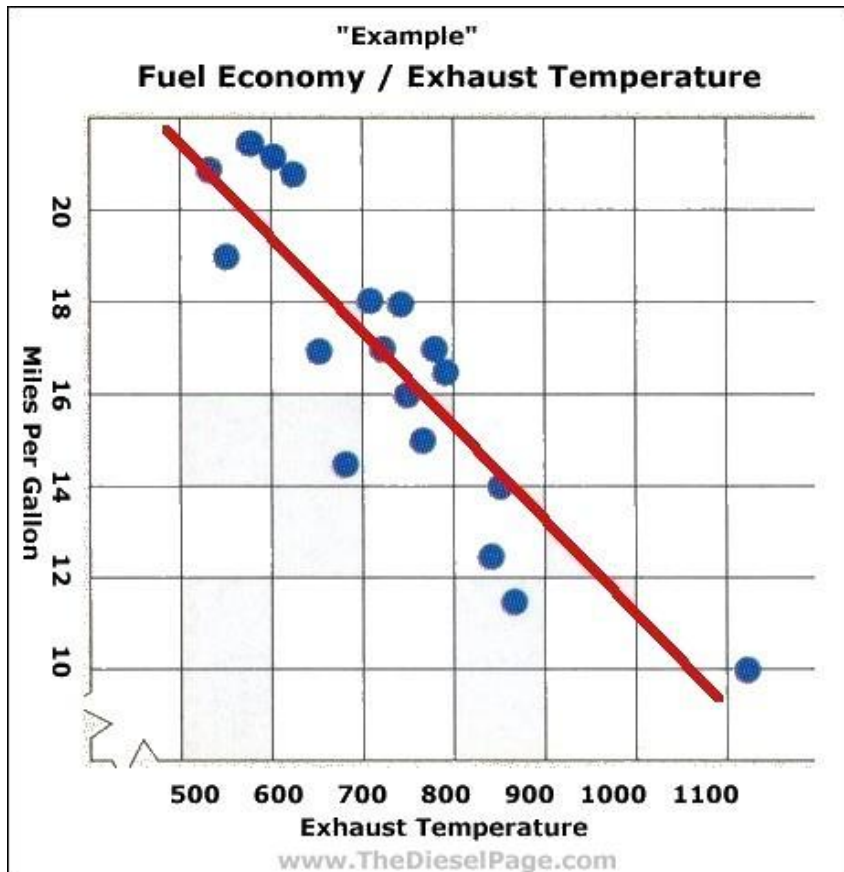


- The scatter plot shows the relationship between two quantitative variables
- Time is the **'independent' variable**; colour intensity is the **'dependent' variable**

Another example of a scatter plot



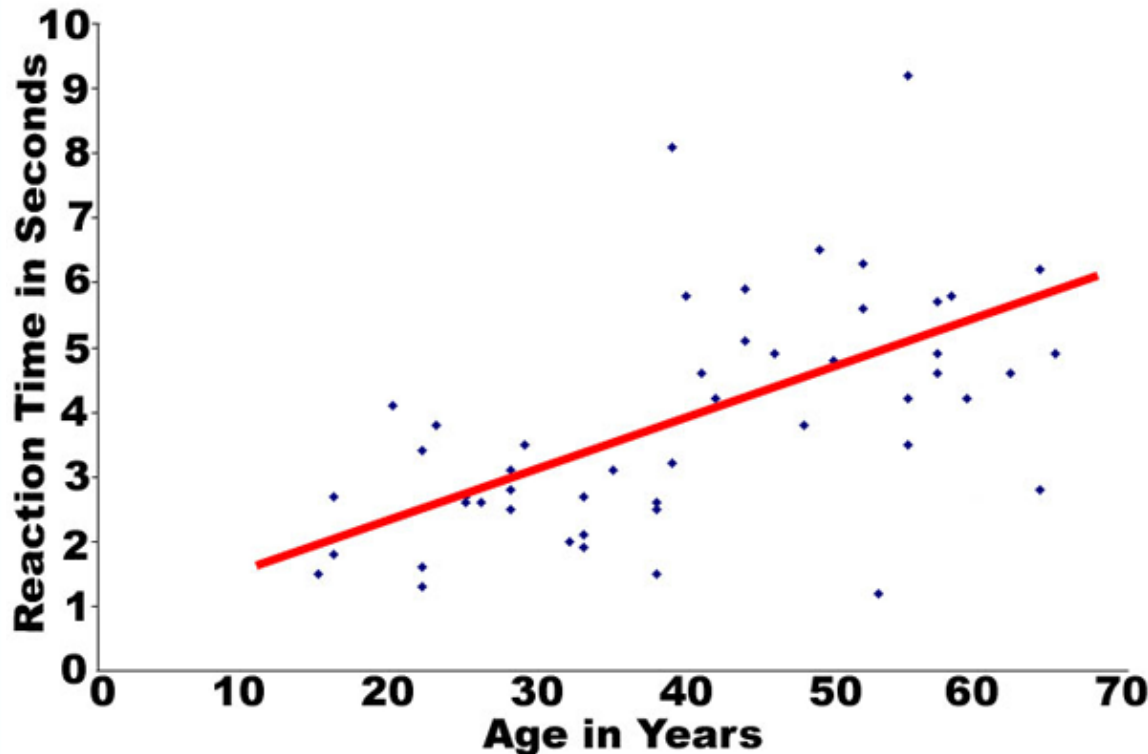
- This plot also shows a 'line of best fit' which shows the relationship more clearly
- This diagram indicates a **positive association** between the variables



- Scatter plot showing a negative association between variables

Another scatter plot example

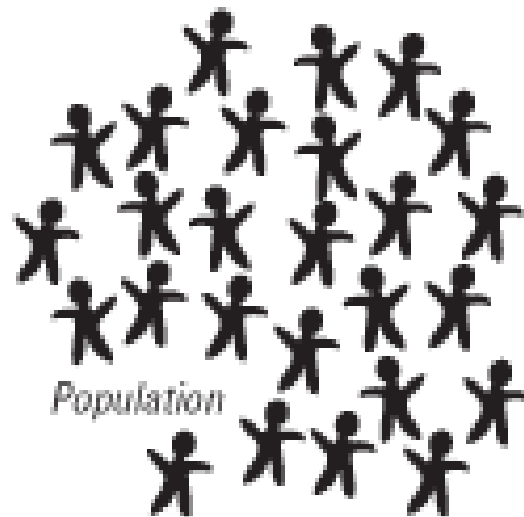
Age and Reaction Time



- What do the data here show?

Drawing conclusions from quantitative data: **inferential statistics**

We want to know about these



Population



Parameter

μ

(Population mean)

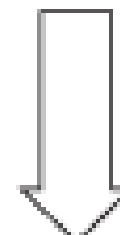
*Random
selection*



We have these to work with



Sample



Inference

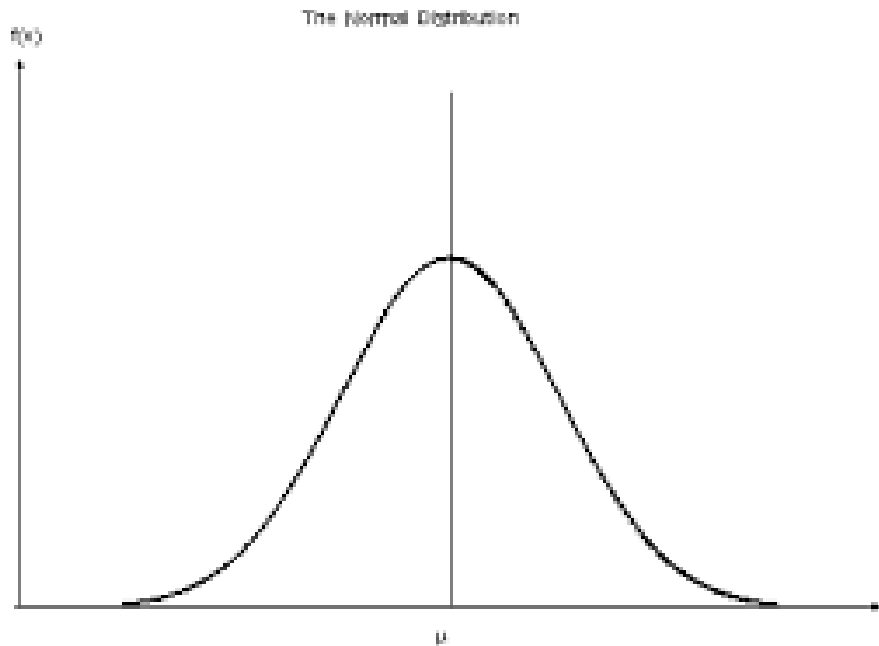
$\bar{x}$

Statistic

(Sample mean)

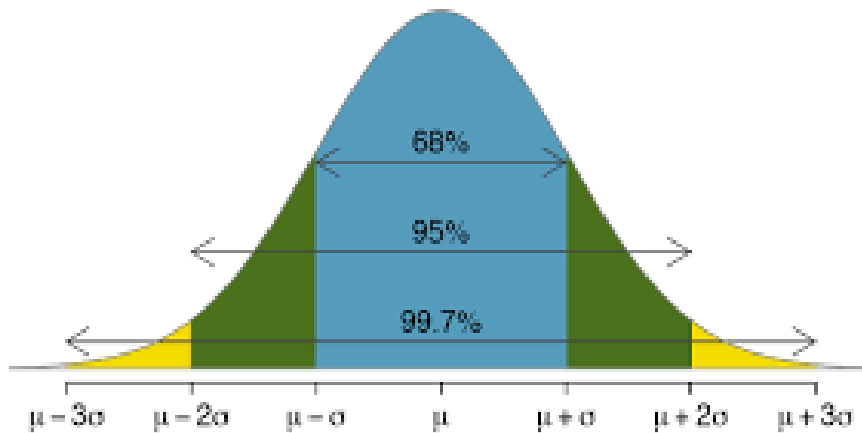
The normal distribution, and how it leads to 'p values'

The normal distribution



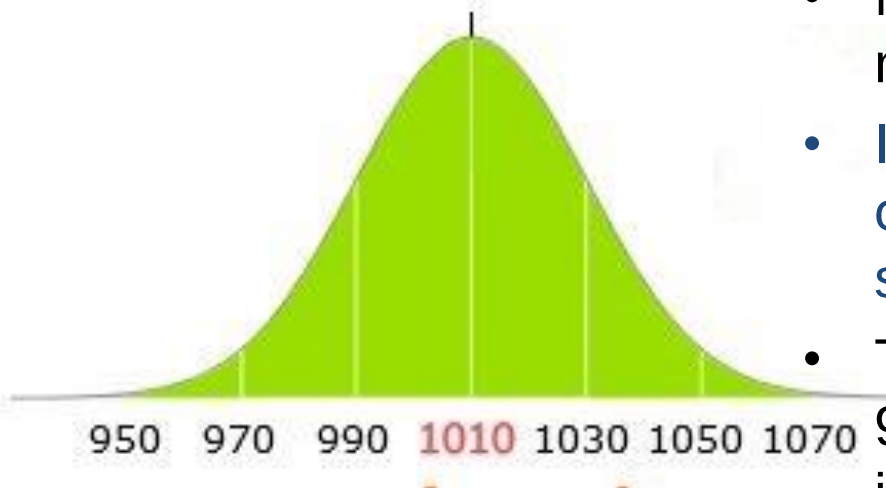
- This ‘bell-shaped’ curve shows the population of a particular value (here, termed ‘x’), eg. height of children in a school; weight of patients in a hospital
- Most values (of x) are clustered around the central value
- The central value is the population mean, μ

Normal distribution, & standard deviation



95% of all values are found within approx $\pm 2\sigma$ (actually, 1.96σ)

Example:



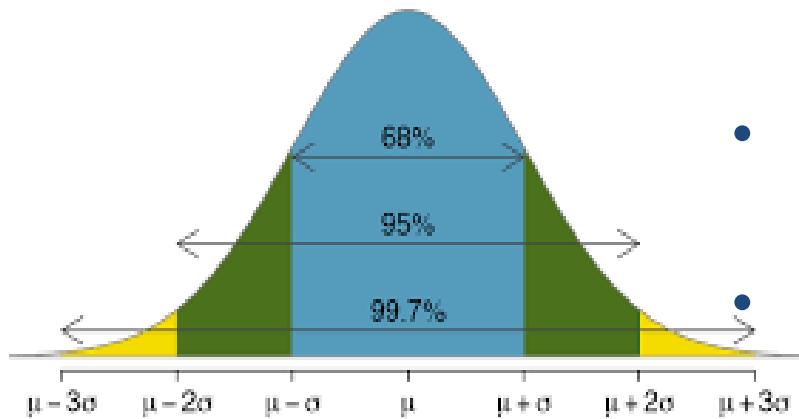
- This represents a population with a normal distribution (eg. height of children aged 6-10 in cm in a particular town)
- The population mean, $\mu = 1010$
- Most heights are clustered around the mean
- If we took many random samples of children's heights, we would find 95% of sample means between 970 and 1050
- The probability of a very low value (eg. 950) belonging to this population is low: it is less than 5%
- We might conclude that a child with height 950cm (or sample with this mean) is significantly different from this population

We can use this idea to infer whether samples are likely to come from particular populations



- You have bought a pack of Welsh-cakes that looks extremely ‘spotty’, ie. with a high raisin content 😊
- You form a hypothesis that it has a significantly different raisin content compared to a typical pack
- How do you prove or disprove your hypothesis?

Example



- You find the mean raisin content of your Welsh-cakes (number per cake, divided by no. of cakes)
- You compare your mean with a typical raisin distribution
- If it falls within the 'extreme' ends (ie. 5% at the edges), we would say that the probability of belonging to this population is <0.05 (ie. Probability, **$p < 0.05$**)
- This means that it is **significantly different**: in this case, the raisin content is significantly higher

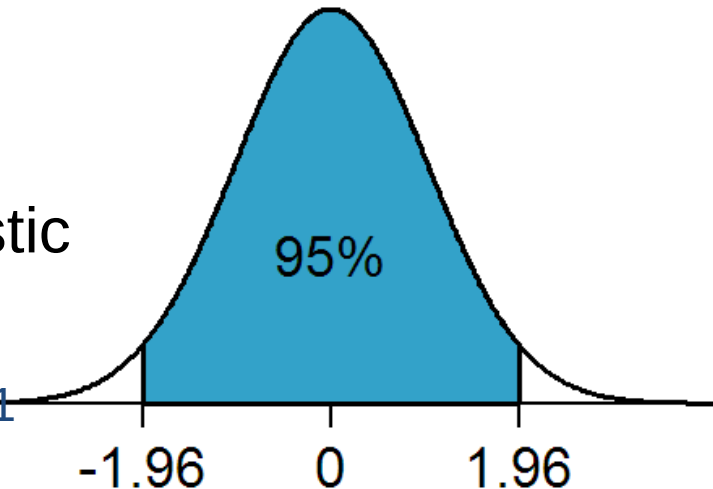
Setting out Hypotheses

Typically, we use the following convention:

- **Null hypothesis** (H_0): there is no significant difference between individuals/groups, and any observed difference is due to chance
- **Alternate hypothesis** (H_1): there is a significant difference

Statistical tests and p values

- Statistical tests do not give definite answers, just probabilities, or '**p values**'
- We decide if these probabilities are significant
- If we find a p value that is *smaller than* 0.05, this implies that the sample statistic is **significant** at this level
- We would then reject H_0 and accept H_1
- Eg. in the Welsh-cake example, we would deduce that our pack ('sample') is significantly different from typical Welsh-cakes ('population')

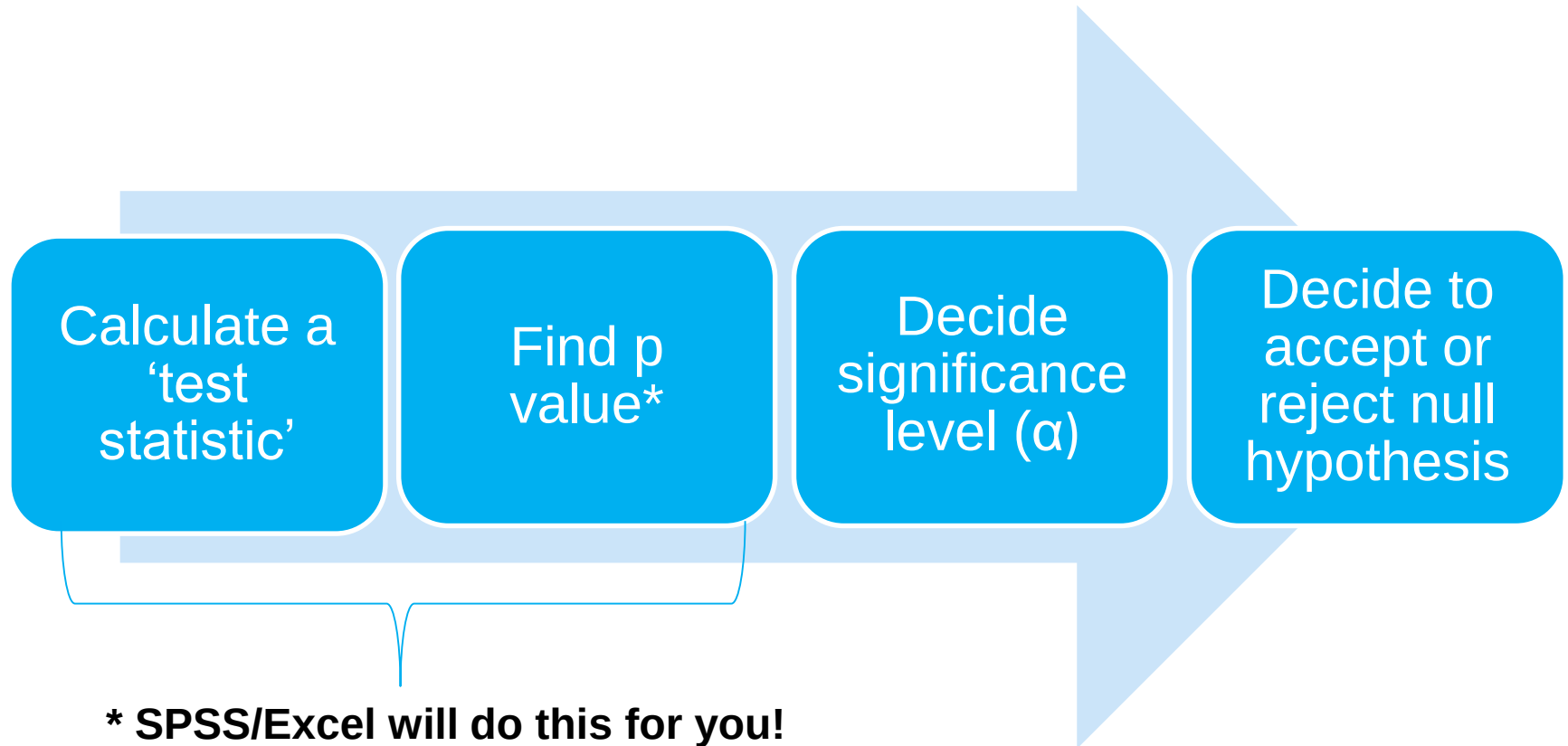


key statistical tests

Two key types of statistical tests

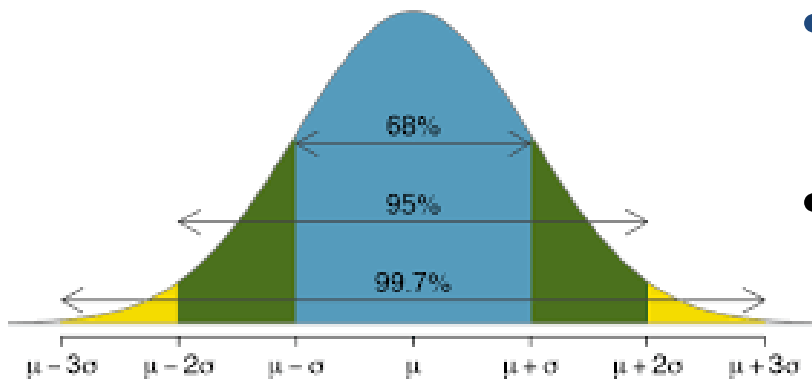
- 'testing' Differences:
- T tests, Chi squared (χ^2), ANOVA
- 'testing' Links/Relationships/ Associations:
- correlation and regression; factor analysis

The process of carrying out a statistical test:



What is a significance level (α)?

- Typically, this is set at 5% (ie. $p < 0.05$; outside the middle 95%)
- This means that we accept H_1 if $p < 0.05$
- If we want to be really sure of a significant difference, we use a 1% significance level, where $p < 0.001$ (outside the middle 99%)



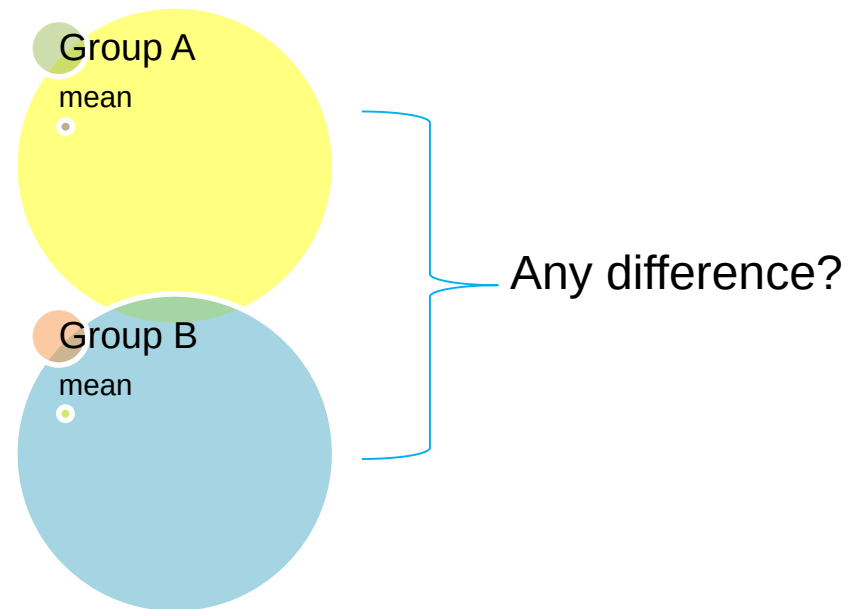
Re-cap video:

https://www.youtube.com/watch?v=e_yknGvncKLw

Inferential Statistics 1: Looking for differences

The T test ('Student's T test')

This is a Statistical test that helps to show if there is a difference between means of **2** groups



Independent (or Standard) T test:

- This test compares means from *two different, or independent, groups*
- Can you suggest an example?
- Eg. heights of oak versus ash saplings of the same age, soil, etc
- What is the research context of this type of statistical test?
- Often, an experimental, or quasi-experimental design. Important to keep possible confounding factors the same between groups



Dependent (or Paired/ repeat) T test

This T test compares means in before/after (ie. 2 groups) situations in the **same individuals**

- Can you suggest an example?
- Eg. heart rate in people before and after drug therapy
- What is the research context of this type of statistical test?
- Often, an experimental, or quasi-experimental design *in the same individuals*. Important to keep possible confounding factors the same between groups



Analysis of Variance (ANOVA)

- This allows the means of **more than 2 groups** to be compared



Mean A



Mean B



Mean C



Mean D

Analysis of Variance (ANOVA)

- This allows the means of **>2 groups** to be compared
- Can you suggest an example?
- Eg. heights of oak, ash, and birch saplings of the same age, soil, etc
- What is the research context of this type of statistical test?
- Often, an experimental, or quasi-experimental design (confirmatory) with >2 groups/treatments. Important to keep possible confounding factors the same between groups

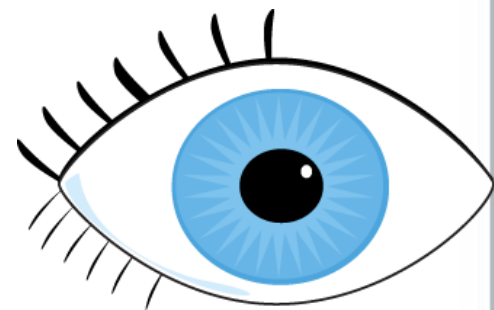
One-way ANOVA

- *>2 groups; one variable* investigated
- Example?
- Eg. height (one variable) of seedlings in three different media (>2 groups)



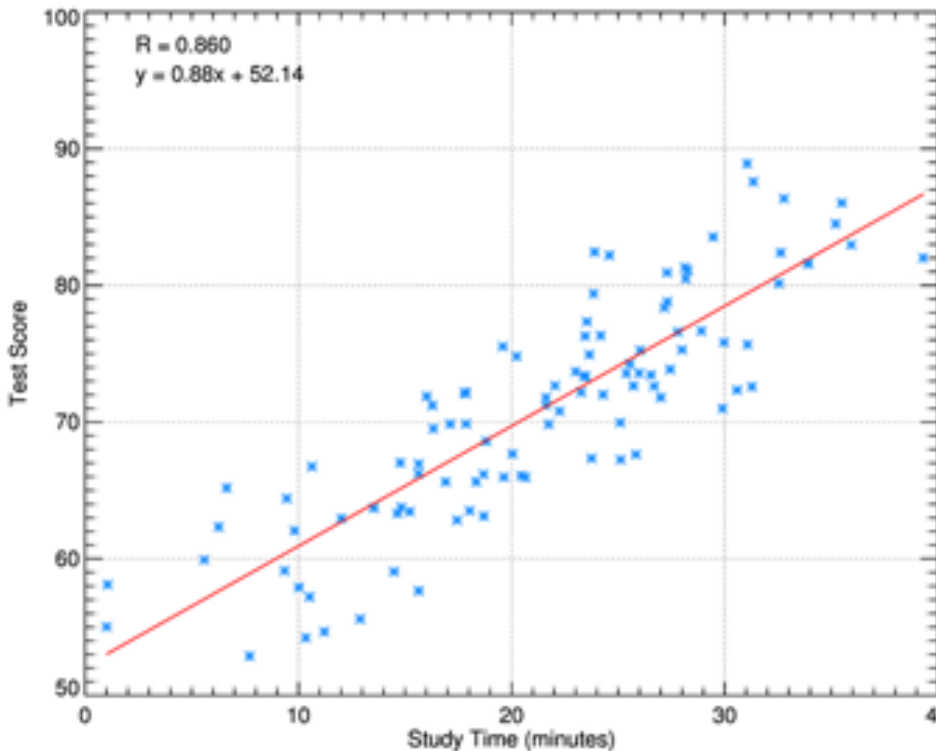
Chi-squared (χ^2) test

- this test compares observed and expected numbers
- Can be used with non-parametric data including nominal data: useful!
- Examples?
- Eg. proportion of people with different eye colours in a town
- Eg2. job types of extroverts and introverts



Inferential Statistics 2: Looking for relationships

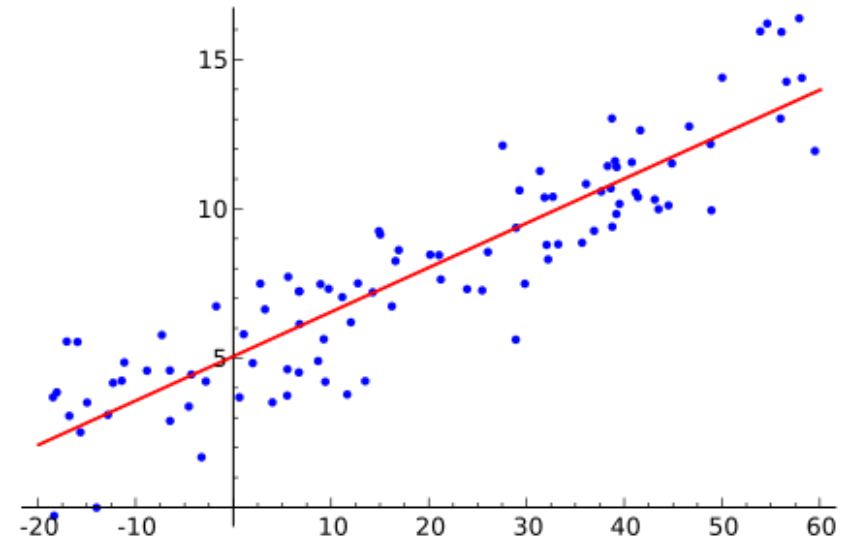
Pearson correlation coefficient



- Two continuous, interval /ratio variables involved
- correlation coefficient, R (-1 to +1) can be computed
- This indicates the closeness of relationship between variables
- This diagram indicates a positive relationship (**correlation**) between the variables
- the '**line of best fit**' (regression line) helps us see the relationship between variables

Regression analysis

- Where two variables are linked in a linear way (correlated), a regression line can be estimated ('line of best fit')
- This is based on an equation that describes the line
- For any value of the independent variable (x), it is then possible to predict the value of the dependent variable (y)



Parametric and non-parametric analysis

Statistical analysis of parametric and non-parametric data should make use of different methods:

Parametric

statistical tests used to analyse data include:

- T tests;
- analysis of variance (ANOVA);
- Pearson correlation

Non-parametric

statistical tests used to analyse data include:

- Mann-Whitney test;
Wilcoxon
- Kruskal-Wallis test;
- Spearman correlation

Now try out the tests!

- Use SPSS
- Or **Excel**