statstutor

statstutor community project

encouraging academics to share statistics support resources

Stcp-marshallowen-7

The Statistics Tutor's Quick Guide to Commonly Used Statistical Tests

www.statstutor.ac.uk

Reviewer: Jean Russell
University of Sheffield

Contents

Introduction

This guide is designed to help you quickly find the information you need about a particular statistical test.

Section 1

Section 1 contains general information about statistics including key definitions and which summary statistics and tests to choose. Use the "Which test should I use?" table to allow the student to choose the test they think is most appropriate, talking them through any assumptions or vocabulary they are unfamiliar with.

Section 2

Section 2 takes you through the most common tests used and those that are usually as complex as the students require. As a statistics tutor, you should be familiar with all these techniques.

Section 3

Section 3 contains tests and techniques that are more complex or are used less frequently. This section is aimed at tutors who have studied statistics in detail before.

Tips for tutoring

We are here to help the students to learn and not to do their work for them or to tell them exactly how to do their work (although this is very tricky sometimes!). Facilitating the understanding and ability to choose an appropriate statistical test is a success, even if the analysis is not as thorough as if we had done it ourselves.

Avoid maths! Most students carrying out project analysis do not need to know the maths behind the technique they are using. You may love maths but a lot of students are maths phobic – even an x can frighten them!

You cannot know everything! Statistics is a vast subject so don't be afraid to say that you don't know. Ask others for help or look up information on the internet to help.

Consider the students ability when advising on the best technique. They have to write up the analysis and therefore need to understand what has been done. Carrying out simple analysis or even just a graph to summarise their results may be enough for their project.

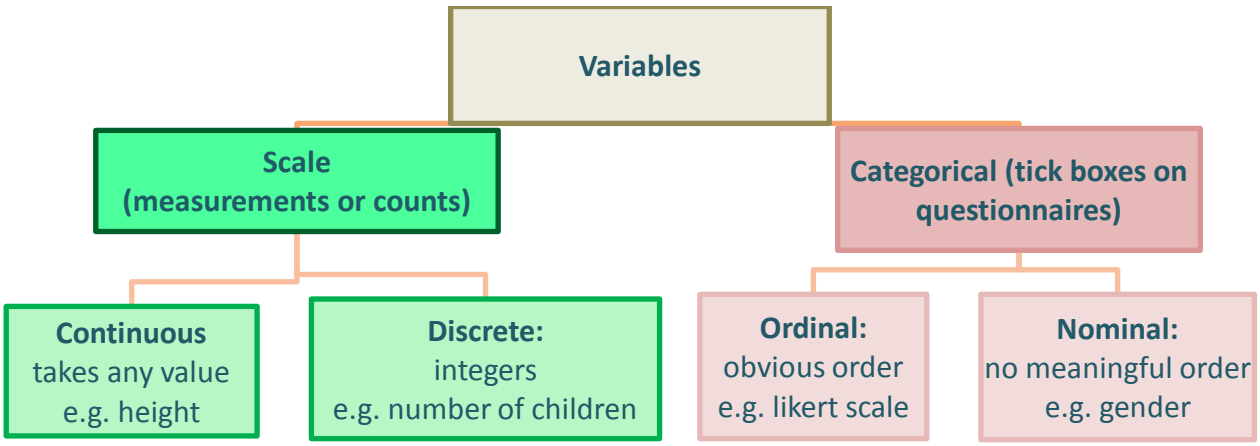
Don't assume that the student knows anything about the technique they are suggesting! Students with no statistical knowledge at all can come in saying their supervisor wants them to carry out multivariate analysis. Get them to explain their project and why they think the technique is suitable if you think that they know very little. In general, start with descriptive statistics and simpler analysis if they have not done any analysis yet. Some students do know exactly why they are doing something and have investigated the topic fully so just help them with the complex technique if you can.

Section 1

General information

Data types

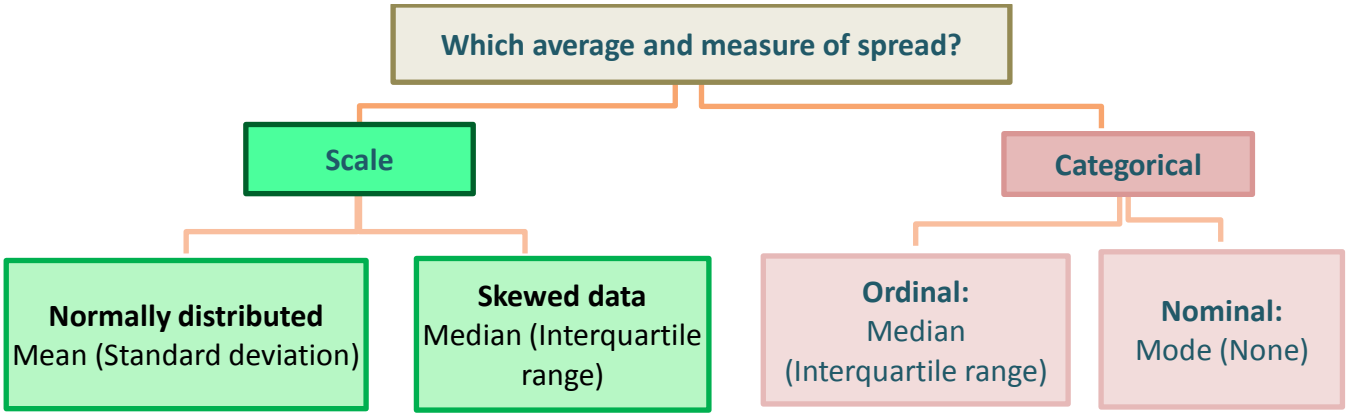
In order to choose suitable summary statistics and analysis for the data, it is also important for students to distinguish between continuous (numerical/ scale) measurements and categorical variables.



Summary Statistics

Students often go straight to the hypothesis test rather than investigating the data with summary statistics and charts first. Encourage them to summarise their data first. As well as summarising their results, charts especially can show outliers and patterns.

For continuous normally distributed data, summarise using means and standard deviations. If the data is skewed or there are influential outliers, the median (middle value) and interquartile range (Upper quartile – lower quartile) are more appropriate.



Asking for the mean, median, minimum, maximum and standard deviation along with producing an appropriate chart will identify outliers and skewed data. A big difference between the mean and median indicates skewed data or influential outliers.

Summary of descriptive and graphical statistics

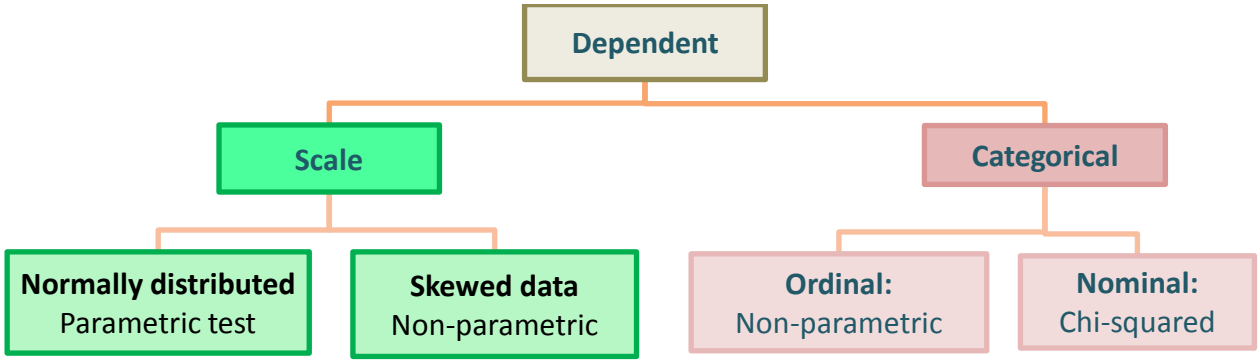
Chart	Variable type	Purpose	Summary Statistics
Pie Chart or bar chart	One Categorical	Shows frequencies/ proportions/percentages	Class percentages
Stacked / multiple bar	Two categorical	Compares proportions within groups	Percentages within groups
Histogram	One scale	Shows distribution of results	Mean and Standard deviation
Scatter graph	Two scale	Shows relationship between two variables and helps detect outliers	Correlation co-efficient
Boxplot	One scale/ one categorical	Compares spread of values	Median and IQR
Line Chart	Scale by time	Displays changes over time Comparison of groups	Means by time point
Means plot	One scale/ 2 categorical	Looks at combined effect of two categorical variables on the mean of one scale variable	Means

Deciding on appropriate statistical methods for research

This is the information you need from a student to help them decide on the most appropriate statistical techniques for their project

What is the main research question? This needs to be able to be defined with specific variables in mind. Which variables (types of measurement) will help answer the research question?

Which is the dependent (outcome) variable and what type of variable is it?



Which are the independent (explanatory) variables, how many are there and what data types are they?

Are relationships or differences between means of interest?

Are there repeated measurements of the same variable for each subject?

Ordinal data

Some departments routinely use parametric tests to analyse ordinal data. As a general rule of thumb, ordinal variables with seven or more categories can be analysed with parametric tests if the data is approximately normally distributed. However, if the students department/ supervisor expect scales of five to be analysed as continuous data, warn them why you think this is not appropriate but let them do it. We are here to advice and help them make decisions. Sometimes, questionnaires have sets of questions trying to measure an underlying latent variable. In that situation, summing or averaging the scores gives a variable which could be considered as scale so parametric tests can be carried out.

Which test should I use?

Common Single Comparison Tests

Comparing:	Dependent (outcome) variable	Independent (explanatory) variable	Parametric test (data is normally distributed)	Non-parametric test (ordinal/ skewed data)
The averages of two INDEPENDENT groups	Scale	Nominal (Binary)	Independent t-test	Mann-Whitney test/ Wilcoxon rank sum
The averages of 3+ independent groups	Scale	Nominal	One-way ANOVA	Kruskal-Wallis test
The average difference between paired (matched) samples e.g. weight before and after a diet	Scale	Time/ Condition variable	Paired t-test	Wilcoxon signed rank test
The 3+ measurements on the same subject	Scale	Time/ condition variable	Repeated measures ANOVA	Friedman test

Tests of association

Relationship between 2 continuous variables	Scale	Scale	Pearson's Correlation Coefficient	Spearman's Correlation Coefficient
Predicting the value of one variable from the value of a predictor variable or looking for significant relationships	Scale	Any	Simple Linear Regression	Transform the data
	Nominal (Binary)	Any	Logistic regression	
Assessing the relationship between two categorical variables	Categorical	Categorical		Chi-squared test

One scale dependent and several independent variables

1st independent	2nd independent	Test
Scale	Scale/ binary	Multiple regression
Nominal (Independent groups)	Nominal (Independent groups)	2 way ANOVA
Nominal (repeated measures)	Nominal (repeated measures)	2 way repeated measures ANOVA
Nominal (Independent groups)	Nominal (repeated measures)	Mixed ANOVA
Nominal	Scale	ANCOVA

Regression or ANOVA? Use regression if you have only scale or binary independent variables. Categorical variables can be recoded to dummy binary variables but if there are a lot of categories, ANOVA is preferable.

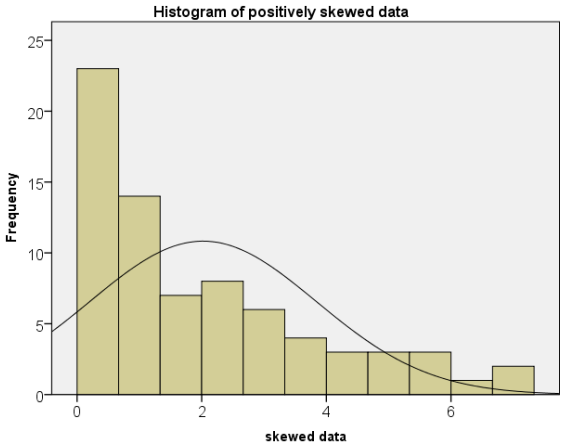
Assumption of normality

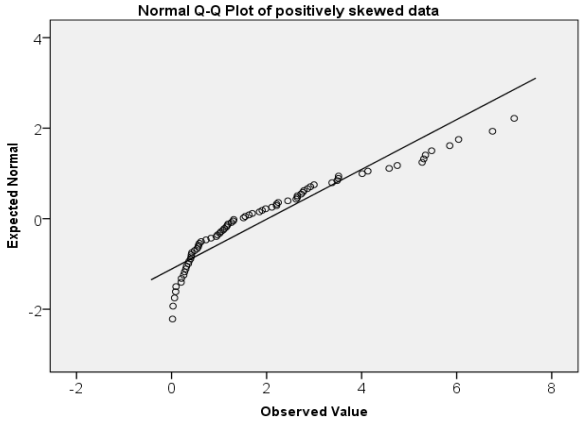
Parametric tests assume that the data follows a particular distribution e.g for t-tests, ANOVA and regression, the data needs to be normally distributed. Parametric tests are more powerful than non-parametric tests, when the assumptions about the distribution of the data are true. This means that they are more likely to detect true differences or relationships that exist.

The tests are quite robust to departures of non-normality so the data only needs to be approximately normally distributed.

Plotting a **histogram** or QQ plot of the variable of interest will give an indication of the shape of the distribution. Histograms should peak in the middle and be approximately symmetrical about the mean. If data is normally distribued, the points in QQ plots will be close to the line.

Below are some examples of very skewed data (i.e. non-normal).





Statistical tests for normality

There are statistical tests for normality such as the *Shapiro-Wilk* and *Kolmogorov-Smirnoff* tests but for small sample sizes (n < 20), the tests are unlikely to detect non-normality and for larger sample sizes (n > 50), the tests can be too sensitive. They are also sensitive to outliers so use histograms (large samples) or QQ plots (small samples).

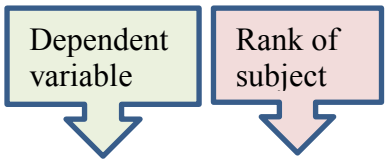
Parametric test	What to check for normality
Independent t-test	Dependent variable by group
Paired t-test	Paired differences
One-way ANOVA	Residuals
Repeated measures ANOVA	Residuals at each time point
Pearson's correlation coefficient	Both variables are normally distributed
Simple linear regression	Residuals

Non-parametric tests

Non-parametric tests make no assumptions about the distribution of the data.

Nonparametric techniques are usually based on ranks or signs rather than the actual data and are usually less powerful than parametric tests.

The example to the right is data on reaction times after drinking either water or alcohol. The reaction times by group were not normally distributed so an independent t-test could not be used. The Mann-Whitney test is more appropriate. It tests the hypothesis that the two distributions are the same. All the data is ordered and ranked from the fastest to the slowest irrelevant of group. The sum of the ranks for each group is used to calculate a test statistic. If there is no difference between the groups the sum of the ranks will be similar. SPSS does all the ranking for you so you don't need to worry about that.



Group	ReactionTime	Rank
Placebo	.37	1
Placebo	.38	2
Placebo	.61	3
Placebo	.78	4
Placebo	.83	5
Placebo	.86	6
Placebo	.90	7
Placebo	.95	8
Alcohol	.98	9
Alcohol	1.11	10
Alcohol	1.27	11
Alcohol	1.32	12
Alcohol	1.44	13
Alcohol	1.45	14
Alcohol	1.46	15
Placebo	1.63	16
Alcohol	1.76	17
Placebo	1.97	18
Alcohol	2.56	19
Alcohol	3.07	20

Non-parametric tests can also be used when other assumptions are not met e.g. equality of variance.

Some people also advise using non-parametric tests for small samples as it is difficult to assess normality.

Other common assumptions

For independent t-tests and ANOVA

Homogeneity of variances: Levene's test

Use: Used to test the equality of variances when comparing the means of independent groups e.g. Independent t-tests and ANOVA.

Note: The violation of this assumption is more serious than violation of the assumption of normality but both t-tests and ANOVA are fairly robust to deviations from this assumption. There are alternative tests within the t-test and ANOVA menus to deal with violations of this assumption.

Interpretation:

If the p-value is less than 0.05 reject H_0 and conclude that the assumption of equal variances has not been met.

For repeated measures ANOVA

Sphericity: Mauchly's test

Use: Tests for sphericity - a measure of whether variances of the differences between all repeated measures are all equal. If the assumption is not met, the F-statistic is positively biased leading to an increased risk of a type 1 error.

Interpretation:

Significant when p-value < 0.05 meaning there are significant differences between the variance of differences, i.e. condition of sphericity is not met. If the assumption is not met, use the Greenhouse-Geisser correction to the degrees of freedom which appears in the standard output.

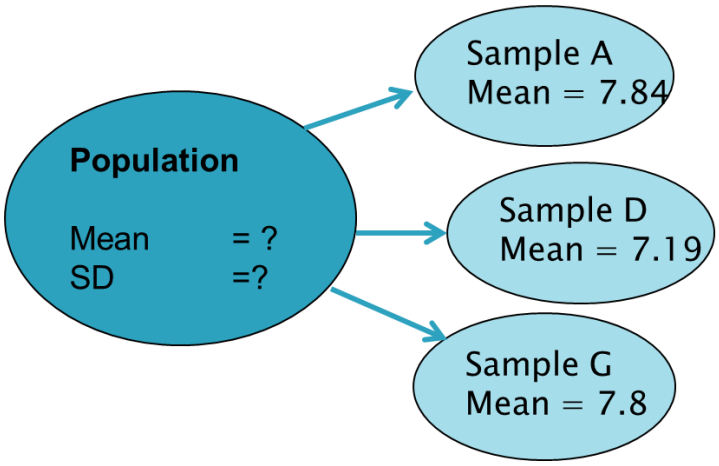
Independent observations

For most tests, it is assumed that the observations are independent. That is the results for one subject* are not affected by another. Examples of data which is not independent are repeated measures on the same subject (use the specific tests for this type of experiment) and observations over time (check the Durbin Watson test for regression). Another situation where observations are not independent is when subjects are nested within groups with a common influence e.g. children within classes who may be influenced by the teacher (use multilevel modelling to include class as an extra RANDOM factor). Time series analysis (which allows for non-independent measurements over time) and multilevel modelling are beyond the scope of most students.

*The subject is the unit of interest which could be a person, an observation, a day etc.

Confidence intervals

Most research uses sample data to make inferences about the wider population. A population is the group of individuals you are interested in e.g. a study into the weight of babies born in Sheffield would use a sample but the results apply to the whole population.



Every sample taken from a population will contain different babies so the mean value varies especially if the sample size is small.

Confidence Intervals describe the variability surrounding the sample point estimate (the wider the interval, the less confident we can be about the estimate of the population mean). In general, all things being equal, the larger the sample size the better (more precise) the estimate is, as less variation between sample means is expected.

The equation for a 95% Confidence Interval for the population mean when the population standard deviation is unknown and the sample size is large (over 30) is

$$\overline{X} \pm 1.96.\frac{s}{\sqrt{n}}$$

$\overline{X}$ = sample mean, n = number in sample, $\frac{s}{\sqrt{n}}$ = standard error

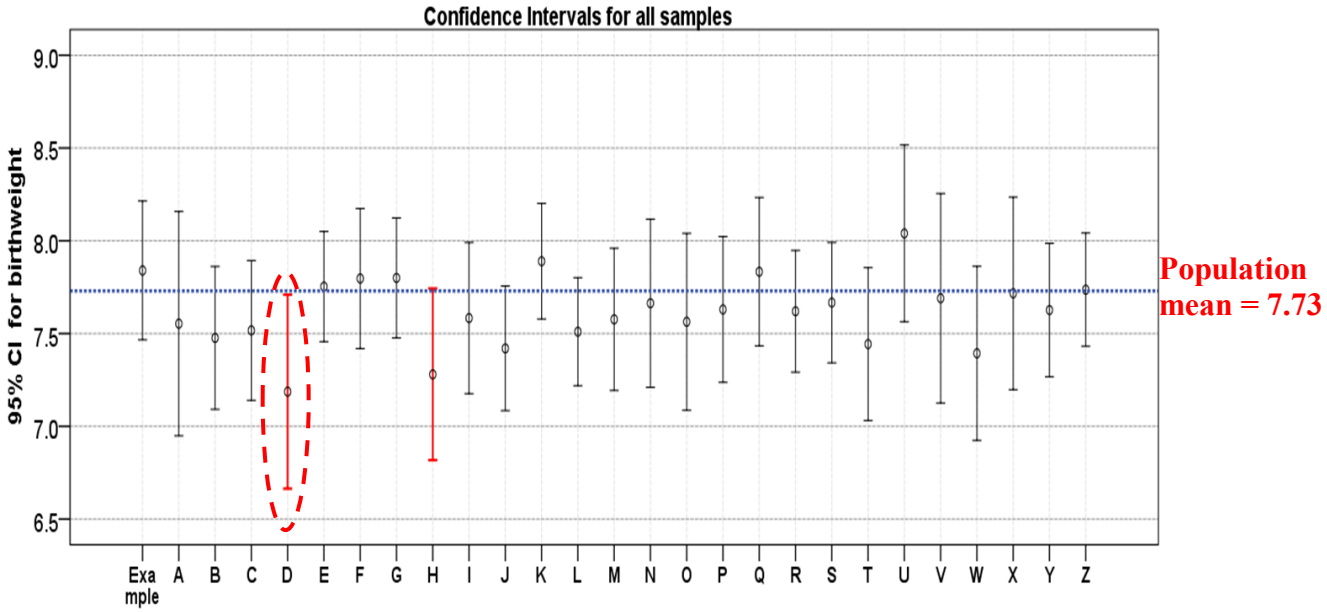
For example, sample D of 30 babies born in 2013 had a mean weight of 7.19lbs with a standard deviation of 1.4, the 95% Confidence Interval for the population mean of all babies is:

$$7.19 \pm 1.96 \cdot \frac{1.4}{\sqrt{30}} = 7.19 \pm 0.5 = (6.7\,,7.7)$$

We would expect the population mean to be between 6.7 lbs and 7.7 lbs.

Confidence intervals give a range of values within which we are confident (in terms of probability) that the true value of a population parameter lies. A 95% CI is interpreted as 95% of the time the CI would contain the true value of the population parameter.

The diagram below shows the confidence intervals for 27 samples of babies taken from the same population. The actual population mean (which is not normally known) is 7.73 lbs. Two of the confidence intervals do not contain the population mean (don't overlap 7.73 lbs) including the one previously calculated.



The website CAST has some great applets for demonstrating concepts to students. This ebook contains core material including an applet for demonstrating confidence intervals http://cast.massey.ac.nz/collection_public.html

There is a strong relationship between hypothesis testing and confidence intervals. For example, when carrying out a paired t-test, if the p-value < 0.05, the 95% confidence interval for the paired differences will not contain 0. However, a p-value just concludes whether there is significant evidence of a difference or not. The confidence interval of the difference gives an indication of the size of the difference.

For more information on the use of confidence intervals see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1339793/.

Hypothesis testing

Hypothesis testing is an **objective** method of making decisions or **inferences** from sample data (evidence). Sample data is used to choose between two choices i.e. **hypotheses** or statements about a population. Typically this is carried out by comparing what we have observed to what we expected if one of the statements (**Null Hypothesis**) was true.

Key terms:

NULL HYPOTHESIS (H0) is a statement about the population & sample data used to decide whether to reject that statement or not. Typically the statement is that there is no difference between groups or association between variables.

ALTERNATIVE HYPOTHESIS (H_1) is often the research question and varies depending on whether the test is one or two tailed.

SIGNIFICANCE LEVEL: The probability of rejecting the null hypothesis when it is true, (also known as a type 1 error). This is decided by the individual but is normally set at 5% (0.05) which means that there is a 1 in 20 chance of rejecting the null hypothesis when it is true.

TEST STATISTIC is a value calculated from a sample to decide whether to accept or reject the null (H_0) and varies between tests. The test statistic compares differences between the samples or between observed and expected values when the null hypothesis is true.

P-VALUE: the probability of obtaining a test statistic at least as extreme as ours if the null is true and there really is no difference or association in the population of interest. P-values are calculated using different probability distributions depending on the test. A significant result is when the p-value is less than the chosen level of significance (usually 0.05).

The court case

Hypothesis testing can be thought of as a court case

Members of a jury have to decide whether a person is guilty or innocent based on evidence presented to them.

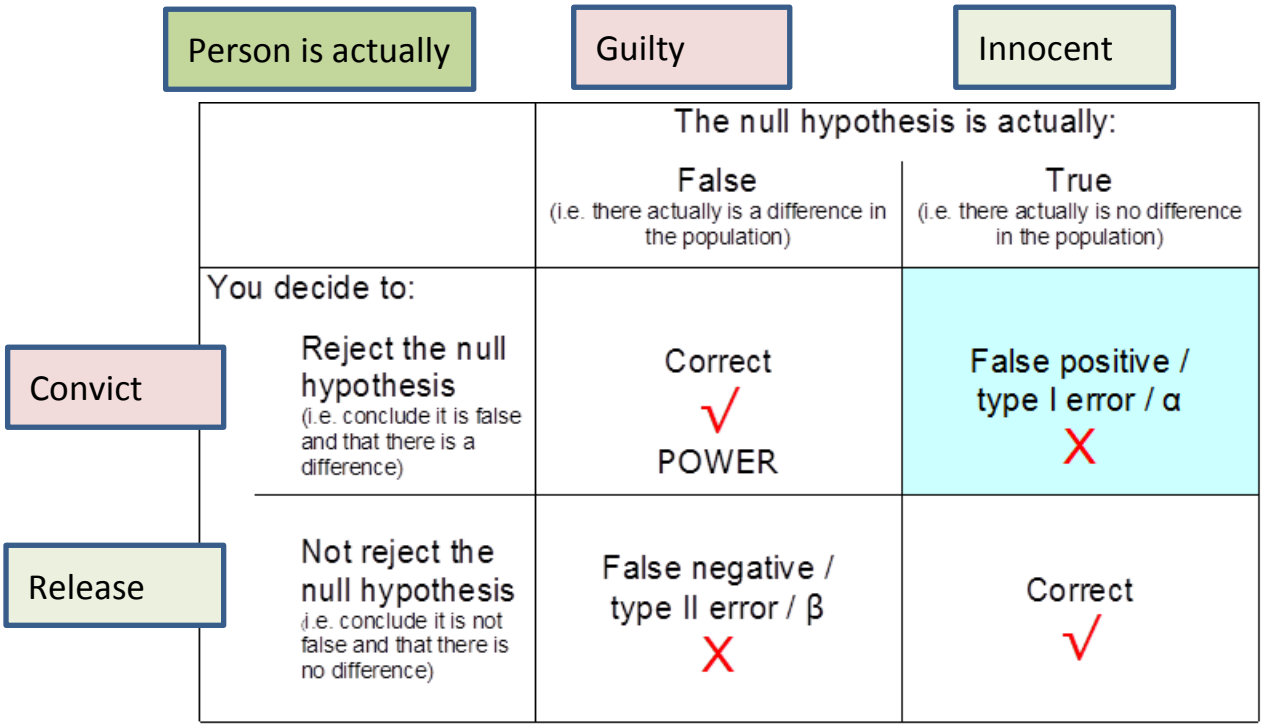
Null: The person is innocent

Alternative: The person is not innocent.

The null can only be rejected if there is enough evidence to disprove it and the jury do not know whether the person is really guilty or innocent so they may make a mistake.

If a court case was a hypothesis test, the jury consider the likelihood of innocence given the evidence and if there's less than a 5% chance that the person is innocent they reject the statement of innocence.

In reality, the person is actually Guilty (null false) or Innocent (null true) but we can only conclude that there is evidence to suggest that the null is false or not enough evidence to suggest it is false.



A type I error is equivalent to convicting an innocent person and is usually set at 5% (the magic 0.05!).

Multiple testing

Some students will try to perform a large number of tests on their data. The chance of a type I error increases with the number of tests. Adjustments to keep the type I error low for a larger number of tests are included as post hoc tests in ANOVA. This will mean less of the results are statistically significant. The most commonly used post hoc tests are Tukey and Sidak although Scheffe's is often used in medicine.

Suggest that the student looks in their notes or papers in their field when choosing. If adjustments need to be made by hand, the Bonferroni adjustment is the easiest to explain although it is the most conservative (least likely to lead to rejection of the null). Either divide the significance level initially used (probably 0.05) by the number of tests being carried out and compare the p-value with the new, smaller significance level. Alternatively, multiply the p-value by the number of tests being carried out and compare to 0.05. This is the standard adjustment made after the Kruskall-Wallis and has a maximum limit of 1 (as it's a probability!).

Sample size and hypothesis tests

The larger the sample size, the more likely a significant result is so for small sample sizes a huge difference is needed to conclude a significant difference. For large sample sizes, small differences may be significant but check if the difference is meaningful.

Effect size

An effect size is a measure of the strength or magnitude of the effect of an independent variable on a dependent variable which helps assess whether a statistically significant result is meaningful.

For example, **for a t-test**, the absolute effect size is just the difference between the two groups. A standardised effect size involves variability and can then be compared to industry standards.

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$

Cohen gives the following guidance for the effect size d, although it is not always meaningful:

0.2 to 0.3 might be a "small" effect, around 0.5 a "medium" effect and 0.8 to infinity, a "large" effect.

Partial eta-squared

Partial eta-squared is a measure of variance. It represents the proportion of variance in the dependent variable that is explained by the independent variable. It also represents the effect size statistic. The effects sizes given in Cohen (1988) for the interpretation of the absolute effect sizes are:

$\eta^2 = 0.010$ is a small association.

$\eta^2 = 0.059$ is a medium association.

$\eta^2 = 0.138$ or larger is a large association.

Section 2

The most common statistical techniques used

Independent t-test

Dependent variable: Continuous

Independent variable: Binary (Group)

Use: A t-test is used to compare the means of two independent groups. Independent groups means that different people are in each group.

Plot: Box-plots (exploratory)or Confidence Interval plots with results

Assumptions	**How to check**	**What to do if assumption is not met**
Normality: dependent variables should be normally distributed within each group	Histograms of dependent variables per group / Shapiro Wilk	Mann-Whitney / Wilcoxon rank sum
Homogeneity of variance	Levene's test * (part of standard SPSS output)	Use bottom row of t-test output in SPSS

*Levenes test: If the assumption of homogeneity is not met, correct for this violation by not using the pooled estimate for the error term for the t-statistic and also by making adjustments to the degrees of freedom using the Welch-Satterthwaite method. SPSS does the automatically in the "Equal variances not assumed" row. Alternatively, a Mann-Whitney test can be carried out.

Interpretation:

If the p-value < 0.05, there is a significant difference between the means of the two groups. Report the means of the two groups or the mean difference and confidence interval from the SPSS output to describe the difference.

SPSS: Analyse → Compare means → Independent-samples T-test

Mann-Whitney test

(non-parametric equivalent to the independent t-test)

Dependent variable: Ordinal/ Continuous

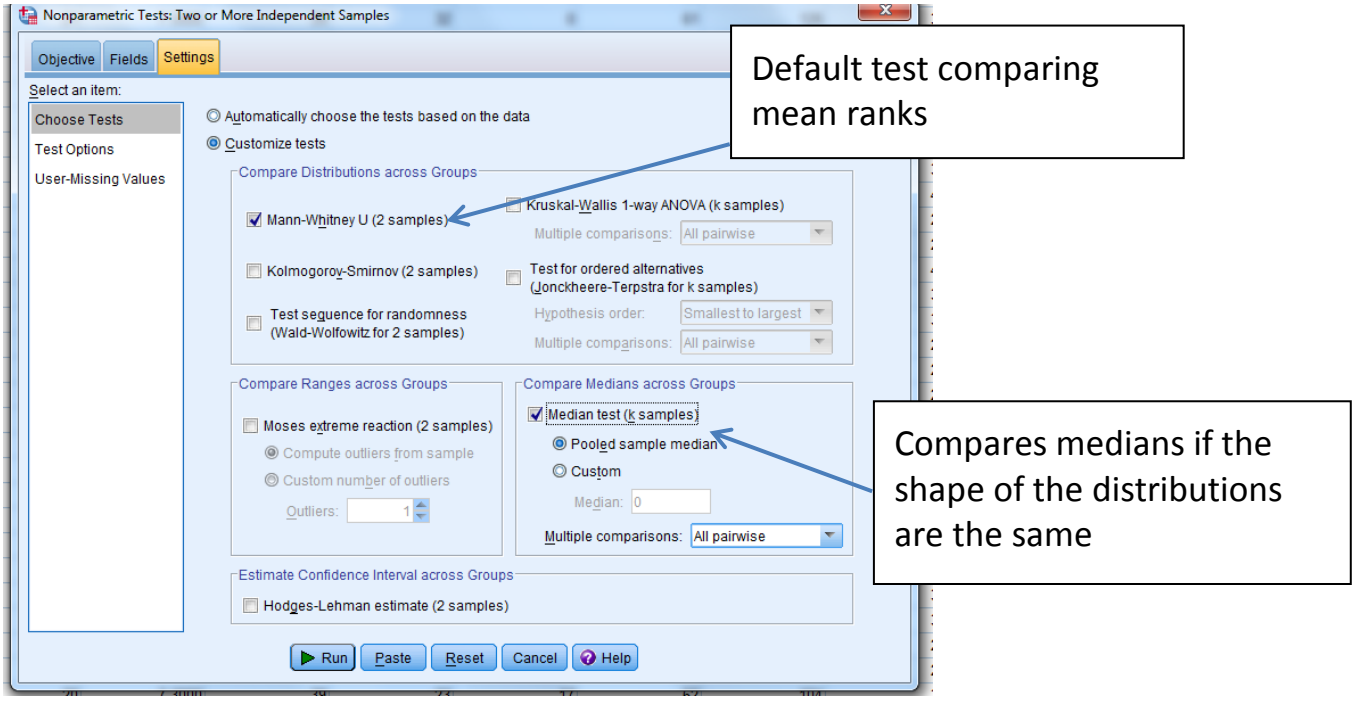
Independent variable: Binary(Group)

The Mann-Whitney test is also known as the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test.

Use: It is used to compare whether two groups containing different people are the same or not. The Mann-Whitney test ranks all of the data and then compares the sum of the ranks for each group to determine whether the groups are the same or not. There are two types of Mann-Whitney U tests. If the distribution of scores for both groups have the same shape, the medians can be compared. If not, use the default test which compares the mean ranks.

Plot: Histograms of the two groups

SPSS: Analyse → Nonparametric Tests → Independent Samples



Paired t-test

Dependent variable: Continuous (at least interval)

Independent variable: Time point 1 or 2/ condition

Use: A paired samples t-test can only be used when the data is paired or matched. Either there are before/after measurements of the same variable or the t-test can be used to compare how a group of subjects perform under two different test conditions. The test assesses whether the mean of the paired differences is zero.

Plot: Histogram of differences

Assumptions	**How to check**	**What to do if assumption is not met**
Normality: paired differences* should be normally distributed	Histogram of differences / Shapiro Wilk	Wilcoxon signed rank

Interpretation:

If the p-values <0.05 then there is a statistically significant difference between the two time points/experiments. Report the mean difference.

SPSS: Analyse → Compare means → Paired-samples T-test

*Paired differences can be calculated using *Transform → Compute variable*

Wilcoxon Signed Rank test

(non-parametric equivalent to the paired t-test)

Dependent variable: Ordinal/ Continuous

Independent variable: Time/ Condition (binary)

Use: The Wilcoxon signed rank test is used to compare two related samples, matched samples or repeated measurements on a single sample to assess whether their population mean ranks differ. It is a paired difference test and is the non-parametric alternative to the paired t-test. The absolute differences are ranked then the signs of the actual differences used to add the negative and positive ranks.

Plot: Histogram of differences

Interpretation:

If p-value < 0.05 then there is evidence that the population mean ranks differ. Report the medians of the two sets of measurements

SPSS: Analyse → Nonparametric tests → Legacy dialogs →2 related samples

Other related tests:

Sign: Compares the number of negative and positive differences

McNemar: Can be used for binary nominal variables when changes in a subjects score are of interest. It compares the number of subjects who have changed their score in a positive direction with those changing their score in a negative direction.

One-way ANOVA

Dependent variable: Continuous

Independent variable: Categorical (at least 3 categories)

Use: Used to detect the difference in means of 3 or more independent groups. It can be thought of as an extension of the t-test for 3 or more independent groups. ANOVA uses the ratio of the between group variance to the within group variance to decide whether there are statistically significant differences between the groups or not.

Plot: Box-plots or confidence interval plots

Assumptions	**How to check**	**What to do if assumption is not met**
Residuals should be normally distributed	Histogram/ QQ plot of residuals / SW	Kruskall-Wallis test (non-parametric)
Homogeneity of variance	Levene's test / Bartlett's test	Welch test instead of ANOVA (adjusted for the differences in variance) and Games-Howell post hoc or Kruskall-Wallis

Interpretation:

ANOVA tests "H_0: all group means are equal" using an F-test. The p-value concludes whether or not there is at least one pairwise difference. If the p-value < 0.05, we reject H_0 and conclude that there is a significant difference between at least one pair of means. Post-hoc tests are used to test where the pairwise differences are. Report the significant pairwise differences and the means.

Post-hoc adjustments:

Tukey or Scheffe are most commonly used but check if their department uses something else. If treatments are being tested against a control use Dunnett. If group sample sizes vary use Hochberg's GT2. If there is a difference between the group variances (Levene's test gives a p-value < 0.05 so we reject H_0) use Games-Howell.

SPSS: Analyse → General Linear model → Univariate (Welch test unavailable)

Or Compare means → ANOVA although dependent variable by group will have to be checked for normality as there's no option for calculating residuals.

Kruskal-Wallis test

(non-parametric equivalent to the one-way ANOVA)

Dependent variable: Ordinal/ Continuous

Independent variable: Categorical

Use: Kruskal-Wallis compares the medians of two or more samples to determine if the samples have come from different populations. It is an extension of the Mann–Whitney U test to 3 or more groups. The distributions do not have to be normal and the variances do not have to be equal.

Plot: Box-plots

Assumptions	**How to check**	**What to do if assumption is not met**
Independent observations	Check data	Friedman
Similar sample sizes	Check data	
>5 data points per sample (ideally)	Frequencies of group	

Interpretation:

When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. The test does not identify where the differences occur or how many differences actually occur. The Mann-Whitney U test would help analyse the specific sample pairs for significant differences. Make sure a p-value correction for multiple testing in is used in the post-hoc tests.

SPSS: Analyse → Nonparametric tests → Independent samples

Note: Double click on the output for the test and a second screen appears with a lot more information including the post-hoc tests with Bonferroni adjustments. Select 'Pairwise comparisons' from the list in the bottom right hand corner if the main test is significant Use the 'adjusted sig.' column which is the Mann-Whitney p-value multiplied by the number of pairwise tests (Bonferroni correction).

One-way ANOVA with repeated measures (within subjects)

Dependent variable: Continuous

Independent variable: categorical with "levels" as the within subject factor

Use: Tests the equality of means in 3 or more groups. All sample members characteristics must be measured under multiple conditions i.e. the dependent variable is repeated. Standard ANOVA cannot be used as the assumption of independence has been violated. This is the equivalent of a one-way ANOVA but for repeated samples and is an extension of a paired-samples t-test. It is used to analyse (1) changes in mean score over 3 or more time points (2) differences in mean score under 3 or more conditions. It separates the variance from measures and from people, hence decreasing the mean squared error.

Assumptions	**How to check**	**What to do if assumption is not met**
Normality of residuals by time point	Histograms of residuals etc	Friedman test (non-parametric)
Sphericity: variances of the differences between each pair of repeated measures are all equal	Mauchly's test	a Greenhouse-Geisser correction or the Huynh-Feldt correction to the df

Interpretation:

If the main ANOVA is significant, there is a difference between at least two time points. The Bonferroni post hoc tests will conclude where those differences are. Report the significant post hoc results and means at each time point.

Post-hoc adjustments:

Of the three post hoc adjustments in 'Options', Bonferroni is most commonly used.

SPSS Analyse → General Linear Model → Repeated measures

Note: The factor e.g. time needs to be specified before the main analysis screen appears. The repeated measures are in different columns and are entered in the main screen.

It is also possible to carry out a two way repeated measures ANOVA and a 'mixed' between – within ANOVA when there are independent groups as well as repeated measures. For the 'mixed' ANOVA, a means plot with time/condition on the x axis and separate lines for each independent group is very useful. For instructions see Brunel ASK video.

https://www.youtube.com/watch?v=duUEb_j9wfY&list=UUdb6U06idJIt7IWNr5YzAdA

Friedman test

(non-parametric equivalent to repeated measures ANOVA)

Dependent variable: Ordinal/ Continuous measured on at least 3 occasions or 3 measures under different conditions

Independent variable: Time/ Condition

Use: The Friedman test is used to detect differences in scores across multiple occasions or conditions. The scores for each subject are ranked and then the sums of the ranks for each condition are used to calculate a test statistic. The Friedman test can also be used when subjects have ranked a list e.g. rank these pictures in order of preference.

Interpretation:

If the Friedman test is significant (p-value < 0.05) then there are differences in the distributions across the time points/ conditions.

Post-hoc tests:

To examine where the differences actually occur, separate Wilcoxon signed-rank tests on the different combinations of related groups are run with the Bonferroni adjustment. The adjusted p-value column is the Wilcoxon p-value multiplied by the number of tests.

SPSS: Analyse → Non-parametric tests → Related samples

Note: Double click on the output for the test and a second screen appears with a lot more information including the post-hoc tests with Bonferroni adjustments. Select 'Pairwise comparisons' from the list in the bottom right hand corner.

Two-way ANOVA

Dependent variable: Continuous

Independent variables: Two categorical (2+ levels within each)

Use: Comparing means for combinations of two independent categorical variables (factors).

There are three sets of hypothesis with a two-way ANOVA. H_0 for each set is as follows:

- The population means of the first factor are equal – equivalent to a one-way ANOVA for the row factor.
- The population means of the second factor are equal – equivalent to a one-way ANOVA for the column factor.
- There is no interaction between the two factors – equivalent to performing a test for independence with contingency tables (a chi-squared test for independence).

Plot: Means plot to look at interaction between the two independent variables. Use the lines plot option but ask for the mean rather than frequencies of the dependent variable.

Assumptions	**How to check**	**What to do if assumption is not met**
Residuals should be normally distributed	Use the 'Save' menu within GLM to request the residuals and then use 'Explore' to produce histogram/ QQ plot of residuals / Shapiro Wilk test	Transform the data
Homogeneity of variance	Levene's test / Bartlett's test	Compare p-values with a smaller significance level e.g. 0.01

Interpretation:

When interpreting the results you need to return to the hypotheses and address each one in turn. If the interaction is significant, the main effects cannot be interpreted from the ANOVA table. Use the means plot to explain the effects or carry out separate ANOVA by group.

Post-hoc adjustments:

Tukey or Scheffe are generally used. For testing treatments against a control use Dunnett, if group sample sizes vary use Hochberg's GT2 and if there is a difference between the group variances (Levene's test gives a p-value < 0.05) use Games-Howell.

SPSS: Analyse → General Linear Model → Univariate

Chi-squared test

(non-parametric)

Dependent variable: Categorical

Independent variable: Categorical

Use: The null hypothesis is that there is no relationship/association between the two categorical variables. The chi-squared test compares expected frequencies, assuming the null is true, with the observed frequencies from the study. When obtaining a significant chi-squared result, calculate percentages in a table to summarise where the differences between the groups are.

Plot: Stacked/ multiple bar chart with percentages

Assumptions	**How to check**	**What to do if assumption is not met**
80% of expected cell counts >5	SPSS tells you under the test	Fisher's exact (usually for 2x2 tables, but can also be used for others) or merge categories where sensible
No cells with expected frequency below 1		

Interpretation:

If p< 0.05, there is significant evidence of relationship between the two variables. Use %'s to describe what the relationship is.

For 2 x 2 tables, use Yates continuity correction.

For comparing 2 paired proportions e.g. proportions of people changing their response given some information about the topic, use McNemar's test.

SPSS:Analyse → Descriptive → Crosstabs → Statistics

Odds and Relative Risk

Odds

The odds of an event happening is defined as follows where p_1 is the probability of an event happening:

$$\frac{\text{probability of an event happening}}{\text{probability of an event not happening}} = \frac{p_1}{1-p_1}, \qquad 0 < p < 1$$

Odds Ratio

The odds ratio is a measure of effect size, describing the strength of association or non-independence between two binary data values. If the probabilities of the event happening in each of the groups are p_1(Group 1) and p_2(Group 2), then the odds ratio is:

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

An odds ratio of 1 implies that the event is equally likely in both groups. An odds ratio greater than one implies that the event is more likely to occur in the first group, while less than one implies that it has more chance to happen in group 2.

Relative Risk (RR)

Relative risk is a similar and more direct measure of comparing the probabilities in two groups. As with the odds ratio, a relative risk equal to 1 means that the event is equally probable in both groups. A relative risk larger than 1 implies that the chance of the event occurring is higher in group 1 and smaller for group 2.

Summary

	Number of times the event happened	Number of times the event did not happen
Group 1	a	b
Group 2	c	d

$$\text{Odds ratio} = \text{OR} = \frac{a/b}{c/d} = \frac{ad}{bc} \qquad \text{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}$$

Correlation

Use: Correlation (r) is used to measure the strength of association between two variables and ranges between -1 (perfect negative correlation) to 1 (perfect positive correlation). Cohen (1992) has the following interpretation of the absolute value of the correlation:

Correlation coefficient value			Association
-0.3 to +0.3			Weak
-0.5 to -0.3	or	0.3 to 0.5	Moderate
-0.9 to -0.5	or	0.5 to 0.9	Strong
-1.0 to -0.9	or	0.9 to 1.0	Very strong

Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159

Plot: Scatterplot

SPSS: Analyse → Correlate → Bivariate Correlation

Pearson's correlation coefficient

Dependent variable: Continuous

Independent variable: Continuous

Pearson's correlation coefficient is the most common measure of correlation.
rho (ρ) = population correlation and r = sample correlation

Assumptions	**How to check**	**What to do if assumption is not met**
Continuous data for each variable	Check data	If ordinal data use Spearman's or Kendall tau
Linearly related variables	Scatter plot	Transform data
Both variables are normally distributed	Histograms of variables/ Shapiro Wilk	Use rank correlation: Spearman's or Kendall tau

Ranked correlation coefficients

Dependent variable: Continuous/ Ordinal

Independent variable: Continuous/ Ordinal

Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is a non-parametric statistical measure of the strength of a monotonic relationship between paired data. The notation used for the sample correlation is r_s.

Assumptions	How to check	What to do if assumption is not met
Linearly related variables	Scatter plot	Transform data

Kendall's Tau Rank Correlation Coefficient

Use for small data sets with a large number of tied ranks

Use: Kendall's tau rank correlation coefficient is used to measure the association between two measured quantities. A tau test is a non-parametric hypothesis test for statistical dependence based on the tau coefficient. Specifically, it is a measure of rank correlation, i.e. the similarity of the orderings of the data when ranked by each of the quantities.

The Tau-b statistic makes adjustments for ties; values of Tau-b range from −1 (100% negative association or perfect inversion) to +1 (100% positive association or perfect agreement). A value of zero indicates the absence of association.

Partial correlation

Partial correlation allows for a third continuous or binary variable to be controlled for. The correlation then measures the association between the independent and dependent variables after removing the variation due to the control.

Regression

There are many different types of regression. The type of regression that is used depends on the type of dependent variable e.g. linear regression is used with a continuous dependent variable, logistic regression with a binary dependent variable and Poisson regression with a Poisson counts dependent variable.

Use: Regression can be used with many continuous and binary independent variables (x). It gives a numerical explanation of how variables relate, enables prediction of the dependent variable (y) given the independent variable (x) and can be used to control for confounding factors when describing a relationship between two variables. Regression produces a line of best fit by minimising the RSS (residual sum of squares) and tests that the slope is 0 for each independent variable. Note: A residual is the difference between an observed y and that predicted by the model.

Linear Regression

Dependent variable: Continuous

Independent variables: Any but categorical must be turned into binary dummy variables

Assumptions	How to check	What to do if assumption is not met
Independent observations (no correlation between successive values)	Durbin Watson = 1.5 – 2.5	Time series – beyond the scope of the tutor and the student!
Residuals should be normally distributed	Histogram/ QQ plot of residuals / SW	Transform the dependent variable
The relationship between the independent and dependent variables is linear	Scatterplot of independent and dependent variables	Transform either the independent or dependent variable
Homoscedasticity: The variance of the residuals about predicted responses should be the same for all predicted responses	Scatterplot of predicted values against residuals	Transform the dependent variable
No observations have a large overall influence (leverage)	Look at Cook's and Leverage distances	Remove observation with very high leverage

Interpretation

ANOVA table: The ANOVA table decides whether the model as a whole is significant. The model is compared to a 'null' model where every observation is predicted to be the same.

Coefficients table: The 'B' column in the coefficients table, gives us the values of the slope and intercept terms for the regression line. For multiple regression, (where there are several predictor variables), the coefficients table shows the significance of each variable individually after controlling for the other variables in the model.

Model summary: The R^2 value shows the proportion of the variation in the dependent variable which is explained by the model. It varies from 0 to 1 but is usually reported as a percentage. The better the model, the higher the R^2 value. The level for a 'good model' varies by discipline but above 70% is generally considered to be good for prediction.

TIP: The majority of students coming to the centre will just be using regression to look for significant relationships so don't confuse them with model selection or comparing models unless they ask about it.

Comparing regression models using R^2

R^2 increases for each additional variable added so the adjusted R^2 is better for comparing models where the dependent variable is the same. The adjusted R^2 takes into account the number of degrees of freedom of the model. When comparing regression models where the dependent variable was transformed or which used different sets of observations, R^2 is not a reliable guide to model quality. There are options for adding variables in blocks in SPSS which enables comparison of models in the output. The change in R^2 is tested formally.

Model selection

There are several methods for model selection (Forwards, Backwards and stepwise) available within SPSS which result in a model only containing significant variables.

Dummy variables

A dummy variable is a binary variable representing one category of a categorical variable e.g. for marital status, code as separate variables Married yes/ no, Divorced yes/ no etc.

Interactions

Interactions need to be created by multiplying the two variables of interest using

Transform → Compute variable.

Logistic Regression

Dependent variable: Binary

Independent variables: Any (Use the 'Categorical' option to specify categorical variables and SPSS creates the dummy variables)

Use: One of the most commonly used tests for categorical variables is the Chi-squared test which looks at whether or not there is a relationship between two categorical variables but this doesn't make an allowance for the potential influence of other explanatory variables. For continuous outcome variables, multiple regression can be used for

controlling for other explanatory variables when assessing relationships between a dependent variable and several independent variables
predicting outcomes of a dependent variable using a linear combination of explanatory (independent) variables

Logistic regression does the same but the outcome variable is categorical. It leads to a model which can predict the probability of the event happening for an individual.

$$\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q \qquad \text{Odds ratio} = \frac{p}{1-p}$$

Note: This a parametric test that does not require the data to be normally distributed.

Interpretation:

If probabilities (p) of the event of interest happening for individuals are needed, the logistic regression equation can be written as:

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}, \quad \text{for} \quad 0 < p < 1$$

Initially look for those variables where the p-value is less than 0.05. These are the factors that have a statistically significant affect on the prediction of the dependent variable. When interpreting the differences, it is easier to look at look at the $\exp(\beta_i)$ which represents the odds ratio for the individual variable.

Section 3

Other statistical tests and techniques

Proportions test (z-test)

Use: The proportions test is used to test whether the proportions of two populations differ significantly with respect to one characteristic.

The test:

The null hypothesis is $p_1 - p_2 = 0$.

The Z-test statistic is calculated as follows:

$$\frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Interpretation:

If the resulting p-value is significant (p-value < 0.05) then there is evidence of a statistically significant difference between the two proportions.

Assumptions	**How to check**	**What to do if assumption is not met**
At least 10 observations in each group	Frequencies	-

One-sample

SPSS> Analyze > Non Parametric > Chi-square (but weight cases)

Two-samples

SPSS> Analyze > Crosstabs > Options – Choose z-test

Reliability

Reliabilty can be divided into two main sections. The reliability between raters (interrater agreement) and the reliability of a set of questions when measuring an underlying variable (Cronbach's alpha).

Interrater reliability

The technique for assessing agreement between raters/ instruments depends on the type of variable being compared. For nominal data, Cohen's kappa should be used and for scale data, the Intraclass Correlation Coefficient (ICC) should be used. Data should be entered with one column for each rater and one row for each subject being rated. Intrarater reliability is when measurements from the same person are being compared.

Cohen's Kappa

Use: Assessing agreement between raters for categorical variables. Kappa compares the proportion of actual agreement between raters (P_A) to the proportion expected to agree by chance (P_C). The P_C values use expected values calculated as in the Chi-squared test for association (row total x column total/grand total).

$$\text{kappa} = \frac{P_A - P_C}{1 - P_C}$$

Interpretation:

There is a test for agreement which tests the hypothesis that agreement is 0 but like correlation, the interpretation of the coefficient itself is more important. The following guidelines were devised by Landis and Koch (1977).

Cohen's kappa	Strength of agreement
< 0	Poor (Agreement worse than by chance)
0 - 0.2	Slight
0.21 - 0.4	Fair
0.41 – 0.6	Moderate
0.61 – 0.8	Good
0.81 - 1	Very good

SPSS: Analyse → Descriptive Statistics → Crosstabs

Select 'Kappa' from the statistics options

Note: Kappa looks at exact matches only so for ordinal data, weighted Kappa is preferable but this is not an option in SPSS. For ordinal data, if exact matches are required, use Kappa or consider using the ICC for scale data for close matches.

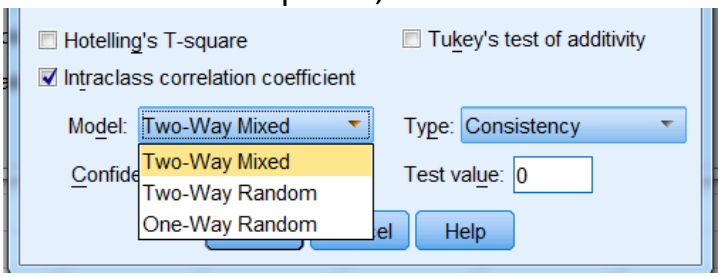
Intraclass Correlation Coefficient

Use: A measure of agreement of continuous measurements for two or more raters. There are several options for the ICC in SPSS. To choose the right combination of model, type and form, you need to ask the student the following questions:

Do all the raters rate all the subjects?
Are the raters the only possible raters (whole population) or a sample of raters?
Do raters need to match exactly or just be consistent (so raters may rank the subjects the same even if their scores don't match)?
Would you normally use just one rater or take an average of several raters?

SPSS:Analyse → Scale → Reliability Analysis

In the 'Statistics' options, select 'Intraclass correlation coefficient'



Models:

One way random	Not all raters have scored all the subjects
Two way random	Random selection of raters and subjects (default)
Two way mixed	Analysis contains whole population of raters

Type:

Absolute agreement	Exact matches on scores/ measurements are required
Consistency	Raters are consistent with their scoring e.g. rater A consistently scores lower than rater B

Form (appears in output): Single measures is used if when not testing for reliability, a subjects score is from one rater only. If a score is usually based on an average of several (k) raters, use average measures.

Interpretation:

Interpret in the same way as Cohen's kappa.

Cronbach's alpha (reliability of scales)

Researchers often use sets of likert style questions to measure an underlying latent variable that cannot be measured exactly. The set of questions is called a scale and the individual questions are called items. The scores for the items can be added or averaged to give an overall score for the scale. It is important that the items are all measuring the same underlying variable.

Use: Cronbach's alpha is a measure of internal consistency (how closely related the items are as a group). If the questions relate to the same issue, participants will be expected to get similar scores on each question.

This measure is not robust against missing data.

Interpretation:

Cronbach's alpha ranges from 0 to 1 and scores are expected to be between 0.7 and 0.9. Below is a commonly accepted rule of thumb for interpreting Cronbach's alpha.

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Very high consistency (the items are so similar that some may not be needed)
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$\alpha < 0.7$	Poor internal consistency

SPSS: Analyse > Scale > Reliability Analysis.

Make sure 'alpha' is selected as the Model in the dialogue box that appears.

Multivariate techniques

Multivariate techniques generally have more than one dependent variable but Multiple Regression and Discriminant Analysis are also referred to as multivariate despite only having one dependent variable. Multivariate techniques tend to involve classification or data reduction. The following table contains some of the more commonly used techniques.

Purpose	Dependent variable (type)	Independent variable (type)	Analysis
Data reduction	2+ (Scale/ binary although ordinal often used)		Principal Components Analysis
			Factor Analysis
	2+ (categorical)		Correspondence Analysis
Identify groups of similar subjects	2+ (Any)		Cluster Analysis
Compare groups	2+ (Scale)	1+ (Categorical)	MANOVA
		1+ (Categorical) & 1+ (Scale)	MANCOVA
Predict group membership	1 (Categorical)	Any	Linear Discriminant Analysis

Principal Component Analysis (PCA)

Use: Principal component analysis aims to reduce the number of inter-correlated variables to a smaller set which explains the overall variability almost as well. It produces new variables which are linear combinations of the original variables called Principal Components (PC's) or factors with the first PC explaining the most variation. These new variables can be used in further analysis e.g. regression. PCA can be based on either the correlation or covariance matrix. If variables are not on the same scale, the variable with the largest variance will dominate the first PC. If the variables are measured on a similar scale, use the covariance matrix, otherwise use the correlation matrix (default in SPSS).

Principal components analysis is not usually used to identify underlying latent variables but if interpretation of which variables contribute most to each PC is of interest, choose a method of rotation of the loadings (correlations between individual variables and the PC's) to identify clearer patterns. The varimax (variance maximising) rotation of the loadings maximises the variability of the new PC whilst minimising the variance around the new variable. It assumes that the PC's are not related.

Dependents: Scale/ binary (preferably mostly scale or mostly binary)
Numerical ordinal variables are often included but the choice of numbering e.g. 1, 2, 3 will impact on the results

Assumptions	How to check
Normality of variables	Histograms
Minimum sample size	Although PCA can be carried out on any number of cases, 5 – 10 cases per variable are often suggested for reliable results. Smaller sample sizes are reasonable if high loadings are achieved (0.8+) but the last few PC's are non-informative and shouldn't be used.
Variables should be adequately related	Correlation matrix shows coefficients above 0.3. Additional checks: KMO > 0.6, Bartlett's $p < 0.05$
No significant outliers	Component scores more than 3 SD's from mean

SPSS: Analyse → Dimension Reduction → Factor

Non-default options to select:

Menu	What to select
Descriptives	Coefficients, KMO and Bartlett's
Extraction	Scree plot (Note: Correlation matrix is the default so select 'Covariance matrix' if variables are measured on a smilar scale)
Rotation	Varimax (most commonly used), Rotated solution, Loadings plot
Factor Scores	Save as variables (if PC scores for each case are required), factor score coefficient matrix
Options	Sorted by size, Supress small coefficients (set min = 0.3)

Interpretation:

Are the variables adequately related?

Correlation matrix: If r > 0.9 for two variables, they are too related and only one is needed. If one variable consistently has correlations under 0.1, it is not related enough to the other variables and is likely to form a PC of its own.

KMO (Kaiser-Meyer-Olkin Measure of Sampling Adequacy): Varies from 0 to 1 but the closer to 1 the better. Should be above 0.6 to use PCA.

Bartlett's test: Bartlett's test for sphericity tests the hypothesis that correlation matrix is an identity matrix (correlations between variables are 0). The test should be significant for PCA to be appropriate.

How many PC's/ factors should be used?

By default SPSS only retains PC's with eigenvalues above 1 (Kaiser Criterion) but this method sometimes selects more PC's than needed. The *Screeplot* is an alternative method for choosing the optimal number of PC's. It plots the eigenvalues for each PC. When the line plateaus, no more PC's are needed. Parallel analysis is an alternative method which is becoming popular http://pareonline.net/pdf/v12n2.pdf .

Other output

Total variance explained: If a rotation has been requested, the percentage of the common variance explained by the retained PC's is in the 'Extraction sums of squared loadings' section whereas it would contain the percentage of total variance if no rotation has been requested. The 'Rotation Sums of Squared Loadings' section contains the distribution of the variation after the Varimax rotation. The varimax (variance maximising) rotation of the loadings maximises the variability of the new PC whilst minimising the variance around the new variable.

Factor scores: The standardised PC scores for each subject will be added to the data set for use in further analysis or to identify subjects scoring highly on certain PC's. Producing a correlation matrix of the scores checks that the PC's are not related. If they are, re-run using the 'Direct Oblimin' rotation instead of the Varimax rotation. It is also useful to examine scatter plots of successive PC's to look for unexpected structure/ subgroups.

Communalities table: The 'Extraction' column contains the proportion of each variables variance explained by the extracted PC's/ factors. The highest coefficients are the strongest variables. Ideally, coefficients should be above 0.5.

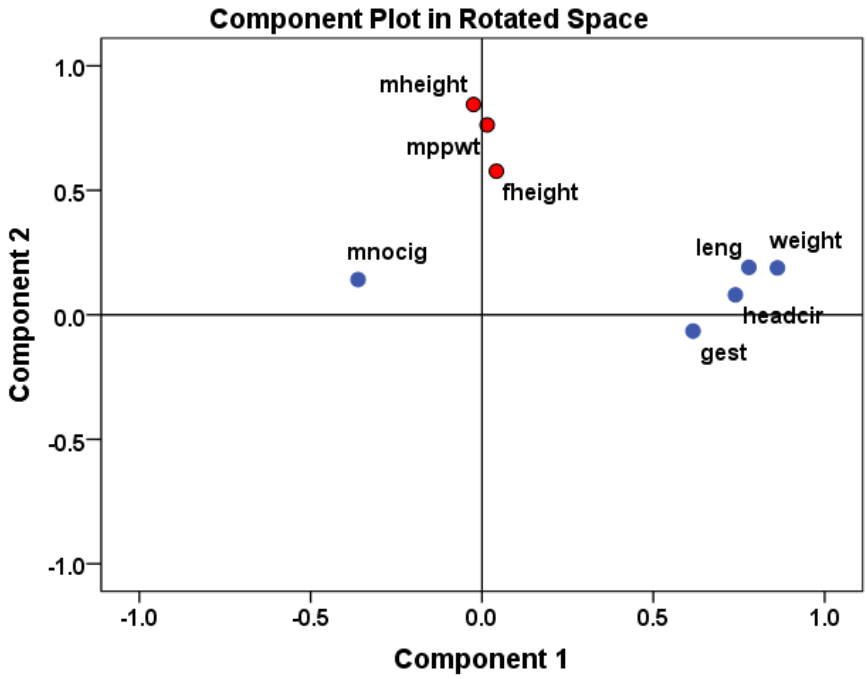
Component matrix: Shows the component loadings (correlations between individual variables and the PC's). Although the values can range between -1 and +1, we chose to suppress coefficients between – 0.3 and +0.3 and concentrate on those with higher loadings. Look at the variables for each PC with the highest loadings. Do similar types of variable all have higher loadings on one PC? The component plot is helpful when grouping raw variables together based on their loadings on each PC.

Component score coefficient matrix: Displays the coefficients for the linear combination of variables for the calculation of individual PC scores.

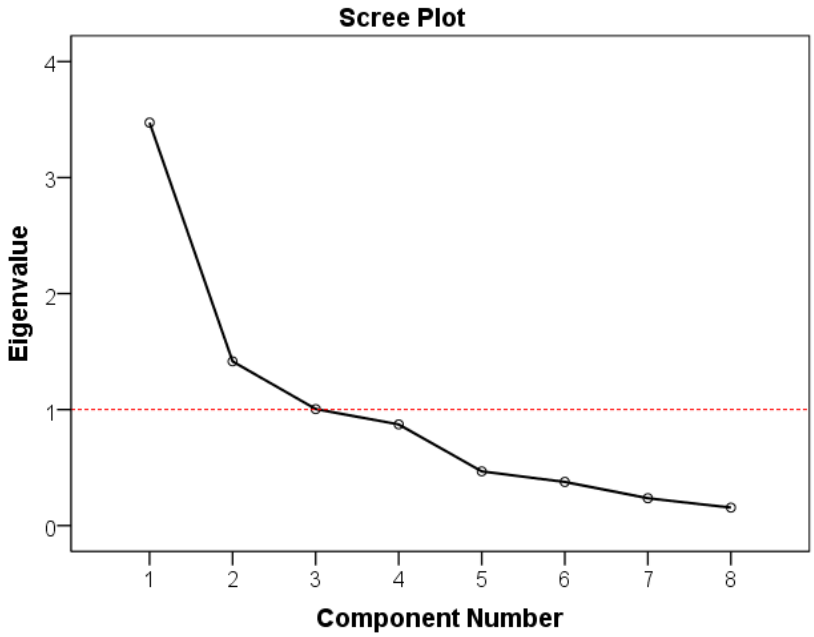
Example data set: Birth weight data set

The results of PCA on a data set containing information on 680 newborn babies and their parents are displayed below. Variables include Baby's length (leng), weight (weight), headcircumference (headcir), gestational age at birth (gest), and mother's height (mheight), pre-pregnancy weight (mppwt) number of cigarettes smoked per day (mnocig) and father's height (fheight).

The component matrix and plot suggest that PC 1 relates mainly to the measurements of the babies (length, weight, gestational age at birth and head circumference) although the number of cigarettes smoked by the mother has a negative loading. PC 2 relates to the height and weight of the parents.



The scree plot suggests that three PC's would be better than the two chosen using an eigenvalue cut off of 1. Further investigation found that the 3rd PC contained only mncig (number of cigarettes smoked by the mother per day) which had very weak correlation with the other variables.



Main references for this section: http://www.statisticshell.com/docs/factor.pdf
http://statistics.ats.ucla.edu/stat/spss/output/principal_components.htm

Exploratory Factor Analysis (EFA)

Use: There are two types of Factor Analysis. Exploratory Factor Analysis (EFA) aims to group together and summarise variables which are correlated and can therefore identify possible underlying latent variables which cannot be measured directly whereas Confirmatory Factor Analysis (CFA) tests theories about latent factors. Confirmatory Factor Analysis is performed using additional SPSS software and is beyond the scope of stats support but EFA is commonly used in disciplines such as Psychology and can be found in standard textbooks.

Exploratory Factor Analysis and Principal Component Analysis are very similar. The main differences are:

PCA uses all the variance in the variables analysed whereas EFA uses only the common (shared) variance between the variables
EFA aims to identify underlying latent variables (factors) rather than just reduce the number of variables

Dependents: Scale/ binary(preferably mostly scale or mostly binary)
Numerical ordinal variables are often included but the choice of numbering e.g. 1, 2, 3 will impact on the results

Factor analysis is commonly used on questionnaire data with likert style questions attempting to measure underlying latent variables such as depression which cannot be measured directly.

SPSS: Analyse →Dimension reduction → Factor

Run the analysis with two possible changes:

Extraction: Change the method to 'Principal axis factoring'
Rotation: Variamax assumes that the factors are not related. If it is likely that the underlying factors do correlate, use 'Direct Oblimin'

Interpretation:

Interpret the output in the same way as PCA although the Component matrix is called the Factor Matrix if the extraction method has changed.

Look at the Rotated Factor matrix to see which variables contribute most to each factor (PC). Variables measuring the same underlying latent variable should all have high loadings on a particular factor and by looking at the raw variables, a sensible name can be given to the factor. The next factor should be measuring another latent variable etc.

The factor plot is useful for assessing grouping of variables on more than one factor. If there are two factors, the variables appear on a scatterplot.

Main reference for this section: http://www.statisticshell.com/docs/factor.pdf

Cluster Analysis

Cluster analysis is a multivariate technique used to group individuals/ variables based on common characteristics. The groups are unknown.

There are three main types of cluster analysis:

Procedure	When to use
Hierarchical clustering	Small data sets of one data type (e.g. continuous) where different numbers of clusters are to be investigated. Both cases and variables can be clustered
K-means	Moderate sized data sets of continuous variables where the number of clusters (k) is specified. Several values of k can be run and optimal chosen.
Two-step	Large data sets or those with a mixture of data types

Hierarchical Clustering

Hierarchical clustering in SPSS is agglomerative (each subject starts in its' own cluster and then clusters are merged until all subjects are in one cluster). The way in which merging occurs depends on the way in which dissimilarity between individuals/ clusters is measured and the method for combining clusters. Once a subject has joined a cluster, it cannot leave. In SPSS, the variables need to be all of one data type.

SPSS: Analyse → Classify → Hierarchical Cluster

Select cases or variables to be clustered.

Statistics:

Proximity matrix (optional – a dissimilarity matrix of distances between subjects)

Range of solutions (a, b) – Output will be compared for a to b number of clusters

Plots: Dendrogram

Method:

Cluster method: Cluster method is how the clusters are merged. The single linkage method joins clusters with the smallest distance between two cases in different clusters (Nearest neighbour) whereas for complete linkage (Furthest neighbour), the distance between two clusters is defined as the distance between the two furthest points. The default method is Between-groups linkage (distance between clusters is the average distance of all data points within these clusters). Wards method (which can only be used with Euclidean measure), calculates the means of all variables within a cluster, then the squared Euclidean distance between cases and means are summed. Clusters are merged where the minimum

increase in sums of squares occurs. Wards is commonly used but susceptible to outliers. Single linkage will isolate outliers into groups which can be removed from the data set and the analysis run again using Wards.

Measure: this is how dissimilarities are measured using distances.

For continuous (interval data) and binary data use Squared Euclidean distance which is the sum of the squared differences between every pair of subjects. This is a dissimilarity measure. For ordinal (count) variables use either Chi-squared or Phi-squared.

Transform: If the variables are measured on different scales, those with larger values/ variances will dominate the distance calculations. Standardisation can be used although variables with more variation are often more likely to separate clusters.

Save: Here the cluster membership for each case is requested. This can be for a set number of clusters or several columns for several cluster numbers.

Interpretation

The main interpretation from the output regards choosing the right number of clusters. The **dendrogram** shows the cases joining (vertical lines) and the distances between clusters. The horizontal axis are the distances scaled to a range of 1 – 25 so are proportional to the actual distances. Long horizontal lines suggest that there has been a large change in the average distance within the cluster so choosing the clusters before the longest jumps is preferable.

The **Proximity matrix** contains the dissimilarity scores between each pair of subjects with small numbers indicating similarity between subjects. The **Agglomeration schedule** shows how the distance measure increases as additional cases are merged into clusters in the 'Coefficients' column. The first step joins the two subjects with the lowest dissimilarity then further steps either join two different subjects or merges a new subject into a formed cluster. The dissimilarity measure increases with the number of steps. Where there is a large increase in the coefficients value, the step merged clusters that were not very similar.

Finally, Hierarchical clustering does not distinguish between the groups but if group membership has been saved for the best number of clusters, descriptives can be calculated by cluster to see which characteristics differ.

K-means clustering

K-means clustering doesn't require a dissimilarity matrix for all pairs as each subject is assigned to the cluster with the mean closest to its value. The means are recalculated at each step (iteration) and subjects can be reassigned to another cluster at any step. The iterations stop when changes in cluster centres don't change much anymore or the maximum iterations are reached. K-means does require a set number of clusters to be specified in the beginning and initial cluster centres (means which are far apart) which are estimated by SPSS if not given. Some people carry out Hierarchical clustering on a sub set of the data to get an idea about the optimal number of clusters and starting cluster means and then carry out K-means. Only continuous variables can be included, Euclidean distances are used and K-means clustering is very sensitive to outliers which may form clusters of their own.

SPSS: Analyse → Classify → K-Means Cluster

In the K-Means menu, specify the number of clusters and request these extras:

Save: Cluster Membership

Options: ANOVA table

Interpretation: The *initial cluster centres* are given in the first table followed by the changes to cluster centres in the *iteration history.* The last row should show negligible change. The *final cluster centres* show how the variables differ in each cluster. It should be clear which variables are most different and therefore define each cluster but the *ANOVA table* shows which variables contribute most to the separation (highest F-statistics) and least. The F-tests should be used for descriptive purposes rather than formal tests as clustering is based on maximising the between cluster to within cluster variation.

Section 4

Suggested resources

Books

SPSS Survival guide. Julie Pallent.

A clear and fairly consise guide to performing the most commom tests in SPSS including assumptions, steps in SPSS and interpreting the output.

SPSS for Psychologists. Brace, Kemp and Snelgar.

We like this book because: It offers students quick examples of using SPSS to undertake statistical analyses and interpret the results.

Discovering statistics using SPSS. Andy Field.

A favourite in Psychology with interesting examples but quite wordy. Good for extra information not included in more consise books but the additional detail can distract from the main points.

Oxford Handbook of Medical Statistics. Janet and Philip Peacock.

Great summary guide covering a wide range statistical techniques and definitions.

Problem Solving: A Statistician's Guide. Chris Chatfield.

We like this book because: It provides ideas and examples of how to go about undertaking a statistical analysis. It also provides a quick overview of many different statistical analyses so students can see if that might be useful.

Elementary Survey Sampling. Scheaffer, Mendenhall and Ott.

We like this book because: It includes simple formula to calculate margins of error (and sample sizes for a required margin of error) from sample surveys and is especially useful where the population being studied is not large.

Multivariate Statistical Methods: A Primer. Bryan Manly.

We like this book because: It gives a good overview of multivariate methods that allows a student to assess whether these are useful. It does include some mathematics but should be accessible to anyone having studied mathematics modules as part of their undergraduate degree, but this could be omitted anyway for others and still be useful.

100 Statistical Tests. Gopal Kanji.

We like this book because: A great resource if you can't remember the details of a particular test. Also useful to find a test for less common situations.

Websites

Statstutor: ***http://www.statstutor.ac.uk/***

A growing collection of statistics teaching resources in different media formats.

STEPS glossary: ***http://www.stats.gla.ac.uk/steps/glossary/index.html***

A useful resource for quick definitions.

BrunelASK videos: ***https://www.youtube.com/user/BrunelASK***

These short videos, usually about using SPSS, were created by Christine Pereira at Brunel University. Once reviewed they will appear on statstutor.

CAST: ***http://cast.massey.ac.nz/collection_public.html***

A collection of computer assisted statistics textbooks including core statistics ebooks and apps for lecturers to use in class.

Statistics Fun ***http://www.youtube.com/user/statisticsfun***

A YouTube channel of statistics videos.

WhatTest: ***http://whattest.lboro.ac.uk***

A website aimed at new researchers who need help deciding on an appropriate analysis plan for a study or experiment.

Online Statistics Education: A Multimedia Course of Study: ***http://onlinestatbook.com/***

A teaching resource with apps for use to demonstrate techniques.

Contributers to the guide

This guide was produced primarily by the Maths and Statistics Help centre (MASH) at Sheffield University as part of a statistics tutors training project funded by SIGMA.

Main authors

Ellen Marshall (University of Sheffield)

Elizabeth Boggis (University of Sheffield)

Other contributers

Chetna Patel (University of Sheffield)

Marta Emmett (University of Sheffield)

Alun Owen (University of Worcester)

Reviewers (University of Sheffield)

Jean Russell and Nick Fieller (Multivariate)