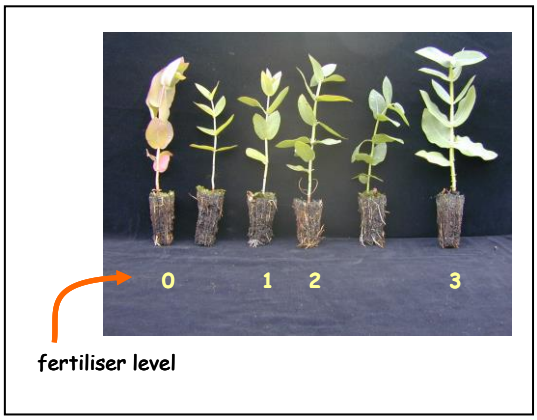
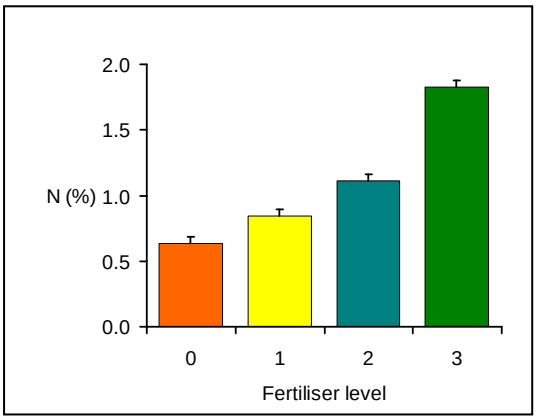
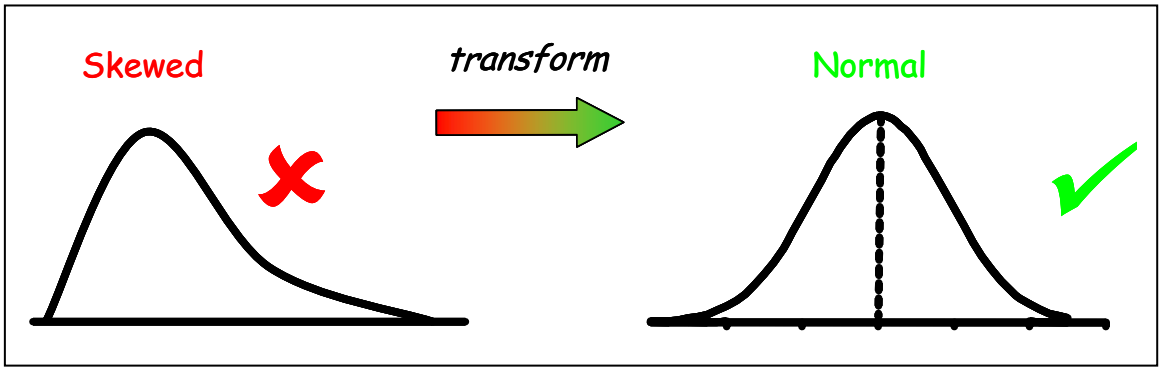
Clare's ~~very~~ Short Guide to Using SPSS

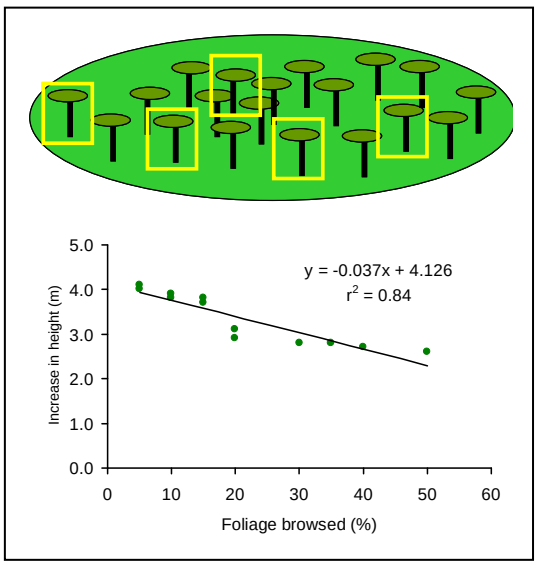
Note: This guide is based on the book:
Field, A. 2009. Discovering statistics using SPSS, 3rd edition. SAGE Publications, London. ISBN 978-1-84787-906-6
which you can buy on the web from Footprint Books http://www.footprint.com.au/.
Page numbers in brackets throughout this guide refer to this book.

But the instructions have been modified to suit later versions of SPSS (Version 21)









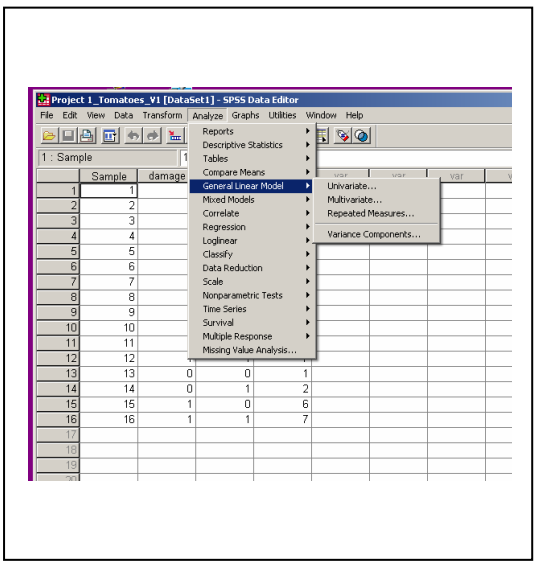


Table of contents

1 Some synonyms

Dependent variable = Outcome variable

Independent variable = Predictor variable

Class data = Categorical data = Nominal data

Continuous data = Scaled data = Interval data

2 Windows in SPSS

You will mainly use two windows:

(1) *Data editor* - contains your data set for analysis. This will either be typed in by you or imported from somewhere such as Excel, and will have any new columns of variables generated in SPSS, such as transformed variables and diagnostics.

(2) *Output Viewer* - output of stats analyses.

Use the *Window* Tab at the top of SPSS to move between the *Data editor* and the *Output Viewer*. If in *Output Viewer* you can also move to the *Data Editor* by clicking on the array symbol with the big red star.

There is a 3rd window:

(3) *Syntax editor* – use this window if you want to run the program using command lines rather than by drop-down and dialogue boxes (it is more complicated, but provides greater flexibility. You can quickly re-do analyses and modify the code for use on other databases)

When SPSS opens, it will be in the *Data editor*. If you have a particular data set already in SPSS (file xxx.sav), you click on "Open an existing data source" then *Open Data* and find the appropriate file or go to the top toolbar and use *File, Open, Data*. If you have output for a particular data set already (file xxx.spv), you open the *Viewer* and the file by *File, Open, Output*. Similarly if you want your syntax file, open it by *File, Open, Syntax* (file xxx.sps).

3 Input data from Excel into *Data Editor*

3.1 Summary

Three parts to this Step:

(1) import your data from excel, if you don't already have it in as an xxx.sav file

(2) define your variables and

(3) save your SPSS data file (xxxx.sav).

If you are importing from Excel, your Excel file:

- must be closed for SPSS to access it (or possibly open if it is a read-only file).
- must have one line (row) for each independent case, with all data connected to one case in columns along that single row.

- Class variables must be allocated a different number for each level (e.g. 0, 1, 2) NOT a name (e.g. male, female)
- Missing values can be blank
- can have headings for each column in the first row

3.2 Step 1 - Import your excel file:

In *Data Editor*, go to

- *File, Open, Data*
    - then in *Files of Type* make sure it says *All files* [or *Excel .xls*], so your excel files appear in the list.
    - Select the file you want to import and click *Open*.
    - You can then choose which worksheet within the file to import
    - If the first row of your file contains the headings, select that in the box "read variable names from first row of data" before you hit *OK*.

3.3 Step 2 - Define your variables

Go to *Variable View* and define the variables of the incoming data set in the order in which they will appear from your excel file (p 70). For each variable, you define:

- *Name* will be the label you have imported from Excel, but you can give it a more comprehensive (or a shorter) name under *Label*.
- *Type* (p 71) will most commonly be numeric. Even class variables must be *numeric* (e.g. 0, 1, 2) [SPSS creates dummy numeric variables for them].
- *Decimals*: will be what is imported, but you should probably adjust for each, e.g. class variable should have no decimals and 10dp are usually useless!
- *Label:* Give it a more comprehensive (or shorter) name than you gave in *Name*
- *Values:* This allows you to define each level of a class variable based on the value it has. A drop-down box will appear when you select the cell, and you must define each level that appears in your data set. E.g. *Value* = 1 then *Label* = "male" then click *Add*; *Value* = 0, then *Label* = female, then click *Add*. Then *OK* once you have defined all levels of that class variable.
- *Missing* (p71, 78): If there are missing values in a variable, you need to define them under *missing* by a number which is completely outside the range of the rest of the data set. E.g. if your data range from 0 – 200, define missing variables as "999" in the *missing* column.
- *Measure* (p 71): use the drop-down menu to select whether data are
    - *Nominal* (names) for class variables;
    - *Ordinal* (ranked data with no idea of intervals between ranks); or
    - *Scale* (for interval data e.g. measured with a ruler, counted accurately).

3.4 Step 3 – Save your SPSS data file

File, Save As etc – do this regularly, and update name (e.g. *File name_Version1*, *File name_Version2*, etc*)*

4 Irksome oddities about SPSS – tricks for avoiding hypertension

There are a few quirks about the program. You could spend hours trying to work out what on earth you have done wrong when it is actually just a result of these quirks.

Below, I have listed a couple of these quirks that I have come across. If you come across more, please let me know (clare.mcarthur@sydney.edu.au) so I can save other poor souls from tearing their hair out.

- Problem with particular Variable names – sometimes you may inadvertently name a variable something which confuses SPSS. This will stop the analysis from running.
    - Example: *Plot* – SPSSv21 seems to accept "Plot".
      Even though this is a commonly used term in ecological and agricultural field studies, old versions of SPSS (Version 15 and earlier, I think) have a tantrum and think you are asking it to plot something. Solution: DO NOT use "Plot" as a variable name. Call it something else, e.g. PlotID.
- Output tables – cells filled with ********* instead of numbers. This is just because the number has too many parts to fit into the output as you read it in SPSS.
    - Solution: (1) cut and paste the table into excel, and the ******** will now appear as numbers! Or (2) double-click on the table and expand the column using the mouse

5 Exploring your data

Before any stats test, you need to check to see whether the data conform to the assumptions of whatever statistical test you have in mind. E.g. if parametric test, this may (but not always) include conformity to a normal distribution and homogeneity of variance.

You can start this from either the *Data Editor* or the *Viewer* (many of the menus are the same), though the output will appear in the *Viewer*

5.1 Summary statistics for dependent variables

- *Analyse*, *Descriptive statistics, Descriptives*
    - *Options*: Use for getting means, s.d. etc for continuous variables
    - This has limited use and you should probably delve further by other means, e.g. while doing the actual analysis.

    - You may find the *P-P Plots* (p135) useful. Move your dependent variable in under Variables, and click ok. Your data should fall along the ideal diagonal line

- *Analyse*, *Descriptive statistics, Frequencies*
    - Use for more detailed summaries of continuous variables than provided by *Descriptives*
    - *Statistics* for summary: *Means, Mode, s.d., s.e. Kurtosis*, *Skewness* etc. Include *Mode* because it will indicate if data are bimodal or polymodal, which may be useful.
    - *Charts* for *Histogram with normal curve*. Also useful for indicating if data are bimodal etc.

NB/ If you want to split the file by group (levels within a class), so you can look at the descriptive stats and distribution within each group:

- *Data*, *Split file*, then split by whatever group you want using "organise output by groups" (p141). Run whatever stats you want on it e.g. the *Analyse*, *Descriptive statistics, Frequencies* with the data like this, then…
- TURN the split-data OFF once you have done whatever stats you want, otherwise SPSS continues with the split groups forevermore! Turn off by *Data*, *Split file*, then *Analyse all cases* or it "Reset".
- If you are testing effects of groups on some dependent variable, as long as each group is normal, doesn't matter if the combined data set is not (p147)

5.2 Check for homogeneity of variance and normality

Homogeneity of variance: For groups of data, SPSS uses (1) Levene's test (p97); for continuous data, you look at graphs. Levene's test (p150) tests the hypothesis that the variances in the groups are equal (so if $p < 0.05$ they are not homogeneous). This test is conservative, so you could get signif result with a large data set even if it is not very heterogeneous.

Homogeneity of variance: Levene's test: read the statistics based on the means (1st line of the table).

You can also look at the (2) Variance ratio. To do this, go to the Descriptives Table in the output and look for the groups with highest and lowest variance (if only two groups then don't have a choice but to use them). Take the ratio of largest over smallest variance; if variance ratio < 2, ok to assume homogeneity of variance.

Normality: You can use formal tests for normality – (1) Kolmogorov-Smirnov, (2) Wilk-Shapiro (p144). If $P > 0.05$ (n.s.), data are normal. NB. These are conservative tests; with large data sets can get signif (non-normal) without huge deviation from normal, so use these tests as a tool along with histograms, box plots.

The best way of checking homogeneity of variance and normality is to go straight to the following (p151):

- *Analyse*, *Descriptive statistics, Explore*
    - This is used when you have class independent data

- It is much more comprehensive than the info you will get from *Analyse*, *Descriptive statistics, Descriptives*
- Put the dependent variables of interest in, add the factor (group) if you want to split the data and look at each group separately
- *Statistics* click on *Descriptives* and *Outliers*
- *Plots*, then
  - in *Boxplots* click on *Factor levels together*;
  - in *Descriptives* click on *Stem and leaf plot* and *Histogram*; and also
  - click *Normality plots with tests.*
  - Under *Spread vs level with Levene test*, click on *Untransformed* first; then if necessary, try various transformations one after the other.

Check the output table showing the Test of Homogeneity of Variance for the Levene statistic and whether it is significant. If it is, you probably need to try transforming the (dependent variable) data

Normality: The Normal Q-Q plot should run along the expected line if normal (p147). If the observed points sag below or lie consistently above the predicted, then it suggests non-normal kurtosis. If there is an S-shaped curve around the predicted line, then it suggests skewness, in which case, try a transformation.

NB/ save your output file.

Formal tests for normality (p148) such as Kolmogorov-Smirnov or Shapiro-Wilk tests can be useful as a complement to a normal Q-Q plot, but they are conservative. So with large data sets they may detect deviance from normality that is not really important enough to worry about.

6 Transforming data

Should you have evidence of significant heteroscedasticity and/or problems of non-normality (for tests of hypotheses where these matter), then you should consider transformations as indicated in your lecture notes. If there is heterogeneity of variance (e.g. residuals against fitted or predicted values are wedge-shaped, not "cloud-shaped"), and / or data are not normally distributed then you may choose to transform (usually the dependent variable) before analysis (p 154)

In *Data Editor*,

- *Transform*, *Compute*
- *Target variable* – type in max. 8-character name for the new variable (e.g. t1VarY for square root transformed VariableY). Common transformations (reflecting increased "strength") are square root ($\sqrt{y}$), log ($\log_e(y)$) and inverse (1/y). Need special fiddling for some transformations of negative or zero values.

- *Type & Label* – define the new variable (type) and give it its full name (label)
- *Numeric expression* – look up the transformation you want e.g. under *Function Group*, select *Arithmetic* and select *Ln* for natural log transformation from the *Function & Special variables* set below. Hit the *up arrow* to place this function into the Numeric expression box. It will end up with empty brackets, within which you then place the variable you want to transform. Do this by selecting that variable from the list on the left hand side and hitting the *arrow across*. When the expression is complete, hit the *OK* button. A new column of data will appear in your database file in the data editor.

You can then test whether this has improved the normality and residuals (homogeneity of variance) by running through *Analyse*, *Descriptive statistics, Explore* again, this time using the transformed dependent variable.

7 Plotting your data

To plot data, you can use

- *Graphs, Chart Builder* (p91):
    - In the *Gallery* Tab, you can choose the graph style you want and build from there. To do this, double-click on the graph you want or drag it into the Chart Preview area
    - *Bar* – (p104) for continuous dependent variable and class independent variable(s). If you have one independent variable, as in a one-way ANOVA, choose the *simple* bar (p105). If more than one, choose the *clustered* bar (p107).
      Drag the variables into the *Chart Preview* region. Use *Element properties*, to display error bars (e.g. *s.e.* or *s.d.* or *95% CI,* p105).
    - *Scatterplot* for x-y plot for continuous dependent and independent variables. NB/ These plots can also be made directly during the analysis (e.g. during regression or correlation analysis) rather than having to do it specially here.
    - When your graph is designed the way you want it, hit *ok*. If not, use the *Reset* button to start again.

There are other ways of plotting data, depending on the stats test you want to do. These are described under each stats section.

8 Data analysis – overview

There are many statistical tests for analysing data, and what you do depends on the questions you are asking and the form of the data (whether your dependent and independent variables are class, counts or continuous). The following is a selection of tests which we commonly use.

9 T-tests – Independent and Paired

The test you use depends on whether the samples for each measure are independent (i.e. different individuals in two groups for the two measures) or dependent (i.e. the same individual has both measures taken on it).

9.1 Step 1 - Plot your Independent data

- go to *Graphs*, *Chart Builder*,
    - then choose *Bar*
    - drag the dependent and independent variable into the appropriate sections of the *Chart Preview* region.
    - Click Element properties to display and add error bars to your bars. Choose whichever you want (e.g. *95 % CI*, *s.d.* or *s.e.*). Hit ok
    - Can double-click on the *Output* graph to finesse further with the details (e.g. get appropriate font size and style on the axes)

9.2 Step 2 – Run the t-test for Independent means

(p324) Also called independent measures or independent samples t-test. Use this when you have a different group of samples for each of the two variables.

Parametric test, so assumptions are (p326) that data are:

- interval, i.e. continuous
- homogeneity of variance (between groups)
- scores are independent

Can check these at same time as doing the t-test (e.g. using Levene's test)

To run the test, go to *Analyse*, *Compare Means*,

- *Independent Samples T-Test*, then
    - Transfer your test (dependent) variable into the *Test Variable(s)* box
    - Transfer your independent (grouping) variable into the *Grouping Variable* box. This makes the *Define Groups* button active, so in this you need to provide the code for each group (e.g. 0, 1) then continue.
    - Hit *OK*.

In the Output,

- check the Group Statistics Table to get the means [and s.d. and s.e.] for each group
- then check the Independent Samples Test Table to find out if the difference between the two groups is significant or not. To know which test to look at

- Check results for Levene's test (see Part 5.2 above) – if significant (P<0.05), then variances are significantly different between groups. In this case, use the adjusted t-test in which variances are not assumed to be equal.
- If the Levene's test is Not significant (P>0.05), then continue using t-test for when variances are assumed to be equal.

9.3 Step 1 - Plot your Paired data (i.e. repeated measures data)

You first need to eliminate the between-subject variability [p319-324], which is not a true reflection of the data error, and which would otherwise be included incorrectly in the error bars if you simply plotted the means and errors as for independent group data. To do this, you normalise the participant means in four Steps:

- Step 1: Calculate the Mean for each Sample, which is the mean of the two variables (e.g. before vs. after; CHO diet vs. PRO diet), by *Transform*, *Compute*, then
  - Type *Mean* (name of new variable) into the *Target Variable* box.
  - Under the *Functions Group* box, select *Statistical*. Then from the *Functions and Special Variables* box, move the *Mean* function up into the *Numeric Expression* box, then
  - Move the two variables at the left into the *Numeric Expression* box in the appropriate spot
  - Hit *OK*. This produces a new column of data, with a variable called Mean, in the *Data Editor* window for each sample.
- Step 2: Calculate the Grand Mean, which is the mean of all scores. This can be done by calculating the mean of all the sample means just calculated in Step 1. To do this: select *Analyse*, *Descriptive Statistics, Descriptives*, then
  - Transfer the Mean variable from LHS into the *Variables* box, then
  - Open *Options* and select only the *Mean*
  - Hit *OK*. This will give you a *Descriptive Stats* box in the *Output Viewer*, and you can use this Grand Mean in the next Step.
- Step 3: Calculate the adjustment factor. To do this, we subtract each sample's mean from the Grand Mean, by *Transform*, *Compute*
  - Type *Adjust* (name of new variable) into the *Target Variable* box.
  - Type the value of the Grand Mean (from the *Output Viewer*, e.g. 8.35) into the *Numeric Expression* box, then create the equation for: Grand Mean – (minus sign) and move the Mean variable at the left into the *Numeric Expression* box in the appropriate spot
  - Hit *OK*. This produces a new column of data, with a variable called Adjust, in the *Data Editor* window for each sample.

    - We now use these adjustment values to eliminate the false between-subject differences in the two variables of interest (i.e. before vs. after; CHO diet vs. PRO diet)

- Step 4: Create adjusted values for each variable, by adding the adjusted mean for each sample to the value of one variable, then the other variable, using *Transform*, *Compute* again:
    - Type *A_variable1* (e.g. for variable CHO) into the *Target Variable* box.
    - Define this new variable ore clearly using the *Type and Label* box, e..g describe it as the "adjusted variable 1".
    - In the *Numeric Expression* box, then create the equation for: *Variable 1* (e.g. CHO) + (i.e. plus) the *Adjust* variable (created in Step 3 above).
    - Hit OK. This produces a new column of data, with a variable called *A_variable1*, in the *Data Editor* window for each sample.
    - Repeat the procedure for the second Variable (e.g. PRO) to create a new adjusted variable (e.g. *A_variable1*).
    - The two new adjusted variables now represent the values adjusted to eliminate any between-subject differences (which shouldn't be there because the data for each sample are within-subjects data). To check this is correct, you can compute a new variable (new column), which is the mean of the two new variables. This new column should show you the Grand Mean which you calculated in Step 2! (if not, you have made a mistake somewhere along the line).

Finally, after adjusting as above, you can plot the graph:

- to go to *Graphs*, *Chart Builder*
    - see 9.1 for ideas

9.4 Step 2 – Run the t-test for Dependent means (Paired)

(p329) Also called matched-pairs or paired samples t-test. Use this when you have used the same group of individuals for each of the two variables (i.e. measured something twice on each).

Parametric test, so assumptions are (p326) that data are:

- interval, i.e. continuous

You can test whether there is a significant difference between two (paired) variables in two ways:

(1) as a paired t-test

Go to *Analyse*, *Compare Means* (p329-330),

- *Paired Samples T-Test*, then
    - Transfer your paired variables into the *Paired Variables* box, by selecting both at once (press control down simultaneously to do this)

- In *Options*, make sure the *Confidence Interval* = 95 %
- Hit *OK*.

In the Output,

- check the Paired Samples Statistics Table to get the means [and s.d. and s.e.] for each variable
- then check the Paired Samples Test Table to find out if the difference between the two is significant or not, with the t-statistic, d.f. and the P-value[Sig. (2-tailed)].

(2) as a single population test of the difference between the variable (compared with zero)

- Create a new variable which is the difference between the two variables of interest, using *Transform, Compute*
- Go to *Analyse*, *Compare Means*, *One-Sample T-Test*
  - Move the new "difference" variable into the *Test Variable(s)* box and hit *OK*.
  - The results are essentially identical to that of the Paired-t-test (and should be!).

9.5 Step 1 - How to finesse with your Plots

It is often best to present data so the y-axis goes through zero rather than simply bounding the data (otherwise the proportional difference between groups can look misleadingly big). SPSS doesn't seem to do this by default.

- In order to change the scale of the y-axis:
  - double-click on the graph to bring up the *Chart Editor*.
  - Double-click on the y-axis to bring up *Properties* and
  - Under *Scale*, change the minimum to zero (if data are all positive!)
  - Under *Number format* – change the number of deco=mal places, if you need, to something sensible if needs be.
- In order to change the axis title
  - Remain in *Chart Editor* (or get back in with a double-click)
  - Click once on the axis title, then again slowly, then change text as you wish
- In order to change other details on the graph, such as the font size of the words on the x-axis,
  - Right-click on one of the words, then select the *Properties Window*
  - Change details as you wish (e.g. under *Text Style* tab, change the font size to 12 or whatever)

9.6 Step 2- Non-parametric alternatives to Independent and Paired t-tests

The non-parametric equivalent:

For independent t-test = Wilcoxon rank-sum test or the Mann-Whitney test

- Go to *Analyse*, *Non-Parametric Tests*, *Independent Samples* Test
    - 1st check your *Objective*
    - Then go to the *Fields* tab and drag the continuous dependent variable into "Test Fields" and the class independent variable into "Groups"
    - Go to the *Settings* tab to choose which test you want to run and hit Run.
    - Go to the Output and double-click on "Hypothesis Test Summary". This gets you to the "Model Viewer" where you will find the details of the tests (e.g. the Mann-Whitney statistic etc)

For paired (dependent) t-test = Wilcoxon signed-rank test

- Go to *Analyse*, *Non-Parametric Tests*,
    - *2 Related Samples* Test

10 1-way ANOVA

10.1 Step 1 - Plot your data for 1-way ANOVA

- Exactly as for plotting data for Independent t-test
- go to *Graphs*, *Chart Builder*
    - then choose *Bar* and continue as described in Section 9.1.

10.2 Assumptions of 1-way ANOVA

Parametric test, so assumptions are (p359-360) that data are:

- interval, i.e. continuous
    - sometimes ok if dichotomous (but better to use a different test)
- homogeneity of variance (between groups) – can transform if not, and try again, but be aware that
    - the test is robust when sample sizes are equal
    - the test is not robust when sample sizes differ – When groups with larger sample sizes have larger variances, then the test is conservative i.e. less likely to detect a difference even if there is one. Converse true too, i.e. groups with larger sample sizes have smaller variances, then more likely to find a difference even if there is NOT one
    - violations of this assumption can be corrected
- scores are independent

Can check some of these assumptions at same time as doing the ANOVA.

10.3 Planned versus Post-hoc pairwise comparisons

For planned contrasts versus post-hoc pairwise comparisons, see p360-364.

- Post-hoc pairwise comparison (p370-374)
    - there are various adjustments you can make to take into account the number of comparisons (i.e. to avoid a Type 1 error).
    - P374 gives a good summary of which test suits which conditions
    - SNK – is very liberal and doesn't control the Type I error
    - Bonferroni & Tukey's – both control the Type I error but are conservative (less likely to detect a difference even if there is one). Bonferroni has more power when the number of comparisons is small.
    - SNK, Bonferroni & Tukey's all perform badly when group sizes are unequal and population variances are different. IN these cases, better to include the Games-Howell procedure too.

10.4 Step 2 – Run the 1-way ANOVA and check the diagnostics

To do this (p375): go to *Analyse*, *Compare Means*,

- *One-Way ANOVA*, then
    - Transfer your test (dependent) variable into the *Dependent List* box
    - Transfer your independent variable into the *Factor* box
    - If you have more than two groups (i.e. levels for the one factor), the under *Post-Hoc* (for post-hoc pairwise comparisons), choose which type of test (e.g. Tukey's)
    - Under *Options, Statistics*, select *Descriptive* and *Homogeneity of Variance test*. Note this last one does the Levene's test here for you, so you can test this assumption here, then if necessary based on the result, transform the dependent variable and re-run the analysis.

11 2-way ANOVA

11.1 Step 1 - Plot your data for 2-way ANOVA

- Similar to plotting data for 1-way ANOVA
- Go to *Graphs*, *Chart Builder*,
    - then choose *Cluster*
    - Under *Data in graph are*, choose *Summaries for groups of cases* (*Summaries of separate variables* are for paired t-tests).
    - Under *Define*, choose move your dependent variable into *Variable*, and your independent variable(s) into *Category Axis*.

- Under *Bars Represent*, choose whichever you want (e.g. *95 % CI*, *s.d.* or *s.e.*)

11.2 Step 2 - Run the 2-way ANOVA and check the diagnostics

- go to *Analyze*, *General Linear Model, Univariate* (p431)
  - make sure you put the independent variables in the right place i.e. Fixed or Random (usually in a classic ANOVA they would both be Fixed factors)
  - then select bits as makes sense!
    (see 1-way ANOVA section for general tips)
  - Go to *Options* (p434). Important to select the independent grouping variable(s) of interest and move them into "*Display Means for*" box. Then under *Display* select *Descriptive* stats, *Homogeneity tests* and *Residual Plot* so you can get the final values for plotting and check your data again.
  - Under *Plots* (p400) you can place one independent variable into *Horizontal Axis* and the second into *Separate Lines* or *Separate Plots* (depending on what you want to display and how). Click *Add* so the plot you want appears in the *Plots* box below.

12 Correlations

12.1 Step 1 – Plot your data

- *Graphs, Chart Builder, Gallery Tab - Scatterplot* (p 117 - 119).
  - You can finesse with your plot, by double-clicking on it in *Output Viewer* to get to the *Chart Editor*.
  - Double-clicking on the symbols o=in the toolbar at the top of the graph will allow you to add and modify things on the graph (e.g. add a linear regression line if appropriate).

12.2 Step 2 - Run the correlation analysis

By either (1) *bivariate* [between 2 variables; e.g. Pearson's correlation coefficient (parametric), Spearman's (non-parametric) rho] or (2) *partial* [relationship between 2 variables with other variables fixed] (p 175).

- *Analyse*, *Correlate, Bivariate*
  - You can generate a correlation matrix with this for a number of variables, giving pairwise correlations for all variables included (i.e. can be > 2 as each pair is done one at a time). This is a useful step for exploring independent variables before multiple regression, because you can then decide which are so tightly correlated that they should not both go in the model.
  - If you click *Pearson's* (r), then your data must be intervals (scaled numeric) and normally distributed (it's a parametric test) (p 177). You

can square the r-value to indicate the % of variation in one variable "explained" by the other.

- *Spearman's* correlation coefficient (r_s) is non-parametric (p 179).
- *Kendall's* tau – also non-parametric. Use in place of *Spearman's* when you have small data set with a large number of tied ranks (p 181).

Or, if you want to test the correlation of two variables while controlling for variation of another variable:

- *Analyse*, *Correlate, Partial*, then see p186-190

13 Simple Linear Regression

This is used to test the linear relationship between one continuous variable and another. You need to have checked your assumptions (normality and homogeneity of variance etc) using the diagnostics of the analysis before accepting the results. You can do this by running the analysis and checking the diagnostics. If the assumptions are not met, revise (e.g. transform the data) and re-run the analysis on the revised variables. Check the revised diagnostics.

13.1 Run the regression and check the diagnostics

- *Analyse*, *Regression, Linear*, see p205-209
  - see the Multiple Linear Regression section for further info on the points made below (p209-212)
  - Select one dependent and one independent variable, keep *Method* as the default *Enter*.
  - *Statistics*: we don't need to look at as many things as for multiple regression (in fact, we can't). But mark *Model Fit*, and under *Regression Coefficients*, mark *Estimates* and *Confidence Intervals*, and under *Residuals* mark the *Durbin-Watson* test and change the *Outliers* to **2** s.d. not the default 3.
  - Do two things in *Plots*:
    (i) under *Standardised Residual Plots*, click both *Histogram* and *Normal Probability Plot*..
    (ii) the **SRESID* (y-axis) vs **ZPRED* (x-axis) scatter plot.
  - Then hit *OK*

13.2 Comments on some of the output

In the output (SPSS *Viewer* window) (p207-208). For further explanation of some of this, see Multiple Linear Regression notes.

- the *Model Summary* Table gives the R^2 and the *Durbin-Watson* statistic (worry if the value is < 1 or > 3)
- the *ANOVA* Table gives the F-test of the regression model. (Is the model with the independent variable in it better than a model using simply the mean?)

- the *Coefficients* Table gives the parameter estimate for the intercept (B-value for the constant with s.e.); and for the independent variable (B-value with s.e.).
    - These values are used to describe the mathematic equation [$Y = ß_0 + ß_1X_1$] between the dependent and independent variable.
    - E.g. if $ß_0 = 6.2$ (intercept) and $ß_1 = 4.3$ (slope), then $Y = 6.2 + 4.3*X_1$
- the *Casewise Diagnostics* Table (p215) gives any individual case whose residual lies outside 2 s.d. of the mean. This is useful to check back and make sure you have not made a typo in putting the data in, and to consider what influence any such case it may have on the results.
- The *Plots* are important for assessing your assumptions:
    - *Histogram* – do the standardised resids look normal?
    - *Normal Probability Plot* – do the data run roughly along the diagonal line?
    - *Scatterplot* – is there any pattern in the standardised residuals plotted against the fitted (predicted) values? This is a check for homoscedasticity (which you want!) – If the resides fall between -2 and +2 in a "cloud" then the data are fine. If there is any pattern (e.g. wedge-shape indicating larger variance with larger values), then the data are heteroscedastic and may need transforming). There's a good description of how to use and interpret the standardised residuals (when plotted against the fitted [i.e. predicted value) on p216.
- Once you have checked the diagnostics – accept the analysis for the regression if they are ok (i.e. fit the assumptions), if not, revise the analysis (e.g. transform dependent data) and try again.

13.3 <u>To plot the Regression with 95% C.I.</u>

To get a nice plot of your regression:

- *Graphs*, *Chart Builder, Scatter/Dot* double-click on the simple scatterplot (top left under the Gallery tab.
- Assign variables by moving
    - Dragging the Dependent Variable → y-axis dashed box on the graph
    - Dragging the Independent Variable → x-axis
    - Hit *OK* , then double-click on the graph to modify it in the *Chart Editor*
    - Click on the *Add line* little diagram (5^{th} graph from left in bottom toolbar in *Chart Editor*) to get the *Fit Line* tab in the *Properties* box.
    - Under the *Fit Line* tab, the *Fit Method* should be "linear" and under Confidence Intervals, mark Mean with 95% CI (confidence intervals). Hit *Apply*.

14 Multiple Linear Regression

14.1 Check correlations between independent variables

You need to have checked correlations between independent variables (if $r \geq 0.7$, choose only one of the two variables, see Tabachnick, B. G., and L. S. Fidell. 1989. Using Multivariate Statistics, 2nd edition. HarperCollins Publishers, New York.).

14.2 Run the regression

Then, to test the relationship between one continuous dependent variable and several continuous independent variables, there are several options (p 212-214):

(1) Hierarchical entry – predictors based on past work and you decide which order they go in (in order of their assumed importance then unknowns)

(2) Forced entry (*Entry*) – all predictors go in together, you make no decision about the order of entry

(3) Stepwise – various (Forward, Backward, Stepwise) – for statistically exploring which variables may be important predictors, not for hypothesis testing unless you have already chosen your subset of variables for good theoretical reasons. Choose Backward over Forward to avoid suppressor effects (p 213) and hence avoid Type II errors for a particular predictor.

You can use class variables in a regression, but if you have > 2 levels within a class, you will need to generate dummy variables (see p 253-255). If you have class and continuous independent variables, this is essentially a general linear model.

14.3 Check the diagnostics

To check the accuracy of your model, you need to check various diagnostics (p 241-251) about (1) residuals AND (2) influence of data points. You MUST check these if ever you run a regression.

- *Analyse*, *Regression, Linear* (p225 onwards)
    - Move the *Dependent* variable over into the right box
    - For the *Independent* variables, you need to have decided on the form of the model and *Method* of inclusion.
      Example 1: in one block with all variables, using one defined method e.g. forced, stepwise.
      Example 2: If you are doing a hierarchical model, you need to build each of the independent variables into a new block. in several blocks, each with a certain type of method which may differ between blocks. It's quite nice to do this, because then the output summarises the model as it builds it up and you can see the Change in R^2 as each block is incorporated. To do this, you put whichever *Independent variable* you want into the 1st Block, then click *Next* and then add the next (set of) Independent variable(s) into the 2nd Block and so on. Switch between blocks using the *Next* and *Previous* buttons.
    - *Statistics*: in general mark all except the covariance matrix (what each does is explained p 227-229). For the *Residuals* bit, change the *Outliers* to **2** s.d. not the default 3 (p 180).
      The *Durbin-Watson* test tests for serial correlations between errors (p

220, 229). However, SPSS does not say whether the statistic is significant or not, so you have to decide yourself. Values can range from 0 – 4, with 2 = not correlated. As a general rule, worry if the value is < 1 or > 3.

- *Plots*: you could do hundreds!
  At the minimum, do three things:
  (i) under *Standardised Residual Plots*, click both *Histogram* and *Normal Probability Plot*. These appear first in the output and are used to check the normality of the residuals. For the NPP, values should fall along the line (more or less).
  (ii) the **ZRESID* (y-axis) vs **ZPRED* (x-axis) scatter plot. This gives a plot of Standardized Residuals versus Standardised Predicted Values of the dependent variable. This appears second in the output and is used to check heteroscedasticity of the residuals ("cloud-shaped" = homoscedastic which is good, a pattern such as wedge-shaped is heteroscedastic and is bad. Standardized residual values (y-axis) mainly should fall between –2.5 and +2.5 (p216). If > 1 % fall outside this range, then the model is a fairly poor fit of the sample data. NB/ Standardised and Studentised residuals are similar but not identical (p217)
  (iii) *Produce all partial plots* (p230) – this plots the residuals of the dependent variable against residuals of each of the predictor variables in the model. These plots are useful because:
  (a) gradient of the regression line = coefficient of the predictor, so obvious outliers represent datum which may have large influence;
  (b) can detect non-linearity more easily,
  (c) useful for detecting co-linearity.
  To do other scatterplots, click *Next*, then add your next 2 variables for plotting.
- *Save*: Click on *Mahalanobis* AND *Cook's* distances. This will calculate these values for each case and put them as new columns in your Data file. You can delete them after you've looked at them if you wish.
- *Options*: IF you are using a stepwise regression model, use this for setting entry and exit criteria.

14.4 Comments on some of the output

- *Descriptive Statistics* Table (P233) – summary only, doesn't help with interpreting the regression
- *Correlations* Table – important to confirm no strong co-linearity between independent variables (should have been tested before you put them in the model anyway)
- *Model Summary* Table*:* (P 234-235) Note that if you blocked your independent variables, then the number of models will equal the number of blocks, and each new model adds the next block into the model. In this output, you are interested in:

- *R^2 value*. [NOT so much the Adjusted R^2, which gives you some idea of how well the model can be generalised. For further details see p 235. Not super important to us.]
- *Durbin-Watson statistic* - tests for serial correlations between errors (p 220, 229, 236). However, SPSS does not say whether the statistic is significant or not, so you have to decide yourself. Values can range from 0 – 4, with 2 = not correlated. As a general rule, worry if the value is < 1 or > 3.

- *ANOVA* Table: (P 237-238) tests whether the model is significantly better at predicting the outcome than using the means as a best guess.
- *Coefficients* Table*:* (p 238-240) this looks at the parameters of the model and so is important!
  - *B-value* - gives the value of the parameter (β_0, β_1, β_2 etc) and the s.e. of this estimate.
  - *t-value* and associated P-value (*sig*.) – tells you whether the variable contributes significantly to the model.
  - *Standardised Beta coefficients* are directly comparable between variables and so can be used to see if the variables have a comparable degree of importance for the model (similar values if they do).
  - *Collinearity statistics*: Important for assessing the assumption of no collinearity (p241-242).
    VIF (Variance inflation Factor, p242) – if largest VIF > 10, worry; if the average VIF >>1 then the regression may be biased.
    Tolerance statistic (= 1/VIF) serious problem if < 0.1, potential problem if < 0.2.
- *Excluded variables* Table*:* see p241 in hierarchical or stepwise models this describes the variables that have not (or not yet) been included in the model. Perhaps useful to see whether it would be good to include.
- *Collinearity Diagnostics* Table: (p242) – complicated. Not essential if you have had a good look for collinearity BEFORE running the model and already excluded co-linear variables (i.e. chosen just one of them).
- *Mahalanobis* and *Cook's distance value* – these will have been added to your Data file in the *Data Editor* (NOT the *Viewer*) as 2 new columns. Check through.
  - *Mahalanobis* (p247) – a measure of leverage of any particular case. Importance depends on sample size and number of predictors in the model. E.g. Rule of thumb: for n = 30 & 2 predictors, a value >11 is a worry; for n = 100 & 3 predictors, a value >15 is a worry.
  - *Cooks'* (p247) – a measure of the overall influence of a case on the model. Rule of thumb: if >1, worry.
  - You can delete these columns after checking them, or save the data as an upgraded file (e.g. File name_Version 2).
- Plots (good examples p 248 & 249):

- Check the normality of the residuals: with the *Histogram* and the *Normal P-P Plot of Regression Standardized Residual*. For the NPP, values should fall along the diagonal line (more or less).
- Check for heteroscedasticity of the residuals with the plot of Studentized Residuals against Standardised Predicted Values of the dependent variable. Results should look like a cloud and have no pattern (e.g. wedge-shaped is bad!) and Studentized Residuals should fall between -2 and +2.
- If these data look non-normal, try a transformation (usually of the dependent variable) and run the analysis again.
- NB/ distributions can look very non-normal in small samples even when they are (p251).

15 Logistic Regression

15.1 Logistic versus multiple linear regression

Logistic regression is like multiple (linear) regression but with an outcome variable that is a categorical dichotomy and predictor variables that are continuous or categorical – we can predict which of the two categories (e.g. dead or alive, present or absent, male or female) a case is likely to be, based on certain other information (the predictor information) (p265).

The equation (p266) for the probability of Y occurring is:

$$P(Y) = \frac{1}{1 + e^{-(B_0 + B_1 x_1 + ... \varepsilon_i)}}$$

You can see it has a somewhat similar form to the multiple linear regression model equation:

$$Y = B_0 + B_1 x_1 + ... \varepsilon_i$$

Values of the parameters in the model are predicted, not by least-squares, but by maximum-likelihood estimation. This selects coefficients that make the observed values most likely to have occurred (p267).

In multiple regression we see how well the model fits the data by comparing predicted and observed values of the outcome, using R^2 (Pearson correlation between observed and predicted). In logistic regression, we also use the observed and predicted values to assess the fit of the model, but use a measure called the Log-likelihood value. This is the sum of the probabilities associated with the predicted and actual (observed) outcomes. The log-likelihood (LL) statistic (p267) is analogous to the residual sums of squares in multiple regression because it indicates how much unexplained information there is after the model has been fitted. Therefore LARGE LL values indicate poor fit.

To test whether a model which includes independent or predictor variable(s) is significant, we compare its LL statistic with the baseline model LL, for details see p268, and get a χ^2 value, where:

$$\chi^2 = 2[LL_{(\text{New model})} - LL_{(\text{Baseline model})}]$$

$$(df = k_{new} - k_{baseline})$$

As with multiple regression, we can also see how much each predictor variable contributes to the model. To do this, check the Wald statistic (equivalent to the t-statistic for each variable in linear regression) (p269).

For interpreting the effect of each predictor variable, the value of Exp(B) is important (p270-271). It is an indicator of the change in odds resulting from a unit change in the predictor. If the value >1, then as the predictor increases, the odds of the outcome occurring increases. If <1, the odds decrease as the predictor increases.

15.2 Run the regression and check the diagnostics

- *Analyse*, *Regression, Binary Logistic* (p278 onwards)
    - Move the *Dependent* variable over into the right box
    - For the *Independent* variables:
        - you need to have decided on the form of the model and *Method* of inclusion.
          Example 1: *Enter* – which is forced entry in the order you present them (i.e. testing an hypothesis, rather than doing an exploratory analysis).
          Example 2: *Forward LR* is forward stepwise regression using Likelihood Ratio. Useful for seeing how the model could be built up.
        - You move them into the *Covariate* box on the right, then
    - Use the *Categorical* box at the top to define if a particular independent variable is categorical. For the *Change contrast* section, keep the *Indicator* default for the *Contrast* and change the *Reference Category* to *First* (p279-280 for explanation).
    - *Save*: this box lets you save various residual variables to examine how well the model fits the observed data (p280). Select both boxes in the *Predicted Values* section (Probabilities and Group Membership) and all in the *Influence* section. The *Predicted Probabilities* are the probabilities of Y occurring given the values of each predictor for a given case. The *Predicted Group Membership* then predicts which group a case fits under given this information. Select at least the *Standardized* box in the *Residual* section. You must examine these residuals after the analysis to see if they are ok!!
    - *Options*: Under *Statistics and Plots*, select *Classification Plots*, *Hosmer-Lemeshow goodness-of-fit*, *Casewise listing of residuals* (for the latter, make sure it is checked for *Outliers outside* **2** *s.d*.), and *Iteration History*. You can also select the *CI for exp(B)* [make sure they are 95% CI].]. Also tick *Include constant in the model*.

15.3 Comments on some of the output:

- *Block 0* – baseline model without any predictor variables.

- In the *Classification Table*, it automatically allocates all cases to the most common outcome (e.g. if more cases have germinated than not, then all cases will be assigned to the "germinate" category), so any that did not germinate are assigned incorrectly). Obviously if your independent predictors are valuable at predicting the probability, then when they are incorporated into the model (in Block 1 information etc), then you should see more cases correctly predicted.
- Useful to look at the table showing "*Variables not in the equation*" (p284). If the *Overall Statistic* [actually the residual Chi-sq statistic, termed the *Score,* at the bottom of the table < 0.05, then it means that some or other of the excluded variables would have improved the fit of the model if they had been included (so in stepwise the procedure would continue, and in the forced entry procedure, you will see the improvement). You can also look at other predictor variables and their *Score* value (p284, Roa's efficient score statistic, which is similar to the Wald statistic and used for a quick run). If this Score statistics is significant for a predictor variable which is not yet in the model equation, then it suggests that that variable will be useful as a predictor.

- *Block 1* –model with predictor variables included – <u>this is the important bit!</u>
  - The *Omnibus Test of Model Coefficients* Table, gives the Chi-sq value for the *Model* and its significance. If signif (< 0.05) then this model, including the predictor variables, is better than the baseline model. So check this value! The Chi-sq test is equivalent to the F-test for the linear regression sums of squares (p285)
  - The *Model Summary* Table gives the log-likelihood for the model (actually -2LL which has a Chi-Sq distribution so can be checked for significance using a Chi-sq table). If this value is (significantly) smaller than the -2LL for the baseline model, then the model is better with the predictor variables included than without.
  - The *Hosmer and Lemeshow Test* Table indicates whether the model is a reasonable fit of the data. If the statistics is N.S., then the fit is reasonable.
  - The *Classification Table* (p286) indicates how well the model predicted both possibilities. It is possible for the model to be better at predicting one outcome than the other. **Sensitivity** is the proportion of true positives, i.e. the proportion of cases correctly identified by the test as meeting a certain condition (e.g. success). **Specificity** is the proportion of true negatives or the proportion of cases correctly identified by the test as not meeting a certain condition (e.g. failure).
  - The *Variables in the Equation* Table (p286) is crucial because it tells us the estimates for the coefficients for the predictors included in the model. The *b-values* are the estimates of the parameters (which you can then place into the model equation) and the *Wald Statistic* tests whether it is significant or not. If significant, it means this predictor variable is a useful predictor of the outcome.

- Remember that the equation for the probability of Y occurring is:

$$P(Y) = \frac{1}{1 + e^{-(B_0 + B_1 x_1 + \ldots \varepsilon_i)}}$$

so you can plug the parameter estimates for the constant (B_0) and the predictor variables (B_1 etc) in there!

- The *Casewise List* Table indicates any case which fell 2 s.d. outside the model and so should be scrutinised in case there was an error in the input.
- If you return to your data set, you'll see that new columns of info have been added against each sample – including the predicted probability based on your model. You can use this to plot your data and/or compute plots for the individual independent variables while keeping other independent variables fixed (e.g. at their mean or median level).

- Examining the residuals – these are diagnostics that you need to look at to (1) isolate any cases which the model fits poorly and (2) isolate points that exert undue influence on the model. To do this go to the *Data Editor* window
  - check the new columns of values for each case.
  - *Cook's D* (distance) (p217, 293) – Should be < 1; if >3, then the case is having a reasonably large influence on the regression coefficient.
  - *DFBeta* is the standardised version of Cook's D – values > 1 indicate possible influential cases (p293).
  - The *Leverage statistic (LEV_1)* (p293) should lie between 0 (the case has no influence at all) and 1 (the case exerts complete influence over the model). The average or expected leverage is calculated as the number of predictors [not including the intercept] plus 1, divided by the sample size. If any values are 2x or 3x greater than this, the data point could be having a large influence.
- Overdispersion: Logistic regression assumes the variance = the mean, i.e. (NOT homoscedastic) so as the mean value increases so does the variance. To check for this, calculate the dispersion parameter by dividing the Chi-Sq statistic for the model by the degrees of freedom. If >1 (by a bit), this is overdispersion and results in possible Type I errors (p276). There are complicated ways of dealing with this (using other link functions, beyond the scope of this guide).

16 Categorical – Class - COUNT data analysis

16.1 What type of data is this for?

Up to now, we have been dealing mainly with continuous dependent data (except for logistic regression which is either/or), and either class (ANOVA and the like) or continuous (regression) independent data.

Sometimes you have data that is all categorical (i.e. class), and want to look at whether there are any patterns (associations) between them.

Example: Are you more likely to catch small mammals in traps with bait than in traps without bait? In this case you will have count data on number of traps with bait, number of traps without bait, and number of each of these trap treatments that did, or did not, catch a small mammal. For this sort of count data, you use contingency tables. Note that sometimes there is what you might call an explanatory variable, but there need not be (i.e. you are not necessarily trying to explain one result from another, just see if there are any patterns, in contrast to a random result).

Example:

		Was the trap baited?		
		Yes	No	Total
Did they catch a small mammal?	Yes	20	5	25
	No	10	25	35
	Total	30	30	60

For this type of data, you will probably be asking two things:

(1) Is there a relationship between the two (or more) variables?

- This is answered by comparing observed frequencies in each combination of categories (i.e. for each cell in your contingency table) with expected frequencies (p689), so that:
  Expected = Row Total x Column Total / (Grand total)
  For the above example:
  Expected frequency for baits with traps not catching small mammals equals: (30 x 35)/60
  NOTE: SPSS does these calculations for you, but it is worth seeing what those calculations are.
- For 2 x 2 contingency tables, you use one of two methods:
  - Pearson's Chi-Sq (p688-689)
    Where $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$
    checked against df of (r-1)*(c-1) and r = number of rows and c = number of columns; so in a 2 x 2 contingency, df = (2-1)*(2-1) = 1.
  - Likelihood Ratio (p690)
    Where $L\chi^2 = 2\sum Obs.\ln(\frac{Obs}{Exp})$.
  - Assumptions of Chi-sq test (p691):
    (1) independent data – each entity/case falls into one cell only;
    (2) no expected frequency < 5 (though for large data sets, can have up to 20 % of expected frequencies <5 rather than none, though power is reduced).

- NB/ Use Likelihood ratio in preference to Chi-sq test, because when sample sizes are small, Chi-sq test is more likely to say there is a significant pattern when in fact there isn't (Type I error) (p691).

(2) How strong is the association?

- If you find a significant association (e.g. the number of traps catching small mammals when you bait is significantly different to the number when you don't bait), then it is useful to find out how much more likely you are to get one outcome than another based on this association (e.g. are you much more likely to catch a small mammal if you bait than if you don't bait, or are you only slightly more likely to do so) (p697-698)

16.2 2 x 2 Contingency Tables & tests of significance (Chi-sq & Likelihood Ratio)

Can enter data to create the contingency table in two ways:

(1) raw scores – this is very tedious if you have lots of counts (p692). Each row of the data editor represents each entity (or single case).

(2) frequency data (weight cases) (p692-695) – choose this method normally – this is fast and essentially creates the contingency table straight away. Here you put in the count (frequency) for each combination of categories.

- To enter raw scores
  - Create 2 coding variables in two columns (e.g. baiting [levels 0=no, 1=yes], and trap small mammal [levels 0=no, 1=yes]). So if you set a total of 60 traps out, you would have 60 rows of data.

With bait?	Catch mammal?
0	0
0	1
0	1
0	0
1	0
1	1
1	1

- To enter frequency data
  - Create the same 2 coding variables [Note both of these **Nominal** data when defining them in *Variable View*], but also create a 3rd variable (a **Scale** variable in *Variable View*) called "frequent" or whatever (p693). In this column you put the number of cases which fall into that specific combination of the two variables.

With bait?	Catch mammal?	Frequent

0	0	25
0	1	5
1	0	10
1	1	20

- NB/ to analyse this type of data, you must tell the computer that "Frequent" column represents the number of cases that fell in each combination of categories. To do this, use the menu path: *Data, Weight Cases*, then select the *Weight cases by* option and move the correct variable into the *Frequency Variable* box (e.g. the variable "Frequent"). Hit *ok*, then *save* the data file.

- <u>To run the Chi-Sq (& Likelihood) analysis</u>:
  - go to *Analyse, Descriptive Statistics, Crosstabs* (p694)
    Move one category (e.g. bait?) into the Row box and the other category (e.g. trap mammal?) into the Column box. If you had a 3rd variable, you could add that into the Layer box.
  - Under *Statistics* (p695), select *Contingency coefficient*, *Chi-sq test*, *Phi and Cramer's V*, *Lambda* and *Kendall's tau-b*.
    Chi-sq test,: tests whether there is a significant association between two categorical variables
    Phi and Cramer's V – measures of strength of association. Use *Phi* if only 2 categorical variables each with 2 levels (categories). Use *Cramer's* if one of the two categorical variables has > 2 levels.
    Lambda – value of 1 = one variable completely predict another variable, 0 = no predictive capacity at all
    Kendall's tau-b. – non-parametric correlation coefficient between the variables.
  - Under *Cells* (p695), important to select *Expected* counts (rule of thumb, all should be >5), *Row*, *Column* and *Total* percentages.
  - Under *Exact*, select *Exact* (p695). This provides <u>Fisher's Exact test.</u>
    These days it is usually always used.
    [Before high computational power, people used to use Chi-Sq (which is an approximation) if counts were high because the result is then similar to the Exact test, but still had to use the Exact test if counts were low, because in that case Chi-Sq is inaccurate.]
    The Exact test is particularly appropriate (i) if you have fixed marginal totals for at least one category (e.g. you have chosen to set 30 traps with bait and 30 traps without bait), and/or (ii) if the expected frequencies are < 5. It doesn't use the $\chi 2$ distribution to test the independence (or the opposite i.e. association) of two categories, but instead answers the question "Given our fixed marginal totals, what is the probability of obtaining the observed cell frequencies and all cell frequencies that are further away from the expected?" (Quinn & Keough p388)

16.3 Interpreting the Chi-Sq analysis Output

- *Crosstabulation Table* – (p696) original plus converted into percentages. Important to confirm that no cell had < 5 expected frequency (check "Expected count" >5).
- *Chi-Sq Table*: (p697) within this, Pearson's Chi-Sq statistic and P value tells you whether there is an association between two categorical variables. Check under this table to see if any of the cells have expected counts < 5 (use Exact test if they do; or use it always since it is more accurate).
- Continuity correction is the Yates' attempt at fixing problems of Chi-Sq test, and can be ignored (p691).
- Likelihood ratio statistics is preferred over the Chi-Sq test (p690-691), so may as well always use this statistic to test whether the association is significant or not.
- NB/ neither Chi-Sq nor Likelihood ratio tells you how strong the association is, just whether it exists. To find out how strong the association is, look in the *Symmetric Measures* Table. Use the *Phi-* value for 2x2 contingency table and *Cramer's* if one category variable has more than 2 categories (p695, 698). These values fall between -1, 0 and +1, the closer to -1 or +1, the stronger the association (a bit like a correlation r-value),
  -1 indicates negative association and +1 indicates positive association.

16.4 Odd ratios

Calculating an effect size: often use the odds ratio instead of strength of association, especially for 2x2 contingency tables (p699-700).

- Step 1 – calculate the odds ratio for 1 pair of cells.
  e.g. Odds of catching small mammal with bait =
  number with mammal with bait
  number with mammal without bait
  = (20/5) = 4
- Step 2 – calculate the odds ratio for 1 pair of cells.
  e.g. Odds of not catching small mammal with bait =
  number without mammal with bait
  number without mammal without bait
  = (10/25) = 0.4
- Step 3 – odds ratio is the ratio of Step 1 and 2:
  e.g. Odds ratio = 4/0.4 = 10
  i.e. 10 times more likely to catch a small mammal if you bait the trap.

16.5 Bigger Contingency Tables & tests of significance (Loglinear analysis)

When you have more than two categorical variables, Chi-Sq doesn't work – you have to go to Loglinear Analysis (p702). Note you could do the reverse, i.e., use Loglinear

Analysis for 2x2 contingency tests, which at least means you are consistent in the analysis you use no matter what the size of your table.

To make the categorical data linear (p703), we model the log of the observed counts against the variables of interest, e.g.:

$Ln(Obs_{ij}) = (\beta_0 + \beta_1 Bait + \beta_2 Catch + \beta_3 Interaction_{ij}) + ln(\varepsilon_{ij})$

This is a saturated model (p705) which we can work out. But, we actually want to determine whether a simpler model is satisfactory at predicting the count frequencies in the two (or more) category variables (p709). i.e. we want to test:

$Ln(Obs_{ij}) = \beta_0 + \beta_1 Bait + \beta_2 Catch$

In all, it's complicated, but our test becomes whether the reduced model still explains as much (or nearly) as the full (or "fuller" model). To test whether the new (current) model has changed the likelihood ratio, we subtract the likelihood statistic of the reduced (current) model from the previous (fuller) one (p710):

$L\chi^2_{Change} = L\chi^2_{Current\ model} - L\chi^2_{Previous\ model}$

16.6 Assumptions of Loglinear Analysis

- Assumptions of Loglinear Analysis (p710): similar to Chi-Square test:
  (1) independent data – each entity/case falls into one cell only;
  (2) no more than 20 % of expected frequencies <5, but all cells must have frequencies >1. If this assumption is broken, don't even bother trying to use this analysis as the results will be completely unreliable.
- Remedies if Assumption 2 is not upheld:
  - Collapse the data across one of the variables (e.g. the one you expect to have least effect). To do this, the highest-order interaction should be non-significant (otherwise not legit to combine two into one!)
  - Collapse the levels of one of the variables
  - Collect more data

16.7 Running Loglinear Analysis in SPSS

- Example of data set:

Site	With bait?	Catch mammal?	Frequent
0	0	0	25
0	0	1	5
0	1	0	10
0	1	1	20
1	0	0	30
1	0	1	10
1	1	0	8
1	1	1	15

- Enter the data in the same way as for 2x2 contingency table. Remember to specify that "Frequent" column represents the number of cases that fell in each combination of categories. To do this, use the menu path: *Data, Weight Cases*, then select the *Weight cases by* option and move the correct variable into the *Frequency Variable* box (e.g. the variable "Frequent"). Hit *ok*, then *save* the data file.
- To generate the Contingency Table:
  Go to *Analyse, Descriptive Statistics, Crosstabs* (p694)
  Move one category (e.g. bait?) into the Row box and the other category (e.g. trap mammal?) into the Column box, and the third variable (e.g. Site) into the Layer box. Don't need to select any details, because this stage is just to generate the contingency table.
- Run it and again, check that Assumption 2 is met, i.e. that the "Expected count" in each cell is >1 and not more than 20 % are <5.
- For the main analysis (once you have checked assumptions) (p712-713):
  - Go to *Analyse, Loglinear, Model Selection*, then select any variable you want to include and move it into the Factor box.
  - For each variable, then *Define Range* (e.g. minimum = 0, maximum = 1).
  - Also go into *Model* and make sure it is selected on the Saturated model.
  - In *Options*, select both *Parameter estimates* and *Association Table* (p714).
  - Back in the main menu, hit *OK* to run the analysis.

16.8 Interpreting the Loglinear output

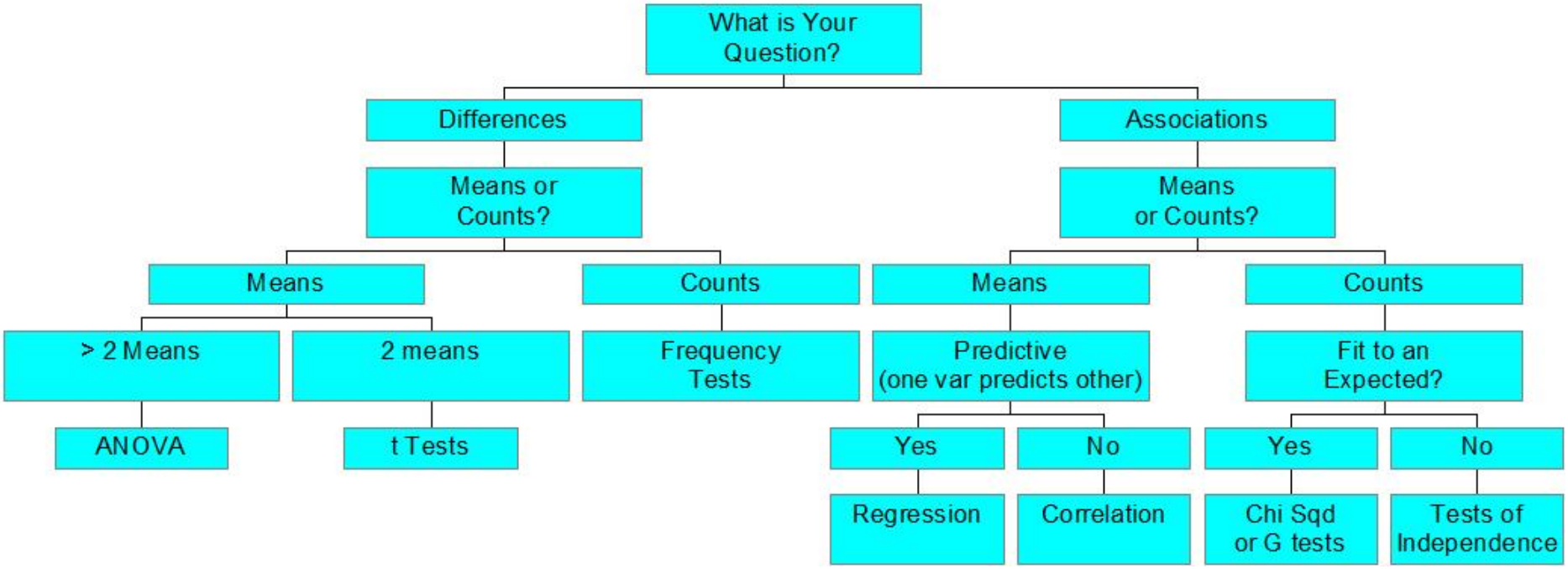
- Contingency Table is the same as for 2x2, but with an extra layer

- Goodness-of-Fit Tests table – note that because you have first tested the saturated model, you should perfectly predict the observed values, so the Likelihood Ratio and Pearson statistic should both= = 0, and the P value is very high ("Sig = ." i.e. the model perfectly predicts the data) (p715)
- What you now want to find out, is what parts of the model you can remove without significantly affecting the fit of the model.
    - K-way and Higher-Order Effects table (p715-716): Check the high-order effects first (i.e. K3 = 3-way interaction), then work your way up the table to look at 2-way interactions (K2) and then last, the main effects (K1). If K3 is significant, then it needs to remain in the model and you stop there. If it is not significant, then it means you could remove it from the model and not affect your predictive capacity.
- Partial Associations and Parameter Estimates tables (p716-717): These both tells us which of the interactions and/or main affects is significant.
    - Partial Associations – look at P-values
    - Parameter Estimates - check the z-score. The higher the value (positive or negative) the more significant the effect, so it is a useful for comparing effects of each factor or interaction.
- The backward elimination statistics Step summary table (p717-718) looks at what happens as each of the highest-order effects are progressively removed from the model. It removes effects only if they have no impact on the model, and stops when the next effect to be removed has a significant effect (i.e. it keeps it in).
- Look at the Goodness of Fit test statistics to see if the final model is a good fit of the data (p718). If it is, the statistic will NOT be significant (i.e. > 0.05 [including "."]).

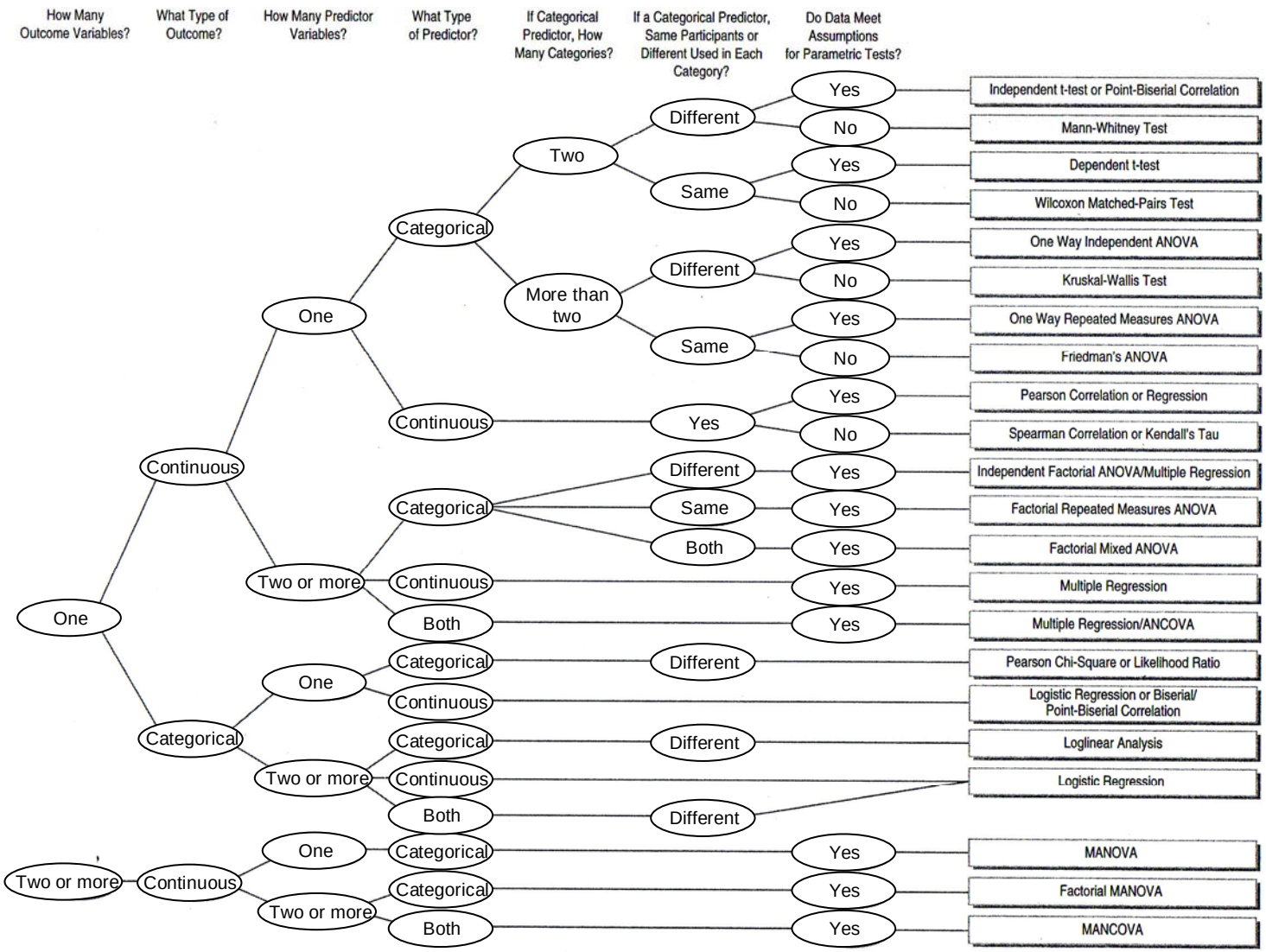
Ross Coleman:

$H_0 : \mu A = \mu B$

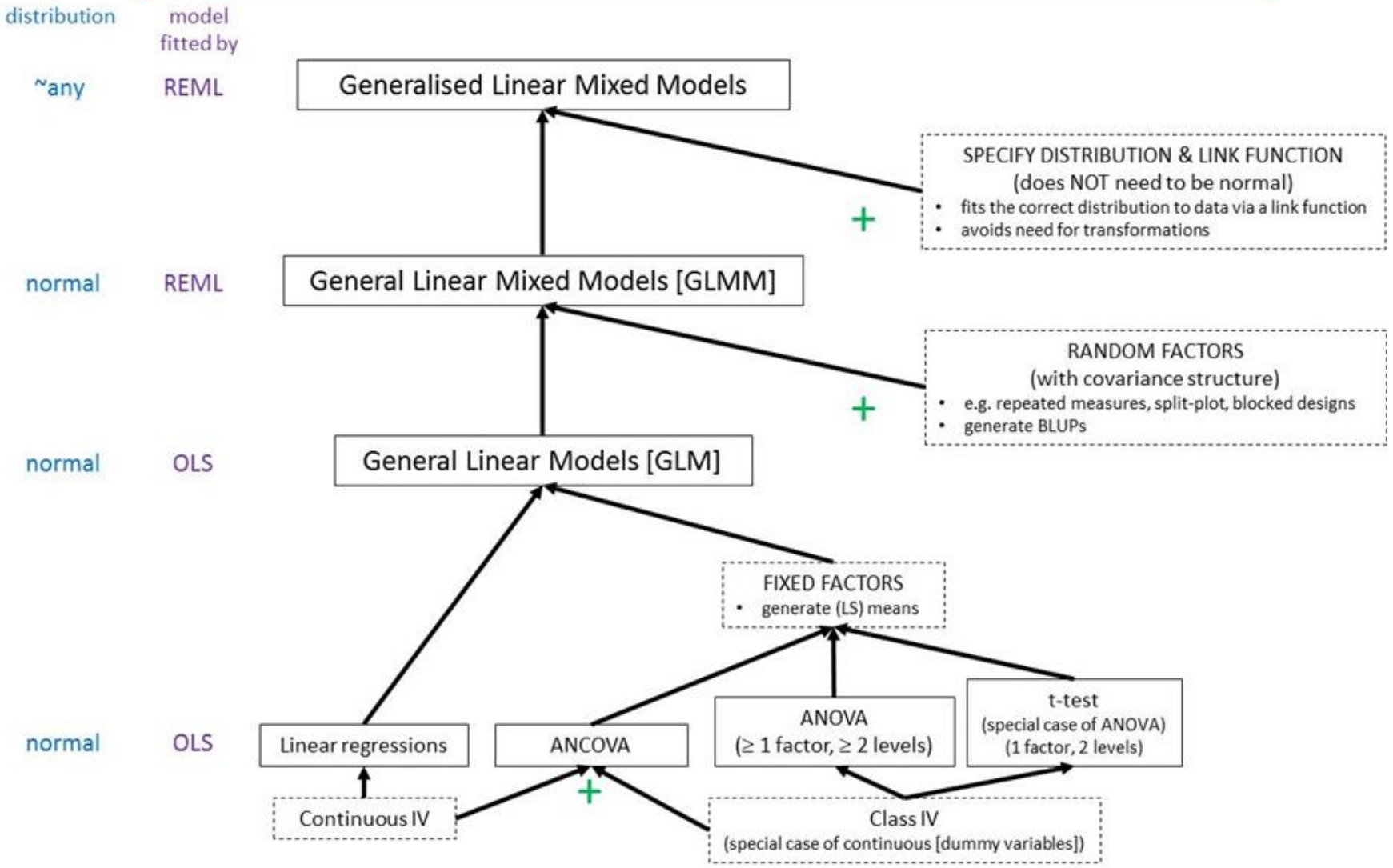
Decisions about tests



Field A (2005) Discovering Statistics Using SPSS. 2nd Edition. SAGE Publications.



Hierarchy of parametric statistical analyses (as Clare sees it)



Based on Gbur EE, Stroup WW, McCarter KS, Durham S, Young LJ, Christman M, West M, Kramer M (2012) Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences. ASA, SSSA, CSSA, Madison USA