APPENDIX E: CHOOSING STATISTICAL TESTS

The myriad of statistical tests and minefield of conditions for their use means that many people find it difficult to remember which statistical procedure or test to use. Therefore, in this appendix we offer a guide to selecting the appropriate statistical procedure for the data that you wish to analyze.

Generally, there are three key factors that you need to take into consideration when choosing the appropriate procedure or test:

1. *The number of variables that the test is to be conducted on*. Statistical tests and procedures can be divided up according to the number of variables that they are designed to analyze. Therefore, when choosing a test it is important that you consider how many variables you wish to analyze. One set of tests are used on single variables (often referred to as descriptive statistics), a second set are used to analyze the relationship between two variables and a third set are used to model multivariable relationships (i.e. relationships between three or more variables).

2. *The level of measurement of the variable(s)*. Statistical tests are specifically designed to be used on variables measured at a certain level and it is therefore essential that when choosing a statistical procedure you are certain which level the variable you intend to analyze is measured at. There are four levels of measurement: nominal, ordinal, interval or ratio. A brief reminder of their meanings is provided in the grey box below.

3. *The aim of the analysis.* Some research is conducted in order to describe a phenomenon. This sort of research will tend to involve the use of descriptive statistics. Other research is exploratory. With this sort of research you are likely to be searching for patterns and relationships in the data. Statistical procedures that reveal patterns or automatically try to fit models of relationships for you are often useful for this kind of research. Finally, a great deal of research involves testing hypotheses about patterns of causation in the population. For this sort of research, n order that you can infer from your data to the population, you will find yourself making use of procedures that use significance tests.

LEVELS OF MEASUREMENT

There are four levels of measurement:

- ***Nominal* measures** are at the lowest level. They use numbers simply as names or as labels for different values, and no mathematical significance therefore attaches to the numbers themselves. A nominal variable consisting of just two categories is referred to as a *dichotomous* variable.

- ***Ordinal* measures** are one level up from nominal ones. Like nominal measures, they use numbers simply as labels, but the numbers do form a rank order of ascending or descending size. On an ordinal scale, something coded zero is less than something coded 1, which is in turn less than something coded 2, but we do not know how much less, and the gaps between each point on the scale are not necessarily equal.

- ***Interval* measures** also use numbers to represent a rank order, but unlike ordinal measures, they also ensure that the intervals between each value on the scale are the same.

- ***Ratio* measures** are the highest level of measurement. Here numbers form part of a scale with equal intervals and a base of zero such that they can be expressed in proportion to one another.

ANALYSIS OF SINGLE VARIABLES (UNIVARIATE ANALYSIS)

In the early stages of the analysis, or if the purpose of the research is purely descriptive, you are likely to want to analyze single variables. With the analysis of single variables the aim is usually to describe the characteristics of each variable such as, the distribution of responses and the number of (non) responses.

Nominal and ordinal level: While the *frequencies* procedure is used on variables of all levels of measurement, the options selected within it vary according to the level of measurement. For nominal variables the procedure is run with the tables displayed in order that we can obtain the number of cases and percentage of cases in each category. For ordinal variables, in addition to describing the proportions in each category, you are likely to want to know about the distribution of responses. Therefore, you will want to request summary statistics to obtain the median, quintiles, quartiles and so on. To display this output in graphical form, bar charts or pie charts can be produced.

Interval and ratio level: You will only want the frequencies procedure to produce summary statistics on the form and shape of the distribution (e.g. the mean, median, standard deviation and skew), so the frequency tables should be suppressed. An additional procedure for variables measured at the interval and ratio level is the *explore* procedure. This procedure enables a more detailed examination of the form and distribution of the data. Because some of the statistics that we will want to use on variables at these levels of measurement make assumptions about the population distribution (i.e. they are *parametric* tests) it is important that they are examined for evidence of non-normality in the data. In addition, we may want to examine the data for patterns and extreme outlying cases. Within the explore procedure you can request stem and leaf plots, boxplots and normality plots in order to help in this analysis. A useful way of graphically displaying the form of the distribution is by producing a histogram (possibly with the normal curve superimposed).

Describing and Exploring Statistics with One Variable (Univariate Analysis)

	Level of measurement	
	Nominal or ordinal	***Interval or ratio***
Appropriate statistical procedures:	Frequencies procedure	Frequencies procedure (frequency table suppressed) Explore procedure
Appropriate ways of displaying the data:	1. Frequency tables displaying raw numbers or percentages (nominal data); percentages, medians, deciles and quartiles (ordinal data). 2. Bar charts 3. Pie charts	1. Summary statistics displaying: means, medians, standard deviations, skew etc. 2. Histograms 3. Stem and leaf plots 4. Box plots 5. Normal plots
Procedures covered in:	Chapter 7	Chapters 7 & 8

BIVARIATE ANALYSIS

When you come to analysing two variables in choosing the appropriate statistical test you need not only to consider the level of measurement of the each of the variables but also the purpose of the analysis. If you are conducting exploratory analysis you are likely to be concerned with patterns of relationships and strength of relationships. If, on the other hand, the primary aim is to test hypotheses then you will want to establish the significance of relationships and you may be concerned to test hypothesized lines of causation.

Nominal by nominal

When both variables are measured at the nominal level the *crosstabs* procedure is used in order to analyze patterns of association. If you are testing a hypothesis then a significance test should be used. In most cases this will be the *chi square test* but if both variables are dichotomous (i.e. have only two values) then *Yate's correction* should be applied to the test. If the sample size is below 20 then *Fisher's exact test* should be used.

A statistically significant result may not represent a substantively significant association so statistical tests can be applied to test the strength of the relationship. There are three measures of this: the odds ratio, the phi coefficient (used for 2x2 tables) and Cramer's V (used for tables with more than 4 cells). If you are testing a hypothesis and it suggests a line of causation, then asymmetrical measures of strength of association such as, *Lambda and Goodman* and *Kruskal's tau*, can be used

Where you wish to explore the effect that a third nominal variable may have on a two-variable relationship (for example, to analyze whether the original relationship is spurious or to examine whether there is interaction between the three variables), the *crosstabs* procedure allows a third variable to be added to the analysis.

Ordinal by dichotomy

The *Mann-Whitney U* test is used to test the statistical significance of an association between two variables where the dependent variable is measured at the ordinal level or higher and where the independent variable is dichotomous. It is a non-parametric test and is therefore sometimes used in preference to the parametric t-test (see below) when the dependent variable is measured at the interval or ratio level but when assumptions of the t-test are violated.

The test compares the difference between the two groups of the dichotomous variable in respect of their average location on a distribution of rank scores on the ordinal variable. If you wish to compare differences in the overall distribution of the ordinal variable on the two values of the dichotomous variable then you should consider using the *Kolmogorov-Smirnov Z test* and *Wald-Wolfowitz runs test*. If you wish to compare extreme responses in each group to see whether the groups differ significantly at the extremes of the distribution then you should use the *Moses extreme reactions test***.**

All of these tests are independent-samples test and assume that the two groups being compared are independent of each other and not matched in any way

Ordinal by ordinal

There are several different statistics which can be used to express significance and strength of association when dealing with two variables measured at the ordinal level. These include *Spearman's rho, Kendall's tau-b* and *tau-c, Goodman and Kruskal's gamma,* and *Summer's d*. All are non-parametric statistics and all are based on the same logic, which involves comparing the values of the two variables for all possible pairs of cases. There is little to choose between these tests although Spearman's rho is more frequently used. They can be obtained from the statistics option within the *crosstabs* procedure.

Interval/ratio by nominal

Providing that the sample has been randomly generated and the variable is normally distributed, when the dependent variable is measured at the interval level and the independent variable is dichotomous, the *t-test* is the appropriate statistical test. It is used to test for a statistically significant difference in means of an interval level dependent variable between two groups. If the two groups are independent of each other (that is, the allocation of an individual to one group is independent of the selection of an individual to his or her group) then you should use the *independent-samples t test*. If the two groups are not independent of each other then you should use the *paired-samples t test*.

As was explained in chapter 12, the t-test is the simplest form of an analysis of variance test (ANOVA). A *one-way Anova* test conducted on an interval level dependent variable and a dichotomous independent variable will produce exactly the same results as a t-test. However, unlike the t-test, the one-way ANOVA can also be used when the nominal independent variable takes more than two values. The procedure uses the *F-test* to determine whether the differences between the groups are statistically significant.

Interval/ratio by interval/ratio

If you wish to calculate the *strength* and *direction* (positive or negative) of an association between two interval or ratio level variables then you should use the *correlation* procedure and request the *Pearson correlations coefficient.*. Correlation is mainly used in exploring data, to see whether variables are associated with each other.

If your analysis involves theory testing, where the aim is to see whether hypotheses concerning the causal impact of one variable on another can be falsified, then the appropriate statistical procedure is *bivariate regression*. It calculates an equation which allows us to predict the value of a dependent variable from the value of the independent variable.

Both procedures assume that any association between the two variables is linear. In order to examine the relationship between the variables to check for this you should run a *scatterplot*. If there is evidence that the linearity assumption has been violated when running a correlation then you should consider using the non-parametric equivalent – Spearman's rho. Regression also makes the assumption that the dependent variable is normally distributed at each value of the independent variable (sometimes called *homoscedasticity*). This assumption must be checked by examining the residuals.

MORE THAN TWO VARIABLES: MULTIVARIATE ANALYSIS

At this level of analysis the concern is usually to fit models of relationships between a number of variables to the data. If the analysis is exploratory then you can allow the procedure to automatically seek the best fitting model for you (although you should note that the final model that is chosen by the procedure can vary according to the selection criteria that are used). If you are conducting multivariate analysis in order to test hypotheses then it is important that you select the models manually according to the hypothesis under test.

A dichotomous dependent variable by multiple interval/ratio variables

When the dependent variable is dichotomous and you wish to build a model that predicts the dependent variable with a number of interval or ratio variables, then you should use *logistic regression*. Unlike in ordinary least squares regression where the independent variables are used to predict the value of the dependent variable, in logistic regression it is the probability of the dependent variable equalling 1 that is predicted. In logistic regression nominal and ordinal variables can be used to predict the dichotomous dependent variable as long as they are transformed into dummy variables first.

Multiple nominal variables

When you wish to understand the pattern of relationships between three or more variables measured at the nominal level you should use *loglinear analysis*. With this

procedure there are no lines of causation and, therefore, there are no dependent variables or independent variables. The aim is to model the simplest set of relationships between the nominal variables that best predicts the observed data.

An interval/ratio variable by multiple nominal variables

The procedure used to predict variations in the interval or ratio variable by multiple nominal variables is the *two-way ANOVA*. This procedure calculates not only whether each of the independent variables is significantly associated with variation in the dependent variable, but also whether each independent variable continues to have an effect once we *control* for the influence of the other.

An interval/ratio variable by a mixture of nominal and interval variables

When you want to compare the difference in mean scores in an interval or ratio dependent variable by categories of a nominal variable while controlling for the influence of an interval level variable which is known to co-vary with them you should use the ANCOVA. It allows us to examine whether differences in mean scores remain statistically significant once the confounding influence of the covariate has been removed.

An interval/ratio variable by multiple interval/ratio variables

An extension of bivariate regression called *multiple regression* is used when we wish to predict the values of a interval or ratio dependent variable on the basis of two or more interval or ratio variables. The logic and procedure of multiple regression is virtually identical to that for bivariate regression, except that you are looking at the relative impact on the dependent variable of a number of different independent variables, rather than just one.

If you are exploring models of independent variables that best predict the dependent variable you can ask the procedure to fit the model for you. However, as is shown in chapter 10, there are a number of different model selection techniques and these may produce conflicting results. If you are testing hypotheses that claim certain

independent variables predict the dependent variable then you should build the model yourself.

One of the features of multiple regression that makes it so useful is that also allows the use of nominal and ordinal variables as independent variables if they are first transformed into dichotomous dummy variables (see chapter 10).

BIVARIATE AND MULTIVARIATE STATISTICS

	Level of measurement of second variable:			
Level of measurement of first variable:	***Dichotomous***	***Nominal***	***Ordinal***	***Interval or ratio***
Dichotomous	**Bivariate:** *Chi square (*with *Yates' correction)* Also: *Fisher's exact test. Phi, odds ratios, Lambda, Goodman & Kruskal's tau* Chapter 8 **Multivariate analysis:** *Loglinear analysis* Appendix I	**Bivariate:** *Chi square.* Also: *Fisher's exact test. Phi, odds ratios, Lambda, Goodman & Kruskal's tau* Chapter 8 **Multivariate analysis:** *Loglinear analysis* Appendix1	**Bivariate and multivariate:** *Logistic regression* (ordinal variable must be transformed into dummy variable) Appendix H	**Bivariate and multivariate:** *Logistic regression* Appendix H
Nominal	**Bivariate:** *Chi square.* Also: *Fisher's exact test. Phi, odds ratios, Lambda, Goodman & Kruskal's tau* Chapter 8 **Multivariate analysis:** *Loglinear analysis* Appendix I	**Bivariate:** *Chi square.* Also: *Fisher's exact test. Phi, odds ratios, Lambda, Goodman & Kruskal's tau* Chapter 9 **Multivariate analysis:** *Loglinear analysis* Appendix I		
Ordinal	**Bivariate analysis:** *Mann-Whitney U test* Also: *Kolmogorov-Smirnov Z test, Wald-Wolfowitz runs test* and *Moses extreme reactions test.* Chapter 9		**Bivariate analysis**: *Spearman's rho.* Also: *Kendall's tau-b and tau-c, Goodman and Kruskal's gamma* Chapter 8	
Interval or ratio	**Bivariate analysis:** *T-test* and *one-way Anova* Chapter 9	**Bivariate analysis:** *One-way Anova* **Multivariate analysis:** *Two-way Anova* Chapter 9	**Bivariate and multivariate:** *Linear regression* (ordinal variable must be transformed into dummy variable). Chapter 10	**Bivariate analysis:** *Bivariate regression* **Multivariate analysis:** *Multiple regression* Chapter 10