

Chapter 12

Data Collection Measurements and Analysis

In planning the collection of data it is necessary first to carefully consider what information is needed to answer the research question. A list of all the possible sources of information (e.g. case records; laboratory data; questionnaire; archival sources, and so on) is first prepared. The next step is to consider what would be the most cost-effective way of obtaining the required information. Whilst listing the sources of information and means of obtaining it, it is helpful to also list all the variables that would be needed for answering the research question, as well as the justification for the use of each in the study. This focuses the mind and one does not end up loaded with irrelevant information, or information which is relevant but inadequate for providing the full answer to the research question.

Bearing in mind that in most research the investigator is examining the relationship between two or more variables, this is also the time for translating hypothesis into operational variables. It requires definitions of terms in the hypothesis, and consideration of how each operational variable would be measured. For example, in a hypothesis concerning nutritional status and immune response the researcher first translates "nutritional status" and "immune response" into operational variables like weight for age, weight for height, body mass index, mid upper arm circumference and so on for assessing nutritional status, and antibody response, B lymphocytes, cell-mediated immunity or other tests as measures of immune response. He then has to consider how each of these would be measured.

The question of **validity** and **reliability** (accuracy and precision) arises with regard to all measurements. Validity is asking, "Are we measuring what we think we are measuring?". In the case of laboratory tests this entails quality control, which is usually done by running the test on a known standard specimen. Reliability is asking the question "How reproducible is the result?" In the case of laboratory tests reproducibility of test results is assessed by running the same test on several occasions on the same standard specimen. Difficulties with regard to validity and reliability arise when one is using a score for making a value judgment. For example, assessing the socioeconomic status (SES). Taking income alone as a measure of SES is not enough, because it is commonly under or over reported. Hence education, occupation, quality of housing, amenities like piped water, sanitation, electricity etc. and material possessions in the house are all taken into consideration. Each item is assigned a score and SES is measured by adding up the individual scores. Intelligence quotient is also measured by scoring on several items. A similar approach is employed for assessing behaviour in school children (Rutter's Score), depression, emotional problems, and so on. Many such 'instruments' have been validated in a variety of sociocultural settings, and a study of the literature would give a good idea as to how well the 'instrument' is likely to perform in the investigator's own setting. In some cases a modification of an existing instrument or development of a new instrument may become necessary. Here validation is crucial prior to application. One way is to take the sum of the scores obtained on half the items chosen randomly and measuring the correlation coefficient against the sum of scores on the other half. A value of 0.6 and above indicates validity. This procedure is known as Cronbach's alpha. Some software

packages (e.g.SPSS) provide a procedure for checking it. If the concern is about an individual item, one correlates the score on that item against the total score. Items which do not correlate well may then be dropped from the list.

Precision and Accuracy. A precise measurement is one, which gives nearly the same value each time it is made. Mean \pm SD gives an idea of the precision. For example, a wide range of the standard deviation makes one wonder how precise the measurement is. The precision of several variables can be compared by the **Coefficient of Variation** (which is $100 \times \text{SD} \div \text{Mean}$). Imprecise measurements have large coefficients of variation.

Precision of measurements can be verified by means of taking paired measurements and checking for:

1. Test - retest consistency.
2. Inter and intra-observer consistency.
3. Internal consistency. For example, concordance between two variables which measure the same general characteristics as in the case of questions designed to assess ability to feed oneself, or walk, or I.Q. etc. The consistency is checked by computing the correlation coefficient.

Accuracy is the degree to which a measurement actually represents the true value. Appendix 12.1 lists ways of reducing systematic errors in measurements.

Every measurement, however carefully made, has built in sources of error. These are related to the observer, the subject, the instrument, and individual variability. Performance of the observer and the subject is affected by the mood, fatigue, background noise, and learning. Through learning and experience the observer performs better, and intra-observer reliability improves. Through repeated administration of a test the subjects also do better, for example tests for measuring the I.Q.

When gathering data which rely on recall a major determinant is proximity to the event. In the case of questionnaires administered by the researcher, there is always an opportunity for checking and double-checking. In self administered or postal questionnaires it is not so and the information becomes that much less reliable. For testing the reliability of any answer the best approach is repeated administration of the questionnaire and comparing how well the answers agree.

When the list of variables, the measurements to be made, and the question of validity and reliability have been all considered, the next step is to establish the ground rules of the study. These should be written down for future reference and to ensure that drifting does not occur as the study progresses. The ground rules relate to the following items:

- Subject eligibility with clear-cut inclusion and exclusion criteria.
- The format for recruiting the subjects. Will subjects be recruited from the in-patient wards, from outpatient clinics or from the community? Would they be recruited from amongst those currently attending the hospital, or would they be chosen from existing registers and records?
- Definitions of key data items should be stated at the outset. (For example, definitions of "low birth weight", "adolescence" "malnutrition" and so on). Bearing in mind that concepts in the hypothesis are first assigned operational definitions for purposes of measurement, analysis and drawing inferences and are subsequently translated back into the concepts carried by the hypothesis, clear definitions of terms are important.

- The scheme for managing the data, coding of the data, and entering the data into the computer for analysis should be determined. This is also the time to determine what variables are to be compared during statistical analysis. For this purpose "dummy" tables are useful. These are tables drawn up with study variables as headings and showing different values of each variable but with empty cells for the data. Such "dummy" tables provide a plan for the lay out of the data for comparison, for assessing significance, and for prediction if any is to be made. If the help of a statistician is to be sought this is the time to do so. At this stage the researcher may also wish to brush up his knowledge about the statistical tests to be employed for testing significance.

Data Analysis

As a first step it is essential to check whether the data have been entered correctly. Almost all computer software have a facility for checking data entry. All inferences drawn and conclusions reached as a result of statistical manipulation of the data rest on the assumption that all measures were valid and reliable, and that the data had been entered correctly. No amount of statistical manipulation can ever make up for errors in data collection and entry.

Analysis begins with exploring and describing the data and confirming its characteristics. The first step in doing so is to obtain the frequencies. An unexpected value in the frequency table (e.g. age 99 in a study of children) indicates incorrect data entry or incorrect recording of the information. One needs to go back and check. Similarly, missing values may become obvious from frequency tables and one may have to go back and do a careful check. This step is referred to as Data Cleaning.

The main reason for performing a frequency distribution is to take a good look at the data, and see how values are distributed. Most computer software also have the facility for doing graphics. Visual display of data in the form of bar charts; histograms; pie charts and so on are helpful in providing a mental image of the data, and for defining the relationships between variables. Visual displays often make us note the unexpected. It is often the case that different types of graphical displays show us quite different aspects of the same data. Depending upon the research objectives the more important tables may be visualized in the form of charts and graphs. Bar charts and pie charts are used for categorical data. Histograms, line graphs and scatter diagrams are similarly employed for visualizing continuous numerical data.

Some software routinely produce the mean and standard deviation at the bottom of each frequency table. This is helpful, but it is advisable not to rely solely on these parameters for two reasons. Many variables depart markedly from the shape of a normal distribution by having skewness, outliers or multiple peaks. Whenever such abnormalities occur in the distribution of the data values, the shape of the distribution rather than the mean or standard deviation is likely to be the most important characteristics. Secondly, one may be misled by reliance on numerical data alone and not taking the shape of the distribution into account.

Shape is best perceived visually. Hence the usefulness of graphics during exploratory data analysis. From the statistical point of view a " Stem and Leaf" plot is an additional help to a bar chart. It retains the numeric representation of data points while arranging them in the shape of the distribution. Another useful plot is the " Box and Whisker" plot. Which arranges the data points

into quartiles. In a Normal distribution the expected positions of the quartiles are at 0.675 standard deviation below and above the mean. The Box contains half the data points (i.e. between the 25th and 75th quartiles) and provides a visual summary of the values in this range while the Whiskers give the details at the ends of the distribution.

Transforming the data

If the data is not Normally distributed (e.g. antibody levels) a useful step is to first transform the data to a scale on which the distribution approaches Normality. Often transforming the readings to the logarithmic scale produces a data set, which is near to Normal in distribution. Other common types of transformations applied to skewed distributions are the square root, the square, and the cubic transformation.

Details in the data

The graphics mentioned in the preceding paragraphs are various ways of looking at the data and for developing a mental image of the findings. The next step is to know what to look for. Skewness, outliers, gaps, multiple peaks or bimodal distribution are some of the important characteristics which one should observe.

The researcher is usually interested in exploring the relationships between two or more variables. The best way of visually exploring relationships between two numerical variables is by means of scatter plots, sometimes also called X-Y plots. These are simple plots where the values of one variable are charted on the horizontal (X) axis and the corresponding values of the other variable on the vertical (Y) axis. In studying scatter plots the researcher is looking for three important features. These are:

1. The shape of the relationship whether linear, U-shaped, J-shaped, S-shaped, or no obvious relationship.
2. The strength of the relationship. By this is meant the extent to which the observed values of the dependent variable (commonly on the X axis) can be predicted from the corresponding values of the independent variable on the Y axis;
3. The direction of the relationship, which means whether high values on one variable are associated with high values of the other variable, or the opposite. Most computer software would do scatterplots. After studying the initial plot the researcher may wish to straighten out the scatter as much as possible. Plotting y^2 , $/y$, $\log y$, $-1/y$ instead of y or doing likewise with x are useful steps in a search for straightness.

In many instances the research question or the data point to a multivariate relationship. In such a situation a gradual stepwise approach is the best policy. To begin with, one first tries to understand each variable as a separate entity. By using "Stem and Leaf", "Box and Whisker", bar charts and so on the outliers, gaps and multiple peaks are explored. Next one looks at pairs of variables exploring relationships by means of scatterplots and linearising transformations where needed. Next one tries to understand the network of relationships in the data by stepwise grouping of variables.

After observing the data and performing the necessary transformations as needed, the next step is to work out the summary statistics of numeric variables like means, standard deviations, median, minimum and maximum values and so on.

Describing the sample and exploring relationships between variables.

After the preliminary steps of checking and exploring the data the researcher is ready to begin the next phase of describing the sample and exploring the relationships between variables by means of cross-tabulation. Cross-tabulations have three main uses depending on the objectives and the type of study:

- Cross-tabulation to describe the sample. The purpose is to describe the problem under study by presenting a combination of variables. Descriptive cross-tabulations are also used to describe the subjects in terms of a combination of background variables like age, sex, profession, residence and so on.
- Cross-tabulations in which groups are compared to determine differences.
- Cross-tabulations that focus on exploring relationships between variables.

Cross-tabulation to describe the sample. In all studies with quantitative data the common practice is to describe the research subjects first. This can be done for individual variables in a frequency table, or for a combination of variables in a cross-table. If the study is largely descriptive but aims at quantification of a problem, cross-tables are useful for presenting the findings. The data in the tables are usually given as frequencies, and there is an option for getting them as percentages as well.

Cross-tabulation to determine differences between groups. In comparative studies (e.g. case-control; cohort; or experimental) some objectives will focus on discovering whether any differences exist between two or more groups on particular attributes (variables). The groups to be compared are, by convention, placed into rows and the outcome variable (dependent variable) in columns.

Cross-tabulation to explore relationships between variables. Such tables are constructed when the aim of the research is to explore possible relationships or associations among variables. A prior decision is needed as to which variables are outcome (dependent) and which are explanatory (independent).

Steps in preparing appropriate cross-tables.

1. Each specific objective and the method chosen for collecting the relevant data are reviewed.
2. Hypothetical sentences are written according to what is considered to be the type of conclusions the investigator expects to reach. The objective is to avoid wasting time on making meaningless tabulations, and to keep the purpose of each tabulation well defined.
3. For each expected conclusion cross-tables are determined. As a first step dummy tables are prepared.
4. Data analysis is performed and the relevant frequency counts are entered in the cells of the tables.
5. Data interpretation is carried out, and conclusions drawn.

General Hints

- All categories of the variables should be specified. There should be no overlap or gaps between categories.
- The column and row counts in the cross tables should tally with the frequency counts of each variable.
- The grand total should correspond with the total number of subjects in the sample.
- For each table there should be a clear title, and headings for rows and columns to avoid misinterpretation.

Statistical analysis.

The first step in explaining any observed difference between groups, or an association between variables, is to consider whether chance bias or confounding might be behind it. Bias and confounding need to be excluded by carefully checking the study design, before complicated statistical procedures are employed to assess the role of chance. Statistical tests of significance are performed to rule out whether the observed results could have occurred by chance. The generally agreed convention is to take 5% and below as the cut-off point. In stating the results authors mention the *P* i.e. probability of chance value by stating whether it was less than 0.05, 0.01 or 0.001.

It is important to note that "statistically significant" does not necessarily mean that an observed difference or association is an important one clinically. Even a very small difference will show "statistical significance" if a big enough sample is taken. On the other hand an important difference may fail to reach statistical significance if the sample is small.

Which Test to Apply?

Before discussing statistical tests let us briefly consider the ground rules that apply to all measurements. In recording measurements what the researcher is doing is to assign numerals to observations. For some observations scales are used. A problem can arise when ordinal scales are treated in the same way as interval scales. When steps on a measurement scale are not of equal length (for example, social class scores) values from different subjects ought not to be added to calculate a mean. By the same token statistical analyses which are based on mean and standard deviation (e.g. correlation, ANOVA etc.) become unreliable.

Secondly, medical research aims to decide whether or not a difference observed is true. Hence other causes of variations like bias, confounding and random variation are assumed to be absent. Only when a sound design has made bias and confounding unlikely would statistical analysis help to differentiate whether an observed difference did not arise by chance.

Now to come back to our question about which test to employ, the answer is "It depends on the aim of the study and the type of data collected".

If the aim of the study is to determine the difference between groups, then the first condition to satisfy is whether the observations are "paired" or "unpaired". In other words was matching done for each individual subject in the sample? In most studies the groups are selected independently (unpaired or unmatched subjects). The next step is to check what types of variables are being compared, whether nominal, ordinal or numerical.

The two most commonly reported tests in the medical literature are the 't' test and the chi-square (χ^2) test. Both the tests are based on the principle of comparing the observed values of the response variable to the expected values. In the case of the 't' test the test compares the observed value of a mean to the expected value in order to determine whether the difference noticed could have arisen by chance. In other words whether the result is significant. The χ^2 test does the same in the case of nominal and ordinal variables.

Both the 't' test and the χ^2 tests are tests of significance. They only assess the possibility that an apparent relationship between variables is not due to chance, and have nothing to do with clinical significance. Statistical tests of significance are based on the size of the difference, size of the groups and the variability of the scores. It is a sobering thought that any observed difference between two groups can be made "statistically significant" by taking a sufficiently large sample! Hence there is no point in a blind adherence to tests of significance. Instead one should carefully first consider the biologic implications of any apparent difference before accepting a finding as truth on the basis of a test of significance.

Parametric and Non-parametric tests

Statistical tests fall into two broad groups viz. parametric and non-parametric. To give an example, the 't' test falls in the group of parametric tests and χ^2 is non-parametric. Parametric statistics like mean, standard deviation, standard error of the mean, and tests like 't' test, analysis of variance etc. are employed for analyzing continuous variables. They are based on the assumption that the variables in question have a normal (Gaussian; bell-shaped) distribution. Non-parametric tests are commonly employed for the analysis of variables measured on the nominal or ordinal scales, and do not assume a normal distribution. Non-parametric tests can also be used for analyzing continuous variables if the values are not normally distributed.

Choice of tests

In the description that follows suggestions are made about what tests to use in what kind of situation. This is also presented in a summary form in Appendix 12.2. No attempt is made to explain the mechanics of the test. For this the reader is referred to books on statistics. Because of the availability of powerful statistical packages which are user friendly, it is necessary to say that applying a battery of tests without knowing their relevance can lead to mistakes in interpretation of results.

Before commencing analysis the essential first step is to identify all the key variables involved in the research, and amongst these the outcome (or dependent) and explanatory (or independent) variables. For each of these variables one then considers the levels of measurement used e.g. nominal, ordinal or numeric. One should attempt to analyze all the information contained

in the data. If we analyze an ordinal variable using techniques, which are appropriate for a nominal variable, we are not using all the information available. A numeric variable contains more information than ordinal variable, and ordered categories provide more information than unordered categories. Information helps us reach correct decisions. The more information that is used in making a decision, the more likely we are to find a correct answer.

Many statistical procedures test the relationship between two variables. One way of looking for relationship between two variables is by means of X-Y plots. Ideally both the variables should be numeric. However, correlation can still be looked for with one of the two variables being categorical. If one variable is categorical and the other ordinal it is not possible to obtain a directional relationship.

For two categorical variables the χ^2 test is used to compare the observed counts in the cross - classification table with counts that would be expected if there was no association between the variables. With small sample size pooling of rows and/or columns can be used to increase expected frequencies and improve the χ^2 estimation.

Testing for relationship between ordinal and a numeric variable depends on how much we know or can assume about the order in the ordinal variable. In some cases it may be possible to assign numeric values to the various categories. For example, levels of education may be dealt with by assigning number of years of schooling to categories like primary, secondary, post-secondary, and so on. If numeric values can be rationally assigned to the ordered categories, the Pearson or Spearman correlation can be computed between these values and those of a numeric variable.

Sometimes it is not possible to assign specific numeric values to ordered categories, but it may be possible to assume equal spacing. For example strongly agree / agree / disagree / strongly disagree might be viewed as equally spaced. In this case the numbers 1,2,3 and 4 can be assigned to these responses following which they can be related to a numeric variable using Pearson or Spearman correlation. Correlation test is unaffected by linear transformation. It makes no difference whether we use numbers 1 through 4 or assign values 10, 20, 30, and 40 to the four ordered categories.

Another commonly employed statistical procedure is to test for differences in scores between two or more groups e.g. sex, age, socioeconomic groups, or subjects on different treatments. If one variable is categorical (e.g. a grouping variable) and the other numeric a two-group t test or one way analysis of variance (ANOVA) may be employed. The t test is used when the categorical grouping variable has only two categories, and ANOVA is used when there are more than two categories. The numeric variable is the one for computing means and variances. The categorical variable is the group with two levels for a t test and more for ANOVA. A significant difference between the means is evidence of a relationship between the categorical and numeric variable. Equivalent non-parametric procedures like the Wilcoxon, Mann Whitney and Kruskal-Wallis test can also be used to test for relationship between categorical and numeric variables. The non-parametric tests are preferred when the numeric variable has non-Gaussian distribution.

Some ordered categories are difficult or impossible to convert to numeric values. For example, urban suburban and rural cannot be reasonably assigned equal spacing. For such variables it may be safer to ignore the order of the categories and resort to one-way ANOVA.

Effect size and the 't' statistic

As noted above the 'z' or 't' tests are commonly employed to test for significance in the difference between two means. The 't' test is used when the sample size is small e.g. <30, otherwise the basis of the two tests is similar. However, most computer packages report a 't' test regardless of the sample size.

A lesser known use of the 't' test is for calculating the effect size. The significance value for 't' does not tell us to what extent the two variables being compared are affected by one another. For example, in a comparison of two treatments for hypertension the new treatment may be more effective and give a significant 't' test result. But how effective? This can be assessed by calculating the effect size i.e. the relative magnitude of the difference between the two means.

One of the several statistics used to calculate effect size is 'eta squared'. Eta squared can range in value from 0 to 1 as follows:

- 0.01 = small effect
- 0.06 = moderate effect
- 0.14 = large effect.

Eta squared is calculated from the formula
$$\text{Eta squared} = \frac{(t - \text{statistic})^2}{(t - \text{statistic})^2 + (N_1 + N_2 - 2)}$$

Where N_1 = Number of subjects in group 1

N_2 = Number of subjects in group 2

The ANOVA procedure

When more than two means are to be compared the test employed is the One-Way Analysis of Variance (or ANOVA for short). One-way because there is one potential source of variability between the groups (e.g. two different treatments for hypertension being compared with the standard one and hence three groups of subjects). The ANOVA procedure computes the variability between the group means (i.e. between group variance) which is weighted by the group sizes and compares it with the random variation within each group (i.e. within group variance) to obtain the F statistic. Thus F is the ratio:

$$F = \frac{\textit{Betweengroup variance}}{\textit{Withingroup variance}}$$

If there is no difference F should be close to 1. If the group means are not equal, between group variance is raised and F is large.

The 't' test is just a special case of ANOVA. If we were to analyze the means of two groups by ANOVA the significance level would be the same as doing a 't' test. In spite of its name One-way ANOVA is actually looking at difference between the means of groups, but does so by using the variances to decide whether the means are really different.

Interaction between two variables (See also Appendix 12.4).

An association, which is statistically significant, does not necessarily mean the existence of a causal relationship. All it implies is that further exploration is needed to establish a causal relationship.

For measuring associations between variables the first requirement is to check whether the data are nominal, ordinal or numerical. For nominal data the Odds Ratio (for case-control studies) and the Relative Risk (for cohort studies) are useful measures of association.

When a continuous response variable (usually called dependent variable; "Y") is being influenced by another quantitative factor (also called independent variable; "X"), the strength of the relationship is determined by Pearson's Product Moment Correlation Coefficient 'r'. Pearson's 'r' is measured on a scale of -1 to +1, and the value of 'r' gives a measure of the increase (or decrease) in the value of Y for each unit change in the value of X; r^2 (by convention written as R^2) gives the coefficient of determination. It is a measure of the amount of variation in "Y" that is explained by the variable "X". It is also a measure of the goodness of fit between the X and Y variables. In that context one should bear in mind that r measures only the linear relationship. Insignificant value of r when a close relationship is apparent means that some other mathematical relationship exists e.g. quadratic.

The Spearman's Correlation Coefficient is the non-parametric equivalent of Pearson's r. It is normally employed to correlate a variable whose values are on an ordinal scale with an ordinal response variable.

Often a variety of factors determine an outcome. Powerful statistical software is now available for carrying out multiple regression analysis. There are different techniques for handling different types of variables - nominal, ordinal, discrete and continuous - in such an analysis. A statistician's advice is helpful in deciding upon the most appropriate regression model. It is always a good policy to first try and list all relevant predictor variables at the time of planning the study, and **before** commencing the data collection rather than retrospectively. At the time of the analysis all the variables originally listed which seem reasonable, as predictors of the response are included in the regression model to start with. The regression model is then progressively refined by addition or subtraction of variables until the best combination is obtained.

A number of modifications are possible on the basic regression concept outlined above to deal with special situations. For example, the inclusion of categorical variables by creating dummy variables; interaction between variables, and so on. Discriminant analysis, logistic regression, and Cox's regression for survival data. Multiple regression and other forms of multivariate analysis is described in Research methods – II : Multivariate Analysis.

Appendix 12.1

Approach to data Analysis

Much medical research is about identifying relationship between variables e.g. smoking and coronary artery disease, bottle-feeding and diarrhoea, nutritional intake and growth, bad parenting and emotional problems in children, and so on. An important start is deciding which variables are explanatory, and which represent the outcome. Confusion is created by differences in terminology. For example explanatory variables may be referred to as “independent” or “predictor” variables, and outcome as “response” variables.

The relationship is often complicated by other factors that can be related to both explanatory and outcome variables. These are the confounding variables. Before commencing analysis the researcher must decide on these three types of variables. With experience one is able to make this decision in the very early stage of planning the study.

Exploring relationships

Often the investigator is not interested in differences between groups, but instead in the strength of the relationship between variables. There are a number of different techniques that can be used for the purpose.

Pearson correlation

Pearson correlation is used when one wants to explore the strength of the relationship between two continuous variables. It gives you an indication of both the direction (positive or negative) and the strength of the relationship. A positive correlation indicates that as one variable increases, so does the other. A negative correlation indicates that as one variable increases, the other decreases. A non-parametric equivalent is Spearman’s rank order correlation.

Partial correlation

Partial correlation is an extension of Pearson correlation—it allows one to control for the possible effects of another confounding variable. Partial correlation ‘removes’ the effect of the confounding variable allowing us to get a more accurate picture of the relationship between the two variables of interest.

Multiple regression

Multiple regression is a more sophisticated extension of correlation and is used when one wants to explore the predictive ability of a set of explanatory variables on one *continuous* outcome measure. Different types of multiple regression allow one to compare the predictive ability of a group of explanatory variables and to find the best set of variables to predict the outcome.

Summary

All of the analyses described above involve exploration of the relationship between continuous variables. If there are only categorical variables, one can use the Chi-square test to explore their relationship. In the Chi Square test we are interested in the number of people in each category, rather than their score on a scale.

Exploring differences between groups

There is another family of statistics that can be used when one wants to find out whether there is a statistically significant difference among a number of groups. Most of these analyses involve comparing the mean score for each group on one or more outcome variables. There are a number of different but related statistics in this group. The following tests are used:

T-tests

T-tests are used when there are *two* groups (e.g., males and females) or two sets of data (before and after) and one wishes to compare the mean score on some continuous variable. There are two main types of t-tests. Paired sample t-tests (also called repeated measures) are used when one is interested in changes in scores for subjects tested at Time 1, and then again at Time 2 (often after some intervention or event). The samples are 'related' because they are the *same* people tested each time. The same applies when one is testing matched pairs of subjects. Independent samples t-tests are used when there are two *different* (independent) groups of people (males and females) and one is interested in comparing their scores.

One-way analysis of variance

One-way analysis of variance is similar to a t-test, but is used when there are *two or more groups* and we wish to compare their mean scores on a continuous variable. It is called one-way because the researcher is looking at the impact of only one explanatory variable on the outcome variable. A one-way analysis of variance (ANOVA) will let you know if your groups differ, but it won't tell you where the significant difference is (gp1/gp3, gp2/gp3 etc). You can conduct further comparisons to find out which groups are significantly different to one another. Similar to t-tests, there are two types of one-way ANOVAs: repeated measures ANOVA (same people on more than two occasions), and between-groups (or independent samples) ANOVA, where you are comparing the mean scores of two or more different groups of people.

Two-way analysis of variance

Two-way analysis of variance allows us to test the impact of two explanatory variables on one outcome variable. There are two different two-way ANOVAs: between-groups ANOVA (when the groups are different) and repeated measures ANOVA (when the same people are tested on more than one occasion). Some research designs combine both between-groups and repeated measures in the one study. These are referred to as Mixed Between-Within Designs, or Split Plot.

The decision-making process

In choosing the right statistic to apply to the data one will need to consider a number of different factors. These include consideration of the type of question one wishes to address, the type of items and scales that were included in the questionnaire, the nature of the data that are available for each of the variables and the assumptions that must be met for each of the different statistical techniques. A logical way of deciding how to set about analysing the data set is described below:

Step 1: What questions do you want to address?

Write a full list of all the questions you would like to answer from your research. Obviously these questions are based on the main research question and are subsidiary to it. It is very likely that some questions could be asked in a number of different ways. For each of the areas of interest write the question in a number of different ways. You will use these alternatives when considering the different statistical approaches you might use.

For example, one might be interested in the effect of age of the mother on her knowledge about health of children. There are a number of different ways in which a question can be framed:

- Is there a relationship between age and level of knowledge?
- Are older mothers more knowledgeable than younger?

These two different questions require different statistical techniques. The question of which is more suitable, may depend on the nature of the data that has been collected. So, for each area of interest detail a number of different questions.

Step 2: Find the questionnaire items and scales that will be used to address these questions

The type of items and scales that were included in the questionnaire will play a large part in determining what type of data has been gathered, and hence which statistical techniques are suitable. That is why it is so important to consider the analyses that one intends to use when first designing the study.

For example, the way in which information about respondents' age has been collected will determine which statistics are available to use. If you asked people to tick one of two options (under 35/over 35), your choice of statistics would be very limited because there are only two possible values for the variable age. If, on the other hand, you asked people to give their age in years, your choices are increased because you can have scores varying across a wide range of values from 18 to 80+. In this situation you may choose to collapse the range of ages down into a smaller number of categories for some analyses (ANOVA), but the full range of scores is also available for other analyses (e.g., correlation).

If the study was in the form of a survey, identify the specific questionnaire items and find each of the individual questions (e.g., age) and the scale that you will use in your analyses. Identify each variable, how it was measured, how many response options there were and the possible range of scores.

If the study involved an experiment check how each of the outcome and explanatory variables were measured. Did the scores on the variable consist of the number of correct responses, an observer's rating of a specific behaviour, or the length of time a subject spent on a specific activity. Whatever the nature of the study, just be clear that you know how each of the variables had been measured.

Step 3: Identify the nature of each of your variables

The next step is to identify the nature of each of the variables. In particular one needs to determine whether each of the variables is (a) an explanatory variable, or (b) an outcome variable. This information comes not from the data but from an understanding of the topic area, relevant theories, and previous research. It is essential that one is clear in one's own mind (and in the research questions) concerning the relationship between the variables— which ones are doing the influencing (i.e. explanatory) and which ones are being affected (i.e. outcome). There are some analyses (e.g., correlation) where it is not necessary to specify which variables are independent and dependent. For other analyses such as ANOVA, it is important that one has this clear. Drawing a model of how one sees one's variables relating is often useful here (see Step 4).

It is also important that one knows the level of measurement for each of the variables. Different statistics are required for variables that are categorical and continuous so it is important to know what you are working with. Are the variables:

- Categorical (also referred to as nominal level data, e.g., sex: male/females);
- Ordinal (rankings: 1st, 2nd, 3rd); and
- Continuous (also referred to as interval level data, e.g., age in years, or scores on the knowledge scale).

Additional information required for continuous and categorical variables

For *continuous* variables one should check the distribution (e.g., are they normally distributed or are they badly skewed). What is the range of scores?

If your variable involves *categories* (e.g., group 1/group 2, males/females) find out how many people fall into each category (are the groups equal or very unbalanced). Are some of the possible categories empty?

All of this information about the variables will determine the choice of statistics to use.

Step 4: Draw a diagram for each of the research questions

Diagrams are useful for clarifying thought. The idea is to pull together some of the information that has been collected in Steps 1 and 2 above in a simple format that will help you choose the correct statistical technique to use, or to choose among a number of different options.

Step 5: Decide whether a parametric or a non-parametric statistical technique is appropriate

Statistical techniques that are available are classified into two main groups: parametric and non-parametric. Parametric statistics are more powerful, however they do have more conditions attached, that is they make strict assumptions about the data. For example, they assume that the underlying distribution of scores in the population is normal.

Each of the different parametric techniques (such as t-tests, ANOVA, Pearson correlation) also has other additional assumptions. It is important that one checks these *before* conducting the analyses. The specific assumptions are listed for each of the techniques covered in the remaining chapters of this book.

What if the data don't meet the assumptions for the statistical technique that one wants to use? Unfortunately this is a common situation. Many of the attributes we want to measure are in fact not normally distributed. Some are strongly skewed, with most scores falling at the low end (e.g. blood pressure); others are skewed so that most of the scores fall at the high end of the scale (e.g., body weight).

If the assumptions of the statistic to be used are not met then there are a number of choices available as detailed below.

Option 1

Use the parametric technique anyway and hope that it does not seriously invalidate the findings. Some statisticians argue that most of the approaches are fairly 'robust'; that is, they will tolerate minor violations of assumptions, particularly if you have a good size sample.

Option 2

You may be able to manipulate your data so that the assumptions of the statistical test (e.g., normal distribution) are met. Some authors suggest 'transforming' the data if their distribution is not normal.

Option 3

The other alternative is to use a non-parametric technique instead. For many of the commonly used parametric techniques there is a corresponding non-parametric alternative. These still come with some assumptions but less stringent ones. These non-parametric alternatives (e.g., Kruskal-Wallis, Mann-Whitney U, Chi-square) tend to be not as powerful; that is they may be less sensitive in detecting a relationship, or a difference among groups.

Step 6: Making the final decision

Once the necessary information concerning the research questions has been collected, the level of measurement for each of the variables and the characteristics of the data are all taken into account one is finally in a position to consider the next move. In the text below, the key elements of some of the basic statistical approaches are described.

Exploring relationships among variables**Chi-square for independence**

Example of research question: What is the relationship between membership of WDI and mothers' knowledge of growth charts?

What do you need?

1. One categorical explanatory variable
(E.g. membership of WDI: Yes / No)
2. One categorical outcome variable
(E.g. knowledge of growth charts: Yes/No)

You are interested in the *number* of people in each category.

Set out data as shown below:

	Member of WDI	Not a member
Knowledge		
Yes		
No		

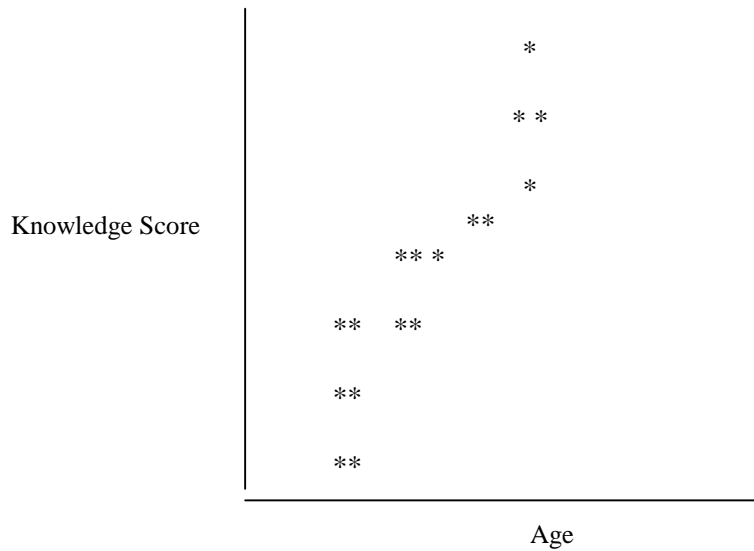
Correlation

Example of research question: Is there a relationship between age and knowledge scores?

What do you need:

Two continuous variables (e.g., age, knowledge scores)

Set out your data as illustrated below:



Non-parametric alternative: Spearman's Rank Order Correlation

Partial correlation

Example of research question: After controlling for the effects of level of schooling, is there still a significant relationship between age and knowledge scores?

What you need Three continuous variables (e.g. age; knowledge score; and level of schooling).

Non-parametric alternative: None

Multiple regression

Example of research question: How much of the variance in knowledge score can be explained by mothers' age, level of schooling, membership of WDI, and socioeconomic status?

What do you need: One continuous outcome variable
Two or more explanatory variables which can be continuous or categorical.

Non parametric alternative: None.

Exploring differences between groups

Independent-samples t-test

Example of research question: Are mothers from villages with WDI more knowledgeable than those from villages without WDI?

What do you need: one categorical explanatory variable with only two groups (e.g., sex: males/females)
One continuous outcome variable (e.g. knowledge score)

(Subjects can only belong to *one group*.)

:
Data is set out as follows:

	WDI Mothers	Mothers without WDI
Mean knowledge score		

Non-parametric alternative: Mann-Whitney Test

Paired-samples t-test (repeated measures)

Example of research question: Does knowledge score improve six months after establishment of WDI activities? Is there a change in knowledge score from Time 1 (pre-intervention) to Time 2 (post-intervention)?

What you need: one categorical independent variable (e.g., Time 1/Time 2)
One continuous dependent variable (e.g., anxiety score)

Same subjects tested on two separate occasions: Time 1 (before Establishment of WDI) and Time 2 (six months later)

Set out the data as below:

	Time 1	Time 2
Mean knowledge score		

Non-parametric alternative: Wilcoxon Signed-Rank Test

One-way between-groups analysis of variance

Ex ample of research question: Is there a difference in knowledge scores for mothers who have no Schooling; primary level schooling; and secondary level schooling?

What you need: • one categorical explanatory variable with two or more groups (e.g. levels of education)
• One continuous outcome variable (e.g. knowledge score)

Set out data as below:

	Level of Schooling		
	None	Primary	Secondary
Mean knowledge score			

Non-parametric alternative: Kruskal-Wallis Test

Appendix 12.2

Ways of reducing systematic errors in measurement.

Source of error	Strategy for reducing error
<hr/>	
Instrument	Calibrate instrument
Observer	Standardize methods.
Observer	Automating the instrument.
Observer	Training the observer.
Subject	Initial preparation of the subject (e.g. overnight fast prior to blood test).
Subject	Making unobtrusive observations (e.g. to check compliance).
Observer Subject	Blinding.

Appendix 12.3

Statistical analysis by types of variables

Outcome variables

Predictor Variable	Continuous (normally distributed)	Continuous (not normally distributed)	Nominal with >2 categories	Dichotomous
Continuous (normally distributed)	Correlation - Linear regression	Spearman Rank correlation	Analysis of variance	Logistic regression
Continuous (not normally distributed) OR Ordinal with >2 categories	Spearman Rank correlation	Spearman Rank correlation	Kruskal Wallis	Wilcoxon rank Sum
Nominal with >2 categories	Analysis of variance	Kruskal Wallis	Chi-squared test	Chi-squared test
Dichotomous	't' test	Wilcoxon rank Sum	Chi-squared test	Chi-squared test

Appendix 12.4

Tests of Significance for determining differences between groups

Data Type	Unpaired Observation	Paired Observation
<u>Nominal data</u>		
Small sample	Fisher's Exact test	Sign test
Large sample	Chi-squared (χ^2) test	McNemar's test
<u>Ordinal data</u>		
Two groups	Wilcoxon test Or Mann-Whitney U-test	Wilcoxon signed-rank test
More than two groups	Kruskal-Wallis test One-way ANOVA	Friedman 2-way ANOVA
<u>Numerical data</u>		
Two groups	't' test One-way ANOVA	Paired 't' test
More than two groups		

Appendix 12.5

Test for Measuring association between variables

Data Type	Test	Significance
Nominal	Chi-squared (χ^2) test	Odds Ratio Or Relative Risk
Ordinal or numerical (without linear relationship)	Spearman's 'r' Or Kendal's tau	Significance of Spearman's 'r' Or Kendal's tau
Numerical (with linear relationship)	Pearson's correlation coefficient	Significance of Pearson's 'r'