INTRODUCTION TO EXPERIMENTAL RESEARCH

Inferential Statistics

In the previous two chapters we have seen how the sample size, mean and standard deviation provide a shorthand description of data that is normally distributed. These descriptive statistics allow comparisons to be made between different sets of data in terms of their typical scores and how the data is distributed. However, researchers often not only want to describe differences and similarities between sets of data, but also want to make inferences. For example, they may wish to test whether any observed differences reflect the impact of some causal factor or test to see if they can infer a particular sort of relationship between the variables that have been measured. To illustrate this, imagine that it is observed that a group of people who undertake a particular set of exercises have lower mean systolic blood pressure than those who do not. A researcher may want to know if it can be concluded that this observed difference is a product of the exercise training rather than just a product of the variability found in systolic blood pressure within the general population. Or, again, if a researcher observes that there is a tendency for those who are more perfectionist also to be more anxious before examinations, the researcher may want to know if this relationship is due to this personality disposition or just a coincidence. While statistics can aid in this decision making it is the logic that informs how the data is collected in conjunction with the statistics that justifies any claim for there being a causal relationship. This is most simply illustrated in the principles of experimental research.

A Brief Introduction to Experimental Research

To understand what is involved in testing predictions concerning data it is worth remembering that any human enquiry is concerned with gathering information to enhance understanding. An experimental approach is just one method, amongst many, of gathering data. For example, information can also be gathered through surveys, psychometric testing or case studies. However, an understanding of experimental research highlights both the role of statistics in aiding decision making and their limitations.

In an experiment the aim is to test a **hypothesis**. A hypothesis is a testable statement that predicts the relationship between two types of variables. These variables are termed the **independent** and the **dependent** variables. The independent variable

is the one that is manipulated and the dependent variable is the one that is measured. For example, if it is predicted that a certain nutritional drink influences performance on an endurance activity, then the consumption of the nutritional drink is the independent variable and the participants' performance is the dependent variable. The experimenter would then measure the dependent variable (i.e. performance) under at least two conditions or levels of the independent variable. These two conditions might comprise a control group (e.g. performing without the nutritional drink), and an experimental group (e.g. performing with the nutritional drink).

Of course, endurance performance is a product of many variables other than the type of drink a participant may have consumed. To rule out the effect of these other variables participants must be tested in a situation where all these other variables, which are known as **irrelevant** or **extraneous variables**, are held constant. These extraneous variables are most often to do with aspects of the situation or the participants that have the potential to influence the dependent variable.

The control of **situational** variables, such as performance conditions, noise and time of day, etc., can generally be achieved with careful experimental design. Advertisements for washing powders provide a simple example of this procedure. Advertisers are usually careful to show that identically soiled clothing is washed in two identical machines and that the only variable that does differ is the washing powder. Sometimes extraneous variables can be a product of the testing procedure itself, for in the above example concerned with a nutritional drink, the testing of the participants has the potential to produce dehydration. If the control group does not have a drink this effect will be larger for this group in comparison with the experimental group. Therefore if the dehydration of the participants in the control group is thought to be a design problem then this group might be given a volume of water equal to that of the nutritional drink.

The other major type of extraneous variables reflect individual differences in participants, consequently they are known as **participant** variables. To illustrate these let us return to the example with the washing machine advertisement. If two articles of clothing are of different levels of cleanliness prior to washing, then any difference in cleanliness after washing may be a product of these original differences rather than the washing powders being tested. In the washing of clothes a soiled garment can be torn in half and one half put in each machine but with human participants individual differences are more difficult to eliminate. The most successful method is to adopt what is known as a **repeated measures design**. In this design the same person is measured on repeated occasions. On each occasion the participant is subjected to a different level of the independent variable. In the context of the previous athletic example concerning the impact of a nutritional drink on endurance performance, this would mean that participants may first perform after consuming a drink of water and then be required to perform again after consuming the nutritional drink. Any differences observed in the dependent variable of performance under the two conditions cannot then be a product of participant variables as the same individuals have been used in both of the different conditions.

Unfortunately this strategy can introduce other **confounding factors** such as **order effects** resulting, for example, from participants being exhausted, bored or having learnt from engaging in the activity in the previous condition. If these effects are symmetrical in that they are the same whichever order the individual is tested in, then counterbalancing may control for them. For example, if there are two conditions then half the participants are tested in one condition first and the other half in the other condition first. In the context of the previous example this would mean that half the participants would first run after consuming the nutritional drink and then run again after the drink of water, whilst the remaining participants would perform in the opposite order.

If the order effects are not symmetrical then they cannot be effectively counterbalanced and a different strategy has to be adopted to control for individual differences. One such strategy to overcome this problem is to employ a **matched pair design**. In this design different participants are selected to perform under the two or more levels of the independent variable but they are matched in pairs across the conditions on the important individual difference, for example running ability, or body mass. This does not mean that the individuals within each group have to be similar, rather each participant in one condition has to have a matched participant in the other condition. For example, there can be a fast, medium and slow runner in one condition as long as there is a fast, medium and slow runner in the other condition(s). This can be a time-consuming procedure because participants have to be pre-tested on the variables of interest before they can be matched. Of greater importance is the fact that the assumption regarding the relevant variable(s) for matching the participants may be incorrect. For example, they may have been matched on skill when physical fitness is the important factor upon which matching should take place. Unless the assumption in the matching forms part of the hypothesis, as it does for example in many studies of intelligence and personality in twins, then the researcher cannot be confident that the correct factor has been chosen. Later on in this chapter the importance of this will be highlighted. It will be noted that the statistical tests that are employed to examine differences between experimental and control groups are exactly the same for data produced by a matched pair design as those that would be employed to data from a repeated measures design. These tests assume that any differences in the groups of data are not a product of individual differences. Unfortunately, if participants have not been matched on the correct variable(s) then any observed differences could be the product of individual differences.

An alternative solution is to adopt an **independent groups design** in which the participants are **randomly** allocated to the various levels of the independent variable. Random assignment means that each participant has an equal chance of entering any group. The logic of this method accepts that there are individual differences but argues that these differences are distributed across the two or more groups by chance. This random allocation means that it is unlikely that there will be any systematic bias in any group. For example, it is most likely that high- and low-fitness participants will be more or less evenly split across two groups and very unlikely that all high-fitness participants will be in one group and low-fitness participants in the other group. The analysis of data from an independent group design employs different statistical tests to those employed for data collected by means of repeated measures or matched pair designs. These tests recognise that the distribution of individual differences such as fitness across the two groups may not be identical and therefore variability within each group is taken into account in the test design. This will be explained in more detail later.

Having hypothesised a relationship between the independent and dependent variables, and chosen a design to control for extraneous factors, the researcher can now collect the data. The data gathered is then subjected to the appropriate statistical analysis. As discussed later in this chapter, this analysis measures the precise probability of getting the observed differences in the dependent variable under the various levels of the independent variable by chance. If there is a relatively small probability that the observed differences occurred by chance, usually less than 5%, then the logic of the experimental design dictates that we attribute the difference to the independent variable. That is, the decision is that there is a **causal** relationship between the independent and dependent variables. To understand this process further the following section explores the logic of statistical inference in more detail.

The Logic of Statistical Inference

In an experiment, a prediction is tested by manipulating the independent variable and measuring changes in the dependent variable while attempting to control extraneous variables. This provides at least two sets of data, one from the control group and one from the experimental group. The decision to be made is whether there is a **significant** difference in the two sets of results that can be attributed to the influence of the independent variable, or whether the difference is due to irrelevant extraneous variables. Thus, statistical testing leads to **inferences** being made from the data as to the **relationship** between the variables.

What is meant by the word 'significant' above? To explain this term, imagine that various weights are put on a mechanical scale and the relationship between the size of these weights and the movement of the pointer is observed. The causal relationship between these two events can be explained by an understanding of the mechanics of the scales. If such an explanation is insufficient the scales can be opened up and the physical relationship between the weights and the pointer movement observed. But in an experiment concerning the effect of a nutritional drink on performance, the human body cannot be 'opened up' in the same way as the scales, nor can a complete explanation of the 'mechanisms' at work be given with the same degree of precision as with the scales. All that can be done is to attempt to keep the irrelevant variables constant and manipulate the independent variable (e.g. the nutritional drink). If a corresponding change in the dependent variable (e.g. performance) is now observed then the logic of the experimental design which has controlled for extraneous variables leads to the inference that a causal relationship exists between the nutritional drink and performance.

Unfortunately, *all* the irrelevant variables can never be kept absolutely constant and thus it can never be proved beyond doubt that chance was not responsible for any changes in the independent variable. Here the word 'chance' is being used as shorthand for irrelevant or extraneous variables. However, by conducting statistical tests on the data, the likelihood that any observed difference is due to chance can be measured. For example, it may be found that the amount of variation in performance between those ingesting the nutritional drink and those who did not ingest the nutritional drink is likely to occur 50 times out of 100 by chance. If this were the case then it would be difficult to credit the nutritional drink as being responsible for the variation. If, on the other hand, the probability of the variation being due to chance was 1 in a 100 then the decision might very well be made to attribute the variation to the independent variable, namely the nutritional drink, for this explanation would be correct 99 times out of 100. The point at which it is decided to reject the chance explanation and attribute responsibility for the difference to the independent variable is called the **significance level**. The significance level is also known as **alpha** (α) for reasons which will become clear later. Most researchers generally choose a significance level of 0.05. That is, as long as the observed differences would only occur by chance 5 times or less out of 100 then the researcher will reject the chance explanation for the observed findings. This is because 95 times out of 100 the observed differences would be due to the experimental manipulation of the independent variable. To some extent the adoption of an α equal to 0.05 (or 5% significance level) is an arbitrary matter and in certain cases one might be far more conservative and choose a level of 0.01 where the probability due to chance is 1 in 100.

Hypothesis Testing

This process of rejecting or accepting the chance explanation is known as hypothesis testing. In the process of statistical inference the researcher is deciding whether:

1. the differences arise because of purely chance fluctuations in the two groups of scores; or
2. the differences are caused, at least in part, by the independent variable.

The first of these is termed the **null hypothesis** and the second is called the **alternate** or **experimental hypothesis**. Thus, in an experiment concerning a nutritional drink and performance, the null hypothesis would be:

> H_0 There will be no significant difference between the mean performance scores of the experimental and control groups.

The alternate hypothesis might state:

> H_1 There will be a significant difference between the mean performance scores of the experimental and control groups.

If, as in the above experimental hypothesis, the direction of the difference is not stated then it is known as a **two-tailed hypothesis**. However, if the alternate hypothesis states the direction of the difference in the two groups of scores, for example the experimental group will score less points for their performance than the control group, then this is known as a **one-tailed hypothesis**.

Unless there are good reasons for doing otherwise, a two-tailed hypothesis should be postulated. One-tailed hypotheses are appropriate when a theoretical perspective based on previous research would suggest that a particular difference is expected or when there is no logical way in which a difference could occur in a direction other than the one predicted.

Type I and Type II Errors

If a significance level of 0.05 is chosen, it is important to realise that on 5% of occasions the observed difference may occur by chance. This means that if the null hypothesis is rejected then the researcher can be 95% confident that the difference found is not the product of chance and there is only a 5% probability of making a mistake. Making the mistake of rejecting the null hypothesis when it is in fact true is known as committing a **Type I error**. The probability of making such a mistake is known as the alpha level and is equal to the significance level adopted.

An obvious solution to the problem of committing a Type I error is to reduce the significance value to, say, 0.01. This would mean that a Type I error would be expected only one time out of a hundred experiments. However, this would result in an increased number of occasions on which the null hypothesis is accepted when it is in fact incorrect. This is known as a **Type II error**. The probability of committing a Type II error is referred to as the beta level (β) and its probability is much more difficult to assess than the alpha level. In spite of this difficulty, it can be seen from the above that a significance level of 0.05 is a compromise that attempts to minimise the probability of committing either of these two types of errors. Another way of expressing the connection between these two errors is in the form of the two scales shown in Figure 4.1.

From these scales it can be seen that in employing a 5% significance level the researcher is accepting that there is a 5% probability that H_0 is correct and a 95%

Scale indicates the probability that H_0 is *correct*

Sig. level

100%..5%..................0%

Decreasing probability of committing a Type I error →

Scale indicates the probability that H_0 is *incorrect*

0%..95%..............100%

← Decreasing possibility of committing a Type II error

Figure 4.1 **The relationship between significance level, confidence level and Type I and II errors**

probability that H_0 is incorrect. It can also be seen that any attempt to reduce the probability of committing a Type I error by reducing the significance level will result in an increase in the possibility of committing a Type II error.

Statistical versus Practical Significance

When conducting research that involves statistical testing there is a great temptation to suggest that statistical significance automatically implies a practical significance. For example, imagine we found that the mean age at death for those who smoked 30 cigarettes per day was 65 and for those who smoked 40 cigarettes per day it was 63. Also imagine that this difference of two years was statistically significant at the $p < 0.05$ level. However, is this a practically significant finding? To answer this question this difference needs to be examined in the context of the greatly reduced life expectancy of both groups compared with non-smokers whose mean age at death was found to be 78 years. Or, again, imagine that heavy users of digital mobile phones have a 20% increased probability of developing a brain tumour and that this increase is statistically significant. This sounds very worrying. However, if the probability of developing a brain tumour in the wider population is 5 in 100,000 then this research would suggest that the risk for mobile phone users is raised to 6 in 100,000. You may feel that the practical implications of this statistically significant increased risk are not sufficient to deter you from continuing to have the convenience of mobile telephony. The distinction between practical significance and statistical significance is very important and the researcher must be very careful to restrict the term significance to its statistical denotation when writing up experimental reports.

Hypotheses Concerned with Differences and Relationships

Each of the examples of hypotheses given so far has included the word 'difference'. For example, a null hypothesis might state that there will be no significant **difference** between the performance scores of the experimental and control groups. This hypothesis predicts that the means of the two groups will not be significantly different.

Sometimes, however, the focus of interest is not in detecting group differences but rather in identifying relationships between variables. For example, our null hypothesis might predict that there will be no significant **relationship** between the length of the training time and the performance scores. This null hypothesis predicts that knowing how much someone trains will not provide any information about how well they perform.

The distinction between hypotheses concerned with relationships and hypotheses concerned with group differences is very important. This is because they reflect the type of question the researcher asks, which in turn will often influence the way that data is collected. Questions concerned with differences usually employ data collected through experimentation involving the manipulation of one or more independent variables. If through the use of the appropriate statistical tests a significant difference is found then the conclusion may be drawn that there is a causal relationship between the independent and dependent variables.

In contrast, research focusing on relationships often involves data collection through survey methods in which no manipulation of independent variables and limited control of extraneous variables have taken place. Attempts to answer questions concerned with relationships employ different statistical tests to questions focusing on differences. If through the use of the appropriate statistical tests a significant relationship is found then the conclusion that there is a causal relationship between the variables concerned may not automatically be drawn, for the lack of control and manipulations of extraneous and independent variables make it difficult for the researcher to rule out the possibility that some other variable caused the observed significant relationships. The distinction between questions that focus on differences and questions that focus on relationships is one very important factor in determining the type of statistical analysis to be employed and the types of conclusion that can be drawn.

An introduction to statistical tests concerned with relationships between variables will not be presented until Chapter 10. However, it is important to appreciate that whenever a significant difference between two variables exists there will always be a relationship between the two variables.

Samples and Populations

It was stated earlier that the problem facing the researcher when analysing data was one of deciding whether there is any significant difference in the results that can be attributed to the influence of the independent variable, or whether any difference is due to irrelevant variables. Before examining the statistical techniques used to make these decisions it is necessary to understand some of the constraints within research.

When experiments or surveys are undertaken it is not usually possible to employ the total **population** as participants and therefore the researcher relies on data collected from a **sample** of this population. These two words, population and sample, represent very important ideas in statistics. The term population refers to all possible objects of a particular type. Such types might be tennis players, cricket bats or characteristics such as fitness levels. Populations are so large, in fact they may be infinite, that a researcher never has access to them and instead often takes a sample of observations from the population. Unless the sample is representative of the population of interest then findings cannot be extrapolated from the sample and applied to the larger population.

To illustrate this imagine that a researcher at a university wanted to understand how first-year undergraduate students perceived and participated in leisure activities. For obvious practical reasons the questionnaire could not be administered to all undergraduate students in the country of interest so instead a sample of those students has to be employed. Unfortunately, were the researcher to take the most convenient course and administer the questionnaire to all students at his/her own university this would not provide a representative sample. This is because students attending any one particular institution are a self-selecting group that are unlikely to represent a cross-section of the student population. If, instead, the researcher took a **random** sample of students from every institution with undergraduates, then as every undergraduate student in the country had an equal chance of being selected, this would usually provide a representative sample. If the researcher found that the mean time spent on leisure activities was 32 hours per month, then he/she would be justified in inferring that this was an accurate estimate of the average amount of leisure time per month engaged in by the student population. If this process was repeated several times the researcher would probably find a slightly different mean value each time as a result of random variations in the process of sample selection. The actual mean of the country's undergraduate population would lie somewhere close to the average of this series of sample means. This is because the means of the samples will be normally distributed around the population mean. Thus, whenever two samples are taken from the same population we would expect the means to be different. This is a very important point whose statistical implications will be explored in more detail in the next chapter.

A clue to this importance can be found by returning to the experiment concerning a nutritional drink and performance. Here, one might have one set of performance scores from a control group who just drank water and another set of performance scores from the experimental group who had taken the nutritional drink. If the means of these two data sets are different what are the possible explanations? From the previous discussion on samples it follows that any observed difference in means could be found because the means from any two samples of a population will often produce different results. This explanation would suggest that the null hypothesis is correct. Alternatively it could be that the means are different because the samples are drawn from two different populations. If this is the case then the experimental hypothesis is correct. The experimental hypothesis implies that one set of performance figures are produced by a population that performs without the nutritional drink and another set by a population that performs with the nutritional drink and these two populations have different performance statistics.

Which of these explanations is correct? The various statistical tests outlined later on in this manual are designed to answer this question. All the tests do this by working out the probability that the results could be obtained from the same population. If this probability is equal to or less than 0.05, then there is less than a 5 out of 100 chance that these results could be obtained from the same population. In this case the researcher would reject the null hypothesis and accept the experimental hypothesis. Before we attempt to understand how these tests calculate this probability we need to know first how to select the appropriate test.

Selecting the Appropriate Statistical Test

There are four major factors that determine the correct test for any particular set of experimental data. These are:

Table 4.1

		Type of design	
Type of test	**Type of data**	**Related samples**	**Independent design**
Parametric	Interval	*t*-test for paired samples	*t*-test for independent samples
Non-parametric	Ordinal	Wilcoxon test	Mann–Whitney test
	Nominal	Sign test	Chi-square test

1. What is the research question?
2. How many dependent and independent variables are there?
3. Was the data gathered by means of an independent, related or mixed design?
4. What type of scale was used to measure the variables?

In Appendix 2 there are two tables that illustrate how these four factors combine to determine the correct choice of test. However, this section concentrates on points 3 and 4 above and their implications for choosing among a limited range of statistical tests. Before examining what this means in detail note that Table 4.1 illustrates how these two factors influence the choice of two categories of statistical tests.

In looking at this table it is important to remember that the term related samples refers to repeated measure and matched pair designs and the term independent samples refers to the independent group design. If you are not clear about these terms you will find a reference to them in the second section of this chapter.

In the table the terms **parametric** and **non-parametric** tests are introduced. These refer to two basic types of statistical tests that make different assumptions about the data being examined. All that can be noted at present is that the choice of tests is determined in part by how the data is collected and what sort of data is collected.

Parametric Tests

Parametric tests make three basic assumptions:

1. It is assumed that each sample of scores has been drawn from a **normally distributed population**.
2. These samples are assumed to have the same **variance**.
3. The dependent variable is assumed to have been measured on an interval scale.

The following section examines each of these assumptions in more detail.

Checking the Assumption of a Normal Distribution

Unfortunately, it is difficult to prove that a particular sample is drawn from a normally distributed population. Previous research can be examined to see if the data previously formed a normal distribution or the existing data could be plotted in the form a histogram to see if it looks like a normal distribution. However, neither of these procedures is very precise. Therefore, to have any measured confidence in our decision we need to look for statistical alternatives. Two measures that are relevant are the concepts of **skewness** and **kurtosis**. You may have noted already from your printout of Exercise 3.2 that the **Explore** option provides a measure of these statistics. Each of these concepts will now be dealt with in turn.

Skewness

Skewness reflects the existence of extreme scores at one end of the distribution. A skewness of zero means that the distribution is symmetrical. Using the z-table (Table 3.1) it is possible to get an exact measure of skewness and to assess the probability that the distribution is symmetrical. Whilst SPSS will produce this value it is important to understand how it was obtained so that its significance can be interpreted.

Skewness is computed by employing the following formula:

$$\text{Skewness} = \frac{\sum z^3}{n}$$

Here z refers to the distance of each score from the mean measured in units of standard deviation. By cubing this value the sign, whether it be positive or negative, will remain the same but the deviation of extreme scores will greatly increase in magnitude. For example, 1 cubed is 1, 2 cubed is 8 and 3 cubed is 27. The reason for dividing by n is to find the average of the cubed deviations and therefore to control for the sample size.

The larger the value obtained from this calculation the greater the skewness. The particular meaning of any value obtained by this calculation is not obvious as theoretically it could be any value between zero and infinity. What is needed is a scale so that the probability that the distribution is skewed or not can be assessed. Without wishing to pre-empt a later discussion on sampling error, the estimate of skewness within the population is based on the sample of data. This sample may not be representative of the population as it may contain some error. To assess the significance of the value of skewness the ratio of the skewness obtained over the error in its measurement is calculated. This error is known as the standard error for skewness. For moderately large samples this standard error for skewness can be computed by the following formula:

$$\text{SE}_{\text{skew}} = \sqrt{\frac{6}{n}}$$

Table 4.2 **Descriptives**

	CLASS		Statistic	Std. Error
Grades	Professor's Class	Skewness	2.276	.491
	Graduate Student's Class	Skewness	.270	.491

If the value of skewness is divided by this error the result is a z-value for skewness that can be interpreted using the standard z-tables:

$$z_{\text{skew}} = \frac{\sum z^3/n}{\sqrt{(6/n)}}$$

Note from the z-table (Table 3.1) that with a move of 1.96 standard deviations away from the mean an area of 2.5% of the normal distribution curve is cut off. To check that a distribution is neither positively nor negatively significantly skewed ($\alpha = 0.05$) both ends of the normal distribution must be examined. Therefore as long as the value of z_{skew} is less than ± 1.96 there is 95% confidence that the population distribution is not positively or negatively skewed. To illustrate what this means in practice consider an example.

The output from the SPSS exercise in the previous chapter includes the information given in Table 4.2. Note that if the standard error for skewness is calculated according to the previously given formula the value 0.562 will result and not the value 0.491 shown in Table 4.2. This is because that formula can only be applied to moderately large samples. Also note that for 'Professor's Class' if the skewnes statistic is divided by the standard error for skewness the result is 4.64. As this is much larger than ±1.96 the conclusion is that these 'Professor's Class' scores are significantly skewed and therefore the distribution significantly deviates from normal. Is this also true for the 'Graduate Student's Class'?

Kurtosis

Kurtosis refers to the peakedness of the curve: that is, is it a flat curve or does it have a sharp point? The procedure for working out the kurtosis is very similar to that employed when working out skewness. The formula is as follows:

$$\text{Kurtosis} = \left(\frac{\sum z^4}{n}\right) - 3$$

Note that here z is raised to the power of 4. This means that, once again, extreme scores will be greatly increased, but this time the signs will all be positive. For example, −1 raised to the power of 4 is 1, −2 raised to the power of 4 is 16, and 3 raised to the power of 4 is 81. The −3 is included in the above kurtosis equation because without it, if there were no kurtosis, this formula would produce a value of 3. Therefore, 3 is subtracted so that if there is no kurtosis a value of zero is produced.

The next step is to work out the likely error in estimating the kurtosis within the population from the sample and this is done as follows:

$$\mathrm{SE}_{\mathrm{kurt}} = \sqrt{\frac{24}{n}}$$

Then to compute a z-value for kurtosis we divide the kurtosis score by the above error term:

$$z_{\mathrm{kurt}} = \frac{\text{kurtosis}}{\sqrt{24/n}}$$

As with skewness, a value between ±1.96 suggests with at least 95% confidence that the distribution is normal. If the value obtained is positive, then it indicates that the curve is more peaked than normal (i.e. leptokurtic), whilst if it is negative, then the curve is flatter than normal (i.e. platykurtic). Again, looking at the SPSS exercise in the previous chapter may be helpful.

Looking at the output in this exercise, note that it includes the information in Table 4.3. Using the 'Professor's Class' values in the table, the kurtosis statistics divided by the standard error for kurtosis produce the value 5.93. As this is much larger than ±1.96 it would be concluded that these 'Professor's Class' scores are significantly kurtotic, that is the distribution significantly deviates from normal. Is this also true for the 'Graduate Student's Class'? Now look at the two histograms from the output of the previous SPSS exercise and see if this graphical representation concurs with this statistical interpretation.

By using these two measures of skewness and kurtosis the researcher can decide, with a precise level of confidence, whether the data conforms to the required assumption of being normally distributed.

Table 4.3 **Descriptives**

	CLASS		Statistic	Std. Error
Grades	Professor's Class	Kurtosis	5.644	.953
	Graduate Student's Class	Kurtosis	.236	.953

Checking the Assumption of Equal Variance

The assumption of equal variance is also known as the homogeneity of variance assumption. Some indication of the equality of the variances of two sets of data can be obtained by looking at a graphical representation such as a boxplot, an example of which was provided in Chapter 2. Although it cannot be proved that two samples have equal variance, it can be determined whether they are significantly different. SPSS employs a variety of procedures to test for equivalence of variance and these tests will be discussed as they become appropriate for particular statistical techniques.

Checking the Assumption of Level of Measurement

In Chapter 2 it was noted that measurement is concerned with assigning numbers to observations. How numbers are assigned to observations determines the arithmetic operations that are permissible. The ways in which numbers are assigned in social research can be classified into three different levels. These are:

1. **Nominal data**. Nominal means in name only. Here numbers are used simply to identify the groups or categories to which various objects belong. For example, the number 1 could be assigned to windsurfing, 2 to dinghy sailing and 3 to land yachting. No numerical operations can be carried out with these numbers beyond noting the frequency with which each occurs. These relative frequencies allow the mode to be identified as an indication of the typical score.
2. **Ordinal scale**. In an ordinal scale objects are not only in different categories, but these categories stand in some relation to each other; for example, students rank their preferences in order for different sorts of leisure activities. However, there is no reason for assuming that the difference between the items ranked first and second is the same as that between those ranked second and third etc. As the data stands in an ordered relationship to one another, the median is used to identify the typical score.
3. **Interval scale**. The relationship between the objects is measured on a scale in which all the intervals are equal. For example, if performance is measured in terms of time then the difference between 28 seconds and 29 seconds is the same as that between 34 and 35 seconds, namely 1 second. This system of measurement allows for the mean to be calculated as the indicator of the typical score.

Because parametric tests are used to examine for significant differences between means, they require that the dependent variable be measured on an interval scale.

Non-parametric Tests

In contrast to parametric tests, non-parametric tests make very few assumptions. In fact they make few assumptions about the distribution of the data and they do not assume homogeneity of variance in the samples. The only assumption that is of concern here is to do with the scale on which the dependent variable is measured. Looking at Table 4.1 above it can be seen that some non-parametric tests require that the dependent variable is measured on an ordinal scale whilst others require nominal data.

As non-parametric tests make so few assumptions compared with the corresponding parametric tests then why use parametric tests? The answer is because parametric tests are more likely to detect a significant difference between two sets of scores than their non-parametric equivalent. This is because more information is available to parametric tests than non-parametric tests. For example, whereas a non-parametric test based on ordinal data takes into account the rankings of the participants' scores, a parametric test would not only have this information but also the magnitude of the intervals between the rankings. It is for this reason that parametric tests are more **powerful** than their non-parametric equivalent. Power in this context is defined as the probability of rejecting the null hypothesis when it is false and is defined in the following way:

$$\text{Power} = 1 - \text{probability of Type II error} = 1 - \beta$$

Although parametric tests make the three assumptions listed above, they will still provide a valid result even when some of these assumptions are not perfectly met. Tests that possess this characteristic are described as **robust** and the reason for this quality will be explained later.

Finally, although non-parametric tests are not as powerful as parametric tests, increasing the sample size can increase their power to that approaching their parametric equivalents.

Summary

Statistics are employed to enable the researcher to draw inferences from samples of a population about whether there are real differences between sets of scores or real relationships between variables. In experimental research, as the design should have controlled for most of the extraneous variables, any observed differences in the dependent variable may be attributed to either changes in the independent variable or error in the measurement due to factors such as individual differences. In deciding whether to attribute these changes to the independent variable statistics are employed to measure the probability of obtaining the differences by chance. This chance explanation is known as the null hypothesis and it postulates either that no relationship exists between the variables or that no differences exist between the sets of scores. If the probability of obtaining the observed differences by chance is equal to or less than the established significance level (e.g. 0.05) then the decision is made to reject the chance explanation and accept the alternative that a true difference exists. If a criterion of 0.05 is employed in rejecting this chance explanation then there is always a 5% probability that the wrong decision has been made. This is known as a Type I error. Alternatively, if the probability was greater than the criterion and the null hypothesis was accepted by mistake, this is known as a Type II error. Even if the researcher decides that there is a real statistical difference between sets of scores this does not mean that this difference is of any practical significance.

The calculation of this probability for accepting or rejecting the null hypothesis varies depending upon the statistical tests. The appropriate test is determined by the research question, the number of dependent and independent variables, the experimental design and the type of scale used to measure the variables. All of the tests covered in this book are known as parametric tests and these tests make three assumptions. They assume that each sample of scores has been drawn from a normal population, that these samples have the same variance and that the dependent variable is measured on an interval scale. The concepts of skewness and kurtosis were introduced as statistical means of evaluating the normality assumption.

EXERCISE 4.1 TEST SELECTION

Below are listed a series of hypothetical situations. In each case indicate which test the researcher should use and why. While this task is made easier if you are familiar with the tests, it is still not that difficult to make the appropriate choice if you ask yourself two things:

A. What experimental design has been adopted?
B. What sort of data has this research generated?

Having answered these questions use Table 4.1 to identify the appropriate test. To keep matters simple, assume that the homogeneity of variance assumption is met and that the data is normally distributed.

1. Which statistical test could be used to test the hypothesis that there is no difference in the number of goals scored by the top 22 first and second division goal scorers? Explain your choice.
2. A geographer who was interested in the apparent reduction of rainfall in East Anglia compared the rainfall in 1991 with that of 2001.
   Which statistical test should she use and why?
3. An ecologist predicted that humidity would influence the amount of time spent foraging by woodlice. He set up an experiment in which a sample of 100 woodlice was randomly assigned to one of two humidity conditions. He then observed how much time they spent foraging over a 24 hour period. The humidity conditions were then reversed for the two groups and the foraging behaviour was once again observed for 24 hours.
   Which statistical test should he use to test his hypothesis? Explain the reasons for your choice.

4. A sociologist was interested in exploring whether there was any association between political choice and religious affiliations. She set up a survey of a representative sample, noting on a tally chart respondents' religion, and which party they usually voted for.
   Which statistical test should she use and why?
5. A psychologist wished to test the hypothesis that the vigorousness of initiation procedures does not influences the degree of affiliation to a group. He persuaded a local employer to assign applicants randomly to one of two methods in selecting staff. One method lasted 30 minutes and the other lasted one day. Ten staff were selected from each group of applicants, the actual selection being based on 30 minutes of testing which was common to both groups. Later the successful applicants were asked to rate the quality of the organisation's work on a scale of 1 to 10.
   What statistical test would be used to test this hypothesis and why?
6. A sociologist wanted to compare the career aspirations of children from working class and professional backgrounds, so 85 'working class' children and 77 'professional' children were randomly selected from a group of homogeneous intellectual capacity. The aspirations of each child were rated by an independent assessor on a scale of 1 to 10 (1 = unskilled, 10 = high-status profession).
   How would one test the hypothesis that the children of parents from the two occupational groups do not have different career aspirations?
7. A newly discovered apple is delicious in flavour. It was decided to test its yielding capacity by planting the new apple adjacent to a standard apple in eight orchards scattered about a region suitable for the cultivation of both apples.
   How would one test for a significant difference between the yielding capacity of the two varieties of apple trees?
8. An historian was interested in changes in the age distribution of a population as a consequence of the Black Death. Focusing on two villages that were matched for size, population structure, etc., before the plague, she noted that one experienced the plague and other did not. She then examined the appropriate census data to see if there was any difference in the number of people aged over 18.
   What test could she use to see if there was a difference in the structure of the population after the plague? Explain your choice.
9. A film on the hazards of smoking was shown to 800 smokers. One week after seeing the film 28 out of the 300 women smokers and 32 of the 500 men smokers were still not smoking.
   What statistical test could be used to test the association between the sex of the participants and giving up smoking?
10. A researcher was interested in whether a particular film on AIDS would influence attitudes towards 'safer sex'. The attitude of a representative sample of students was assessed on a seven-point scale before and after viewing the film.
    What test could one use to see if there was a difference in their attitudes on these two occasions?

EXERCISE 4.2 USING THE NORMAL DISTRIBUTION

This is another exercise that employs the concepts introduced in Chapters 2 and 3. It is included here because a sound understanding of descriptive statistics and the normal distribution is essential for anyone who wishes to employ statistical tests.

1. If the mean amount of beer drunk by male students is 15 pints per week and the standard deviation is 3 pints, then:
   (a) What percentage of students consume more than 22 pints per week?
   (b) What amount of beer do more than 75% of students consume per week?
   (c) What percentage of students consume less than 7 pints per week?
   (d) What percentage of students consume between 10 and 20 pints per week?

2 If 40% of students spend more than 20 hours per week reading and the standard deviation is 6 hours per week, then:
   (a) What is the average amount of time spent reading?
   (b) If the university expects students to read for 30 hours per week, what percentage of students fulfil this requirement?

3 In a population of 230,000 students, the mean debt that they have accumulated by the time they graduate is £6400 and the standard deviation is £1600.
   (a) How many students graduate with a debt greater than £10,000?
   (b) How many graduate with a debt of less than £1000?
   (c) How many graduate with a debt of between £4000 and £8000?