Department of Psychology
University of York

Year 1 Module
Data Analysis

Statistics Notes

©Andrew Monk

October 2000

CONTENTS

Chapter 1

Experiments and variables

Statistical concepts introduced in this Chapter:

Continuous and nominal variables, independent variable, dependent variable, levels of an independent variable, random effect, fixed effect.

An experiment

Psychologists and other behavioural scientists use statistics to test hypotheses. This section will discuss how an hypothesis is turned into an experiment. Consider the idea that, despite the claims of some manufacturers, people prefer butter to margarine. Now imagine you are devising an experiment to test such a hypothesis. Let us say that we are interested in measuring preference and that we have each taster put a mark on a 100 mm line as depicted in Figure 1. Measuring the distance of the mark from the 'very nasty' end gives us a preference rating, a measure of how much they like what they have tasted. We can then get two groups of tasters, one given margarine and the other butter. Our hypothesis will be supported if the average rating of the group having butter is consistently higher than that of the group having margarine.

Very nasty |____| Very Nice

Figure 1.1 Preference rating scale

Variables

What could affect the results of this experiment? Temperature might be important. Perhaps people prefer margarine at room temperature and butter slightly chilled, or vice versa. In this experiment temperature can be described as a 'variable' which needs to be 'controlled'. One might decide to control the temperature that at which the fats are tasted as 19^{o}C (room temperature). This would be recorded in the report on the experiment. If someone suspected that different results might be obtained with cooler fats they could then repeat the experiment with that single change.

Temperature is a 'continuous variable'. It can take values 15^{o}C, 19^{o}C, 19.5^{o}C and so on. Some of the other variables which may be important in determining the results of this experiment do not have values which are numbers, rather the values of the variable are categories or names. These are known as 'nominal variables'. Take for example the variable method-of-presentation. This could take values such as, on-a-salty-biscuit, on-bread, on-its-own and so on.

Table 1.1 lists three variables to be controlled in the experiment, they are quantity-of-fat in grams (continuous), presentation-of-fat (nominal) and temperature-of-fat (continuous). In addition Table 1.1 gives the 'independent' and 'dependent variables'. These terms are explained in the next section.

Quantity of fat in mouth - controlled: 2 gm

Presentation of fat - controlled: spread on white bread

Temperature of fat - controlled: 19°C

Kind of fat - independent variable:

level 1 'Koma Quality Margarine'

level 2 'Country Glow Slightly Salted Butter'

Preference rating - dependent variable (score between 0 and 100)

Table 1.1 Variables in an experiment.

The independent and dependent variables

A good experiment seeks to control all the relevant variables in some way. Some, quantity, temperature and presentation here, will be held constant. Others will be manipulated i.e., controlled at two or more values. In Table 1.1 kind of fat is the variable which has been chosen as the manipulation. Tasters will get 'Koma Quality Margarine' or 'Country Glow Slightly Salted Butter'. For somewhat obscure reasons this is called the independent variable. Notice that the independent variable is said to have levels. You may find it strange to think of kind of fat as a variable having levels 'Koma Quality Margarine' and 'Country Glow Slightly Salted Butter'. The analogy is with levels of a continuous variable. For example, Table 1.2 shows how temperature might be used as an independent variable. Here the levels of the independent variable are 15 and 19 degrees Celsius. Type of fat is controlled at a single level.

Quantity of fat in mouth - controlled: 2 gm

Presentation of fat - controlled: spread on white bread

Temperature of fat - independent variable:

level 1 15°C

level 2 19°C

Kind of fat - controlled: 'Koma Quality Margarine'

Preference rating - dependent variable (score between 0 and 100)

Table 1.2 An alternative experiment.

The variable in which the result of the experiment is expressed is called the dependent variable. In this example the dependent variable is the preference rating. The distance from the mark made by the taster on the 100 mm line, measured from the 'very nasty' end of the scale, gives a score from 0 to 100. 100 is a very positive preference and 0 a very negative preference.

Experimental control

This example is applied behavioural science so the values chosen for variables should be typical of the real situation we are trying to mimic. Accordingly we chose 19 degrees Celsius as the temperature and 2 gms as the quantity. These values will be recorded in the report of the experiment so that someone else repeating it can see how the variable was controlled. It is most important that none of these variables varies with the independent variable. Such a situation is known as 'confounding'. Quantity might, for example, inadvertently have varied with fat type. Perhaps the person running the experiment liked butter much more than margarine and gave everyone who tasted butter more than everyone who tasted margarine. Were this to be the case we can no longer interpret a difference in preference rating. Type of fat is confounded with quantity of fat. It may be that people prefer the margarine or it may be that they prefer being given less fat.

To summarise so far, we have seen three kinds of variable: (i) the most common kind of variable is controlled at a single level so as not to bias the results of the experiment; (ii) one variable is manipulated to two or more levels, this is the independent variable and forms the basis of the experiment; (iii) one variable is chosen to express the outcome of the experiment, this is the dependent variable. In the process of deciding how to control all these variables the original hypothesis has been refined quite considerably. We started off with 'people prefer butter to margarine'. Table 1.1 describes an experiment to ask the much more specific question - whether people give higher preference ratings to Koma Quality Margarine or Country Glow Slightly Salted Butter when both are presented on white bread at 19ºC there being 2 gms of fat in the mouth when tasting. This additional specificity is necessary to get a reasonably definitive conclusion from an experiment. It is a strength of the experimental method and also a weakness. The specificity makes it possible to answer the question, it also forces the scientist to think clearly about the implicit assumptions being made. The weakness, the cost that has to be paid, is that the conclusions made are specific to the experimental situation. For example, the ingenuity required to control the amount of fat in the mouth of the taster may make the experiment rather atypical of real tasting situations.

Subject variables

Psychologists used to call the people who take part in their experiments 'subjects'.Now it is more common to call them "Participants", nevertheless the word remains in use as a technical term in statistics. Table 3 contains a list of some of the variables on which participants could vary and why they could affect the results of the experiment. We will refer to these as "subject variables".

Recent eating history (Have they eaten anything with a strong flavour in the past hour?)

Long term eating history (Do they normally eat margarine?)

Age (the senses, including taste, dull with age)

Health (a blocked nose will interfere with smell which is intimately associated with taste)

Attitude to food (Do they care about the taste of food?)

Table 1.3 Some subject variables

Recent eating history could conceivably be controlled at a single level by asking participants not to eat or drink anything before they did the experiment. The other variables in Table 1.3 could only be controlled by selecting participants. It would make sense to select people who were healthy but what about long term eating history and attitude to food? Even if we could assess the values of these variables satisfactorily it would not be sensible to select participants so as to control them at a single level as we want the results of the experiment to be generalisable. Ideally we would like our conclusions to apply to people in general, not just to 25 year olds who eat 150 gms of butter per week and who rate their attitude to food as 'positive'.

The problem is finessed by specifying the population we wish to generalise to and then sampling at random from it. So, we might define our target population as psychology undergraduates at the University of York. We then sample at random from this population. The definition of a random sample is that each member of the population has an equal probability of being in it. This means that, within the limitations of chance, the proportion of participants in the sample who have particular characteristics will be representative of the proportion of participants having the same characteristics in the population. The procedure automatically assures that there will be a representative range of ages, attitudes to food and long term eating history. This is a clever solution to the problem of controlling participant variables as it even controls variables we don't know about. Let us say that, after we have completed our experiment some psycho-physiologist discovers an enzyme, found in the saliva of some people but not others, that strongly affects the ability to taste fats. We do not have to repeat the experiment selecting the appropriate proportions of participants with and without this enzyme in their saliva. Because we sampled randomly the sample should already be representative.

Technically what we have done in randomly sampling from a population is to declare a new variable 'Subjects'. This variable has the levels 'John Smith', 'Fred Jones', 'Jean Brown' and so on. This variable is called a 'random effect' as the levels chosen for the experiment are sampled randomly from the complete set of possible levels in a population. This can be contrasted with the independent variable in an experiment, such as type of fat, which is called a 'fixed effect' as the levels chosen are fixed for the experiment. Were we to repeat the experiment we would obtain a new random set of participants but we would use the same brands of fat.

Random sampling presents certain practical problems which need not concern us at the moment. How these problems can be circumvented is discussed in Chapter 2 and Chapter 7.

Technical note - types of measurement

Variables may be classified in a number of ways. One of interest to statisticians concerns the kind of information given. At one end of the continuum there are nominal variables such as eye colour or sex. At the other end there are ratio variables which have numerical values e.g., weight. Table 1.4 gives four types of variable ordered by the amount of information given. At the bottom of the table, nominal variables only tells you whether two individuals are the same or different. At the top of the table a ratio variable quantifies how much they differ, one can make statements such as individual A is twice the weight of individual B. An interval variable also takes a numerical quantity but does not have a meaningful zero point so one can only make statements about intervals rather than ratios (this is a rather fine statistical point without important practical implications in this book). Ordinal variables simply order individuals, one cannot make statements about intervals or ratios.

For the purposes of this book we only need to distinguish between continuous variables (ratio and interval) and nominal variables. Strictly speaking a continuous variable can take any value within some range, e.g. 19.5638°C. A measure which can only take specific numerical values is said to be 'discontinuous', e.g., number-of-children can only take values which are whole numbers. In practice most of the measures used by psychologists are discontinuous, in the strict sense. If the number of possible values which the variable can take is not unduly restricted then they are normally treated as if they were continuous. We shall return to the issue of types of measurement in Chapter 6.

Type of variable	Example of use	Example of variable
Ratio	X is twice as big as Y	weight in gms
Interval	A - B = C - D	temperature in °F
Ordinal	X is greater than Y	ranking
Nominal	A is the same as B	marital status

Table 1.4 Types of variable

Summary

The Chapter has shown how an initial hypothesis is made more concrete by thinking about the variables which need to be controlled in order to test it experimentally. 'People prefer butter to margarine' was progressively refined by defining an independent variable (what we mean by 'margarine' and 'butter') a dependent variable (what we mean by 'distinguish'). Other variables need to be controlled if the experiment is to be unambiguously interpretable. Some of these (e.g. temperature and quantity of fat) are fixed at a single sensible value. Others, subject variables, are controlled by sampling from a specified population.

The concepts introduced are summarised in the following glossary:

Continuous variable: in theory a variable that can take any value e.g., mass in gms. In practice variables which can take only some values, such as the number of words recalled in a memory test (you can only recall a whole number of words), are treated as continuous.

Nominal variable: a variable whose values are simply names e.g., the variable eye colour with values: blue, grey, brown or the variable sex with the values: male or female.

Independent variable: the variable used to form the experimental question being asked e.g., fat tasted (butter versus margarine), temperature (15°C versus 19°C) or sex (male versus female).

Levels of an independent variable: the different experimental conditions specified by an independent variable are known as its levels. In our example the independent variable was type of fat and the levels 'Koma Quality Margarine' and 'Country Glow Slightly Salted Butter'.

Dependent variable: the variable which is the 'output' of the experiment. In our example this was a preference rating. Most commonly the dependent variable is a continuous variable such as a score. Sometimes it is a nominal variable e.g., 'succeeded' or 'failed'. Many people have difficulty remembering which is the independent variable and which is the dependent variable. The dependent variable is the one that is dependent on everything else that you do, for 'dependent variable' read 'result' or 'score'.

Participant: the person who takes part in an experiment, as distinct from the all powerful experimenter who runs it (in more old fashioned texts these people are called "subjects").

Chapter 2

Summary Statistics

Statistical concepts introduced in this Chapter:

Contingency tables, probability, mean, median, mode, range, semi-inter-quartile range, standard deviation and variance, standard error of the mean.

Summarising nominal variables - contingency tables

Large quantities of data are difficult to take in. For this reason we normally compute summary statistics. The commonest of these is the humble average. First we will consider another way of summarising data, the contingency table

Some fictitious data about the patients passing through an acute psychiatric ward for depressives are presented in Table 2.1. This gives the number of days they were there before they were discharged, their sex and the treatment they received. ECT stands for electro-convulsive therapy.

	Number of days	Sex	Treatment
Patient 1	15	M	ECT
Patient 2	3	F	Anti-depressants
Patient 3	5	F	ECT
Patient 4	4	F	Anti-depressants
Patient 5	11	M	Anti-depressants
Patient 6	4	F	Anti-depressants
Patient 7	12	M	ECT
Patient 8	3	F	Anti-depressants
Patient 9	10	F	ECT
Patient 10	4	F	Anti-depressants

Table 2.1 Abstracts from patient records (fictitious data)

There are various ways we could tabulate these data. Let us concentrate on the sex of the patients and their treatment. We could count the number of male and female patients and put the answer in a table:

Number of patients

Male	Female
3	7

There are more females than males in this sample.

Similarly we could count the number of patients receiving the two kinds of treatment:

Number of patients

ECT	Anti-depressants
4	6

Finally we can count the number of males receiving ECT the number of females receiving ECT and so on:

Number of patients

	ECT	Anti-depressants
Male	2	1
Female	2	5

This is known as a contingency table. It shows how one variable is contingent on the other. For example, the males are much more likely to be given ECT (I should stress that all the examples used in this book are fictional!)

There are more females than males so an additional table giving the percentages of each sex having the two treatments would aid interpretation.

Percentage of patients of each sex receiving a particular treatment

	ECT	Anti-depressants
Male	67	33
Female	29	71

Probability

The above results could have been expressed as proportions. The proportion of males getting ECT in this study is 0.67. Proportions like this are sometimes referred to as 'a posteriori' (after the event) probabilities, or sometimes 'empirical probabilities'. The a posteriori probability of getting ECT if you are female is lower (0.29) i.e., the proportion of females being treated with ECT is .29.

Probabilities refer to events in this case the occurrence of a patient who is female being treated with ECT. An a posteriori probability then is a number between 0 and 1 indicating how often an event has occurred. It is fairly rare to use a posteriori probabilities in this way as it is more straightforward to think of the result as a proportion.

'A priori', or theoretical probabilities are much more important as we will see in later Chapters. This is again a number between 0 and 1, but this time it expresses a prediction as to how likely some event is to happen. For example, if one is willing to assume that:

(a) it is equally likely a coin will come up heads or tails

(b) heads and tails are the only events possible, (it never lands on its edge or falls down a crack in the floor for example) and

(c) it can't come up heads and tails at the same time, i.e., these are mutually exclusive events, then one can compute the theoretical or a priori probability of one of these events occurring as 0.5. These are reasonable assumptions and empirical investigation with coins will prove it to be a good prediction in that the proportion of heads in a long sequence of coin tosses will be approximately 0.5. In Chapter 5, and the Chapters which follow it, we will compute a priori probabilities where the event to be predicted is getting a particular result from an experiment. As with the coin example we will have to make assumptions though they are slightly more complex than (a) to (c) above.

The probability of something not happening

Readers who have learned how to manipulate probabilities in maths courses will know that if two events are mutually exclusive then their probabilities can be added to get the probability of one *or* the other happening. Let us say that we have computed the a priori probability that the next card in the deck is a king to be .10 and that the a priori probability it is a queen is .15. These events are mutually exclusive, the card cannot be a queen and a king at the same time. This means that we can add the probabilities. The a priori probability it is a king *or* a queen is .25 (.10 + .15).

You will only need to use this rule in one very specific situation, that is to work out the probability of something not happening. If we know that the probability of the next card being a queen is .15 then the probability it is not a queen is .85. This follows from the above rule because the card must either be a queen or not a queen. One of these events must happen, the probability of being 'a queen or not a queen' is 1.00 i.e., certainty. In addition, being a queen and not being a queen are mutually exclusive events. In summary

Probability the card is a queen = .15

Probability the card is a queen + the probability the card is not a queen = 1.0 (must be true)

Therefore

Probability the card is not a queen = 1.00 - the probability it is a queen = 1.00 - .15 = .85

The same logic can be applied to a posteriori probabilities. Knowing that the probability of getting ECT if you are male is .67 and that everyone either got ECT or anti-depressants (but not both), it is possible to deduce that the probability of getting anti-depressants if you are male is

1 - 0.67 = 0.33

Measures of central tendency: mean

Now to return to the summary statistic you are all probably already familiar with, that is the idea of an average. An average summarises a set of numbers as some sort of 'central tendency'. Table 2.1 includes the number of days spent in the ward. Even though there are only ten data points it is difficult to see whether say patients receiving ECT tend to spend more time in hospital than those getting anti-depressants. With larger, more realistic, samples this is even more of a problem. What is needed is some statistic that summarises the 'central tendency' of a set of scores. The most commonly used in the *mean*. This is computed by summing the scores and then dividing by the number of scores summed.

Mean stay of patients receiving ECT (15+5+12+10)/4 = 10.5 days

Mean stay of patients receiving anti-depressants (3+4+11+4+3+4)/6 = 4.8 days

You may know the mean as the average of a set of numbers. The mean is a measure of central tendency. It gives a single value which is typical or central as a way of summarising a set of values. As we shall see there are other measures of central tendency that are sometimes also described as averages and so we shall always use the more precise statistical term, mean.

Measures of central tendency: median

Another measure of central tendency is the median. The *median* is a value having as many scores above it as it has below it. The median of the set of scores

41 53 37 31 64

is obtained by first putting the scores in order

31 37 **41** 53 64

41 is the central value in this order so the median of this set of scores is 41 (no measure of central tendency is going to be very meaningful with samples as small as this but by using a small data set the procedures should be easier to follow). If there is an even number of scores in the set then the value with an equal number of scores either side of it will be somewhere in between the middle two scores. By convention it is set as the mean of these two values i.e., half way between the two. So the median of the scores

34 52 46 75 60 61

which sorted are

34 46 **52 60** 61 75

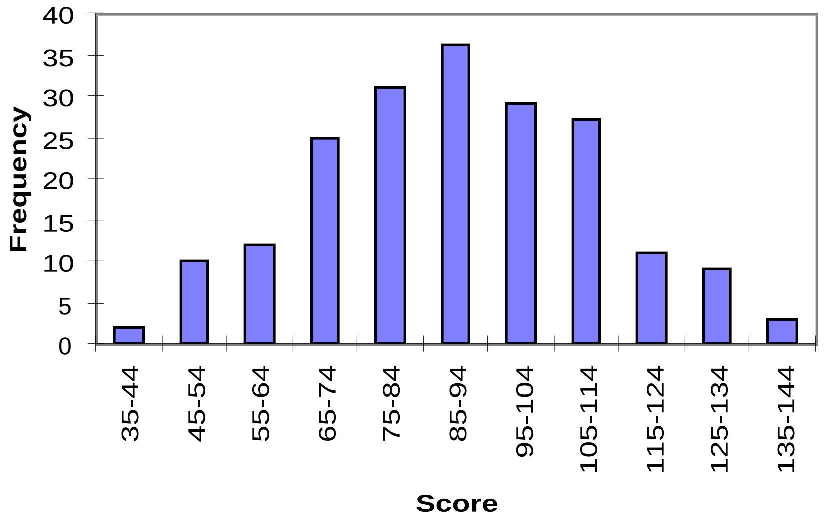
is (52+60)/2 = 56

If some of the values are the same (ties in the ordering) we proceed just as above. The median stay of patients receiving anti-depressants (3 4 11 4 3 4) or sorted as (3 3 **4 4** 4 11) is (4+4)/2=4

Measures of central tendency: mode

The mode is defined with respect to a frequency histogram such as Figure 2.1. It is only a meaningful measure of central tendency with large samples. The *mode* is the value, or range of values, which most frequently occurs.

Figure 2.1 Frequency histogram for a population of scores - modal range 85-94



You will have plotted frequency histograms in school maths courses. Figure 2.1 is such a histogram representing the data from 160 participants. The horizontal axis is divided into ranges of scores. You might like to think of these as buckets. Any score between 35 and 44 is put into the first bucket, anything between 45 and 54 is put in the second, and so on. The vertical axis gives the number of scores falling into each bucket. So there is 1 score between 35 and 44 and 10 between 45 and 54. The fullest bucket, that is the range of scores within which scores most frequently occur, is the mode. The modal value is usually taken as the centre of this range, in this case 90.

Sometimes there is more than one peak in the histogram. In Figure 2.2 this happens because there are two few data points to make the use of a mode sensible. Most of the scores occur with a frequency of only 1 or 2. Increasing the size of the buckets would not help because one would end up with two few of them to be useful. In Figure 2.3 there are two modes, in the jargon a 'bimodal distribution'. This may be the result of mixing two rather different populations of scores. The mode would not be a representative measure of central tendency for the data displayed in Figure 2.2 or Figure 2.3.

Figure 2.2 Frequency histogram - these data are not sensibly summarised by a mode as the size of the population is too small.

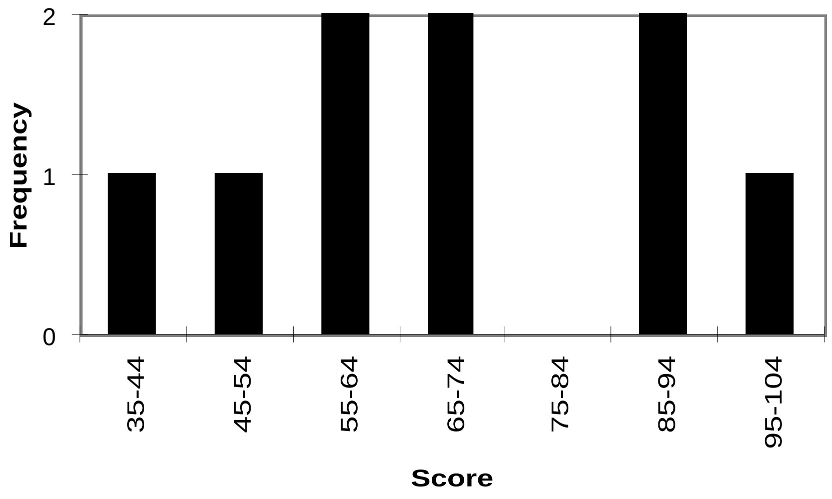
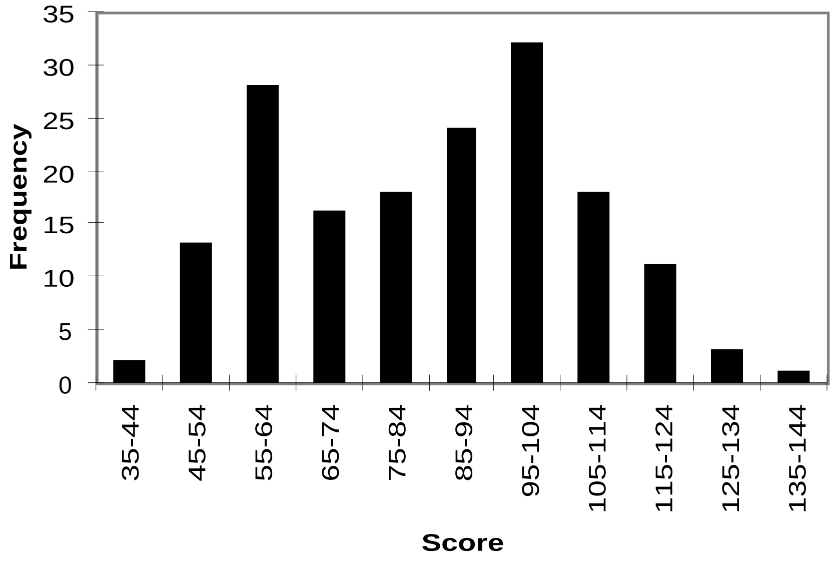


Figure 2.3 Bimodal frequency distribution - see text for explanation



Relationship between mean, median and mode

With a perfectly symmetrical distribution the mean will have an equal number of scores either side of it so it will correspond to the median. In many natural distributions the central values are the most common and the mean is the mode also. This state of simplicity is approximated in Figure 2.1. The mean for the data upon which it is based is 83.45, the median 84.0 and the mode 90. Figure 2.4 is not symmetrical. There are more values at the lower end than at the higher end. This is often the case with small measurements of time. It is much harder to reduce one's reaction time from say .9 seconds to .5 than from 2.9 to 2.5. It is impossible to reduce it below 0 seconds! With asymmetrical distributions like this the mean median and mode do not correspond. This distribution can be described as 'positively skewed'. It is as if a symmetrical distribution has been stretched at higher or 'positive' end of the scale. This can be seen in the relationship between mean and median. The median is .50 and the mean .77. The median being less than the mean is a sign of a positively skewed distribution. Were the median to be greater than the mean that would indicate a negatively skewed distribution. Which measure of central tendency is most representative of a skewed distribution is debatable. Most people would probably take the median.

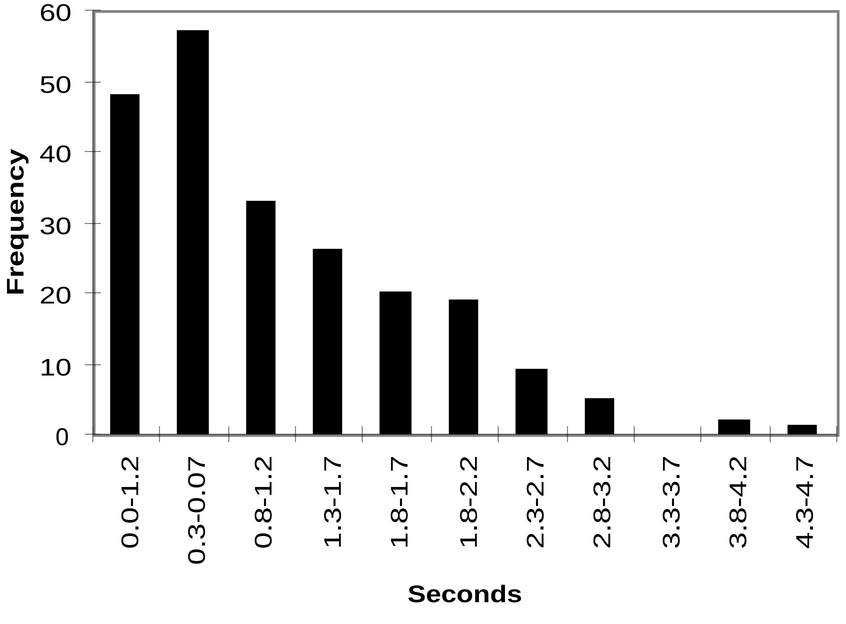
In small samples the mean can differ from the median because of outliers. Outliers are values which are atypical because they are very much larger, or smaller, than all the others. Consider the following two sets of data.

(a) 3 5 7 5 4 5 6 7 7 4 29 Mean 7.45, Median 5.00

(b) 7 6 9 8 7 6 7 5 7 9 7 Mean 7.09, Median 7.00

Set (a) contains an outlier (29). All the other values are in the range 3 to 7. 29 is clearly aberrant as it is 4 times that range from the next highest value (a rule of thumb for identifying outliers is given in the section on box-plots below). Outliers distort the mean much more than they do the median. For this reason the median is generally considered to be a more representative measure of central tendency than the mean when there are outliers.

Figure 2.4 A positively skewed frequency distribution



It is unusual for means and medians to give different conclusions (according to the means (a)>(b), according to the medians (b)>(a)). Where there are large differences between the mean and the median the validity of the data should be questioned and it may be possible to remove outliers, without biasing the results of the experiment, by applying some objective rule. The issue of outliers and skewed distributions will be reexamined in Chapter 6.

Measures of variation

The following two sets of data have means which are almost the same yet they differ in an important respect, the variation in scores.

(c) 42 41 50 59 46 63 74 35 75 42 46 Mean 52.09

(d) 31 34 5 99 56 87 63 57 40 25 61 Mean 50.73

Variation, sometimes known as dispersion, can be summarised in several ways, the simplest is the range. Data set (c) has a minimum value of 35 and a maximum of 75 so the *range* (i.e., maximum - minimum) is 40. Sample (d) has a minimum of 5 and a maximum of 99 so its range is larger at 94.

Sometimes the range may not be a very representative statistic because it is based on the two most extreme values and extreme values may be flukes in some way. The inter-quartile range is a related but slightly more sophisticated measure of variation devised to get round this problem. To explain how this works we need to explain the concept of a 'quartile'.

The median is defined as the value having half the scores below it. On a similar basis the first quartile is defined as value having one quarter of the scores below it. The third quartile is the value having three quarters of the scores below it (on this basis the median could be thought of as the second quartile). By taking the difference between the first and third quartiles we get a measure of variation which does not depend on the possibly aberrant minimum and maximum scores. Thus, the inter-quartile range is the difference between the first and the third quartile. To find the inter-quartile range for data set (c) proceed as follows:

1. Find the first and the third quartile. To do it by hand rank the data and pick the scores with one and three quarters of the scores below them respectively

35 41 **42** 42 46 46 50 59 **63** 74 75

2. Compute the difference between the two quartiles. This is the inter-quartile range

inter-quartile range = 63 - 42 = 21

It is easy to compute the first and third quartiles in this example because they are represented by actual scores (42 and 63). This will not always be the case. Just as the median of an even number of scores will always fall between two scores there will be occasions where quartiles fall between two scores. In such cases Minitab extrapolates. For example, if the first quartile should really be at rank 2.75 then it will work out the value three quarters of the way between the score with rank 2 and the score with rank 3. Those who are interested may consult the Minitab reference manual for further details of how this is done.

Box-plots

The inter-quartile range is the basis of a box-plot. Figure 2.5 contains four box-plots, one for each of the data sets (a) to (d). The boundaries of the box are the two quartiles so it contains the central 50% of the data. The cross marks the median value. The lines going out from the box, known as whiskers, stretch to the minimum and maximum values. Where there are extreme outliers (conventionally 1.5 times the inter-quartile range above or below the relevant quartile) they are represented by individual points (see for example the '0' in the box plot for data set (a) in Figure 2.5).

Box plots summarise central tendency and variation and so are useful for comparing data. Consider first the box plots for data sets (c) and (d). The minimum and maximum of these data sets are not outliers by the above definition and so the whiskers contain all the data. The similarity of central tendency and the difference in variation between data sets (c) and (d) is apparent in Figure 2.5. Now consider the box plots for the data sets (a) and (b). The outlier in data set (a), 29, is plotted as a separate point ('0'). Otherwise the two data sets can be seen to be very similar in central tendency and variation.

The standard deviation

An alternative way of thinking about variation is in terms of the average deviation from the mean. The second column in Table 2.2 gives the deviation of each value in data set (c) from the mean of 52.09. Similarly the last column gives the deviation of each value in data set (d) from its mean of 50.73. The absolute size of these deviations is much larger in the case of data set (d).

```
          ------
(a)    ---I+   I                                           O
          ------

              ---
(b)        ---I+I---
              ---
     ------+---------+---------+---------+---------+---------+
         5.0      10.0      15.0      20.0      25.0      30.0


                          -----------
(c)                    ---I +       I-------
                          -----------

                     ----------------
(d)     -------------I           +  I-------------------
                     ----------------
     +---------+---------+---------+---------+---------+-----
     0        20        40        60        80       100
```

Figure 2.5 Box plots of data sets (a) to (d).

If either of these sets of deviations from the mean are added up the negative deviations will balance the positive deviations and the net total will be zero. We need to average the absolute size of the deviations i.e., ignoring whether they are positive or negative. For reasons too complicated to go into here it turns out that the mathematically most appropriate way of doing this is as follows:

(i) square the deviations (this gets rid of all the negative values e.g., -10.09 x -10.09 = 101.81)

(ii) sum the squared deviations

(iii) divide by N-1, where N is the number of scores in the data set

(iv) square root the result

The statistic thus computed is called the *standard deviation*. Its square, or what you get if you stop after step (iii) is called the *variance*. The sum of the squared deviations for data set (c) is 1868.9 and for (d) it is 7506.2 so the standard deviation is

$$\sqrt{1868.9/10} = \sqrt{186.89} = 13.67$$

in the case of (c) and

$$\sqrt{7506.2/10} = \sqrt{750.62} = 27.40$$

in the case of (d).

This computation may seem complex but you will always have Excel to do it for you. It may also seem very devious (statistical pun!) but the standard deviation is mathematically convenient and for that reason it is an important summary statistic. In general it will be approximately half the inter-quartile range and so to make the latter statistic imitate its more important cousin it is usual to quote the semi-inter-quartile range rather than the inter-quartile range calculated above.

(c)	Deviation	(d)	Deviation
42	-10.09	31	-19.73
41	-11.09	34	-16.73
50	-2.09	5	-45.73
59	6.91	99	48.27
46	-6.09	56	5.27
63	10.91	87	36.27
74	21.91	63	12.27
35	-17.09	57	6.27
75	22.91	40	-10.73
42	-10.09	25	-25.73
46	-6.09	61	10.27

Table 2.2 Deviations from the mean, data sets (c) and (d)

You may be puzzled why we divide by N-1 instead of N when computing the standard deviation. The explanation for this depends on the statistical concept of a sample to be described in the next section.

Populations and samples

Imagine it was required to measure the mean height of male students at some large university of college. Measuring all of them would be very time consuming and is beyond the resources available. The alternative is to choose at random say 200 and to take the mean of this sample as an estimate of the mean of the whole population of students. The accuracy of this estimate will depend on the size of the sample. Clearly a sample of 200 is likely to give a better estimate than a sample of say 10. It is also important that the sample is truly random and not biased in some way. For example, selecting the sample solely from the physical education department, where physique is an entrance requirement, may result in a poor estimate.

The mean of a sample is usually signified with the symbol $\bar{X}$

(capital X with a bar over it, which is said 'x bar'). The standard deviation of a sample is usually signified by s (not capital) so the variance is written as s^2. For example, for set (c) one might write

$\bar{X}$ = 52.09, s = 13.67, s^2 = 186.89.

For somewhat obscure mathematical reasons it turns out that computing the standard deviation by dividing by N-1 rather than N gives an unbiased estimate of the population standard deviation. Dividing by the apparently more reasonable N gives, on average, an under estimate. Some calculators give you the choice of computing the standard deviation of a set of data using N or N-1. Only use the former if you have measured the complete population (something you are very unlikely to do).

To summarise, the procedure for estimating population statistics using a sample is as follows:

(i) define the population (e.g., male students at a particular university of college);

(ii) sample at random from this population (at random means that each individual in the population has an equal probability of being included in the sample);

(iii) compute summary statistics for the sample, these are taken as estimates of the population statistics.

How well a sample statistic estimates the 'true' population statistic depends on the size of the sample. The standard error of the mean, explained below, is a way of quantifying how accurate the sample mean is as an estimate of the population mean.

The standard error of the mean

Let us say that instead of taking a sample of 200 students we were only able to sample ten. How accurate would the mean of this sample be as an estimate of the population mean compared with the 200 student sample? The problem is that it is always going to be impractical to measure the true population mean. However, we could measure the variation in the means derived from a number of ten- and 200-student samples. In this (hypothetical) evaluation of sample size we might take say 20 10-student samples, compute a mean for each of these samples and then compute the standard deviation of these means. Doing the same thing for say 20 200-student samples would allow us to compare the variation in sample means directly. This is a hypothetical exercise as it is unlikely to be practical to take 20 200-student samples. However, statisticians have shown that, if one makes certain reasonable assumptions about the data, the standard deviation hypothetically computed above can itself be estimated from the standard deviation of the sample. That is, the standard deviation of the means of a number of samples of size N is approximately $s/\sqrt{N}$ where s is the standard deviation of one of the samples.

This mathematical finding gives rise to a measure of accuracy known as the standard error of the mean

$$\text{standard error of the mean} = \frac{\text{standard deviation of the sample}}{\sqrt{N}}$$

For example, say the mean height of our sample of 200 is 1.65 metres and the standard deviation of the sample is .18 metres. Then the standard error of the mean is

$$\frac{.18}{\sqrt{200}} = .013$$

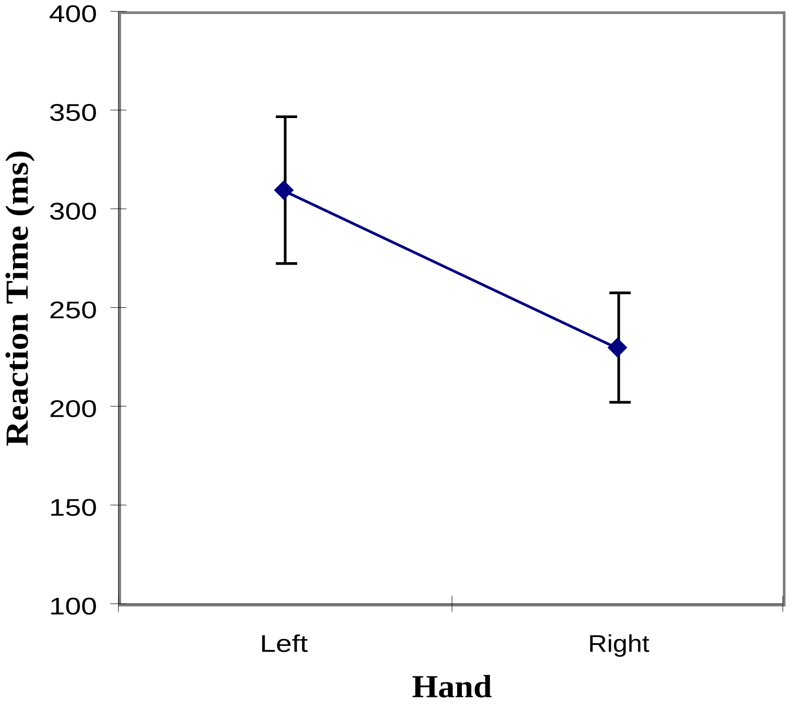
If the sample had been of only 10 individuals the standard error of the mean would have been .057, that is over 4 times larger. The standard error of the mean is used to indicate precision of measurement. Unlike the standard deviation, which is independent of N, it gets smaller the more data points go into the mean.

Figure 2.6 has error bars of one standard error of the mean plotted above and below the mean value. This is a common way of showing how good the sample mean is as an estimate of some population mean. In this case reaction time is repeatedly measured for one individual using either the left or right hand. In such a case the population is the total set of responses the individual could have produced.

Practical considerations when sampling

The procedure for estimating population statistics described above is somewhat idealised because it is extremely difficult to get a truly random sample from an arbitrarily defined population. For example, psychologists depend on volunteers who are willing to take part in experiments. Although they would like to generalise their conclusions to humanity in toto they often sample from a population which should properly be defined as 'psychology students at the University of X who volunteer for experiments'. This is not usually a problem. Statistical arguments permit them to generalise from the results obtained from the sample to the limited population it is drawn from and then arguments based on psychological knowledge of how these effects vary from individual to individual allow them to generalise from this population to a more general one.

Figure 2.6 Graph of mean reaction ti when reponses are made with the left right hand - the vertical "error bars" plus and minus one standard error of mean



Take for example an experiment on colour perception. An experimenter compares the ease with which text can be read when it is printed in blue as opposed to black ink. A consistent difference in reading speed is observed in all twenty of a sample of students. The standard error of this mean difference is small and so it is concluded that the mean difference observed in the sample is a good estimate of the population mean difference. The experimenter knows of no experiments which indicate that members of this population (undergraduates at his university who volunteer for experiments) differ in their colour perception from the population at large and so generalises the conclusion appropriately.

The problem faced by opinion pollsters cannot be solved in this way. People who readily answer questions such as who they will vote for in the next election probably do have systematically different opinions to those who are unwilling to answer such questions. However there is a solution and again it involves the use of arguments outside of statistics. Pollsters know what kind of subject variables affect people's intention to vote. They are the region they live in, their socio-economic status and so on. They also know what proportion of the population of the country falls into each possible combination of the levels of each of these variables. They can then adjust their sampling techniques to make sure that they get a representative number of each type of person. This is no longer random sampling but as long as they really do know all the most important factors having an effect on voting intentions it will have the same effect.

Summary

1. It is possible to summarise the characteristics of some set of data by computing certain statistics.

2. Nominal variables are summarised by constructing *contingency tables* giving the frequency, proportion or percentage of participants falling into different classes or combinations of classes (e.g., males receiving ECT).

3. An *a priori or theoretical probability* expresses how likely some event is to occur. A probability is represented by a number between 0 (will never occur) and 1 (will always occur).

4. The *mean* (written $\bar{X}$) and *median* are measures of central tendency. The mean is computed by summing the scores and dividing by N (the number of scores). The median is the number having an equal number of scores above and below it. The median may be more representative when there are outliers (atypically high or low scores).

5. The *mode* is only meaningful with large data sets where a frequency histogram is computed. The mode is the most frequently occurring score or range of scores in the histogram.

6. The *variance,* s^2, is a measure of variation or dispersion. It is computed by summing the squared deviations of each score from the mean and then dividing by N-1.

7. The *standard deviation*, s, is the square root of the variance.

8. Other measures of dispersion are: the range, the difference between the largest and the smallest score; the inter-quartile range, the difference between the first and third quartile (these are the scores with one and three quarters of the scores below them respectively); and the semi-inter-quartile range which is half the inter-quartile range.

9. It is often convenient to assume that the set of scores obtained is a sample from some larger population. The sample mean and standard deviation are seen as estimates of some theoretical 'true' population mean and standard deviation.

10. The standard error of the mean is an estimate of the standard deviation of the sample mean. That is, although we have only a single sample of size N and standard deviation s, were we to take numerous samples from the same population and compute their means, the standard deviation of those means would be $s/\sqrt{N}$. This standard error of the mean is often quoted as a measure of how good the sample mean is as an estimate of the theoretical true population mean, i.e., how precise a measurement it is.

Chapter 3

Relationships between variables

Statistical concepts introduced in this Chapter:

Graphs, tables and equations, the linear function, scattergram, cumulative frequency plot, percentile points.

Much of the material in the next four sections may be familiar to the reader. Nevertheless, it should be read as preparation for Chapter 4.

Tables and graphs

Tables and graphs can be used to supply essentially the same information. They make it possible to 'map' between one variable and another. Table 3.1 and Figure 3.1 both describe the relationship between search time and the number of stimuli that have to be searched. These fictitious data might have come from an experiment where participants have to find a single letter 't' in an array of other letters. The density of the letters is manipulated so that there are between 20 and 200 letters to search through on each page. Eight participants each search 20 pages of each density and the mean time taken to find the 't', averaging across participants and trials, is computed.

Number of stimuli	Time (seconds)
20	1.92
40	2.54
60	3.16
80	3.78
100	4.40
120	5.02
140	5.64
160	6.26
180	6.88
200	7.50

Table 3.1 The relationship between search time and the number of stimuli to be searched.

To use the table to map between number of stimuli and search time simply find the appropriate row. For example, to find out the search time when there are 100 stimuli the fifth row is consulted and the answer found to be 4.40 seconds. To map in the opposite direction, say to find out how many stimuli will take about 5 seconds to search, the same procedure is followed. Here the sixth row is consulted and the answer is just under 120.

Figure 3.1 Search time plotted against number of stimuli from Table 3.1

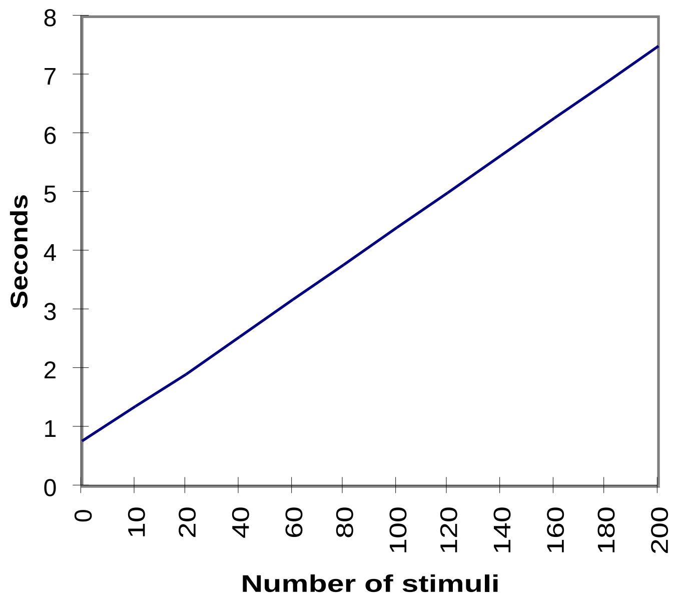


Figure 3.1 does the same job. To find the time taken when there are 100 stimuli, using the graph, first find the value on the relevant axis (in this case the horizontal one) and then read off the corresponding value on the other axis (see dotted lines on Figure 3.1. The same answer of 4.40 is obtained.

As they present equivalent information tables can be transformed into graphs and graphs into tables. To transform a graph to a table list values of one variable as the first column of the table and then read off the corresponding values for the second column from the graph. To transform a table to a graph plot a point for each pair of values in the table.

Tables, graphs and formulae

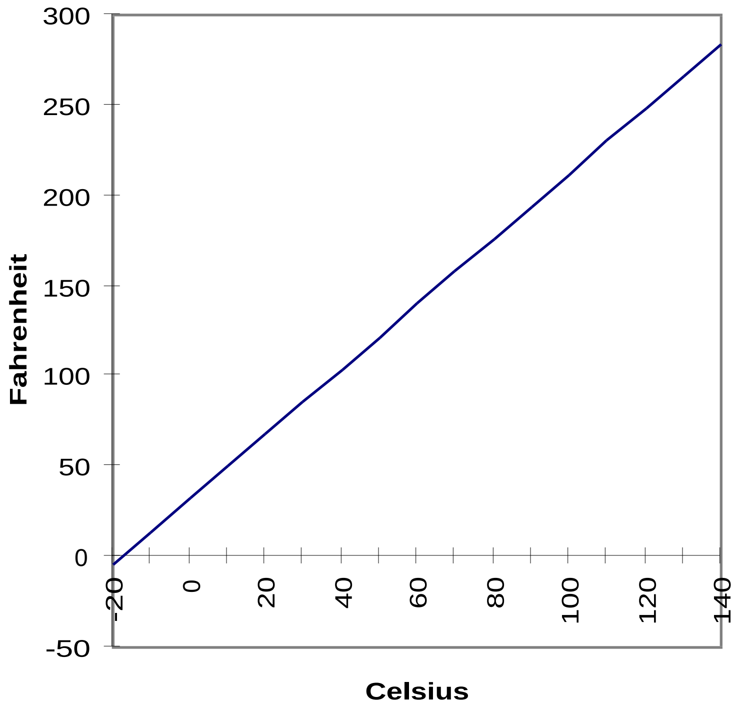
Some relationships commonly displayed as graphs or tables can also be described by a formula. Take for example the relationship between the two temperature scales Celsius and Fahrenheit. It is possible to map between these two scales using Figure 3.2 or Table 3.2, or alternatively the formula

$^{o}F = 32 + 1.8\ ^{o}C$

Degrees Celsius (^{0}C)	Degrees Fahrenheit (^{0}F)
-20	-4
-10	14
0	32
10	50
20	68
30	86
40	104
50	122
60	140
70	158
80	176
90	194
100	212
110	230
120	248
130	266
140	284

Table 3.2 Table for converting degrees Celsius to degrees Fahrenheit

Figure 3.2 Plot for converting Fahrenheit to Celsius



Let us say that it is required to find what 40 degrees Celsius is in degrees Fahrenheit. Table 3.2, row 7 shows that the answer is 104. Figure 3.2 gives the same answer (see dotted lines). To get this value from the formula the given value (40^{o}C) is 'substituted into' the formula i.e.,

$^oF = 32 + 1.8\ ^oC$

$^oF= 32 + 1.8 \times 40$

$^oF = 32 + 72$

$^oF = 104$

The equation is written for converting Celsius to Fahrenheit but it can be used in the opposite direction as well e.g.,

$104 = 32 + 1.8\ ^oC$

$104 - 32 = 1.8\ ^oC$

$72 / 1.8 = {}^oC$

$40 = {}^oC$

As one can translate tables into graphs and graphs into tables one can also translate formulae into tables or graphs. A table can be constructed from a formula by putting some values into one column of the table and then working out the corresponding values to go into the other column using the formula. The table can then be transformed into a graph if necessary.

Linear functions

The formula for converting Celsius to Fahrenheit is of a particularly common type known as a linear function. This is because when two variables related by a linear function are plotted as a graph all the points fit on a straight line. Linear functions have the form

$y = c + mx$

y is conventionally the variable one wishes to compute or estimate from some given value of x. In the Celsius to Fahrenheit example the given variable is degrees Celsius and the variable to be computed degrees Fahrenheit. So for

$^oF = 32 + 1.8\ ^oC$

y is ^oF

x is ^oC

m is 1.8 and

c is 32

(When plotting a graph it is conventional, in the social sciences, to plot the given variable on the horizontal axis.)

There was a straight line relationship between number of stimuli to be searched and mean search time evident in Figure 3.1. The formula describing this line is

seconds = 0.8 + .031 number_of_stimuli

The graph was transformed into a formula as follows. In terms of the formula $y = c + mx$,

(i) y is the time in seconds

(ii) x is the number of stimuli

(iii) m is the slope of the line i.e., how much it goes up for each unit along the horizontal axis (seconds per stimulus added to the page). It is determined by taking two arbitrary points on the line and dividing the change in seconds by the change in number of stimuli as one moves from one to the other. To minimise the error arising from reading the graph widely separate points are normally taken (see Figure 3.3). When there are 40 stimuli it takes 2.5 seconds, when there are 180, 6.9 seconds. So when the number of stimuli is increased by 140 the time is increased by 4.4

$$m = \frac{6.9 - 2.5}{180 - 40} = \frac{4.4}{140} = .031$$

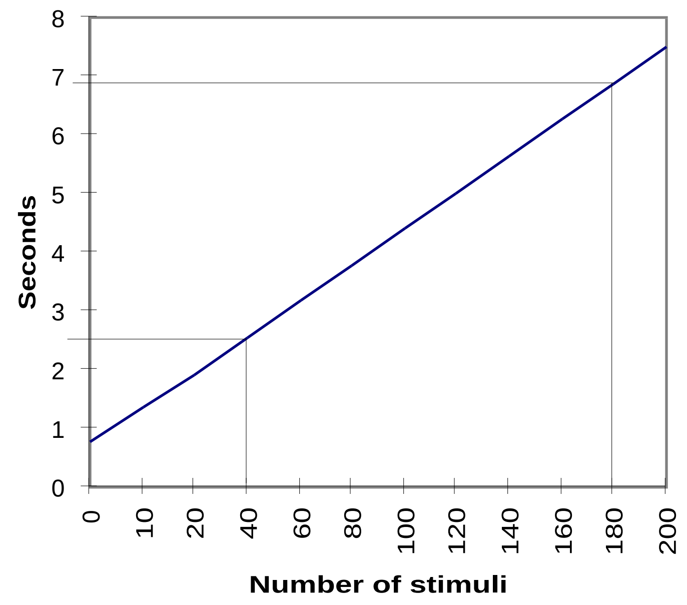


Figure 3.3 Deriving the linear function from a graph (see text for explanation)

(iv) c is called the intercept because it can be read from the graph as the y value where the line crosses the vertical axis. Every point on the vertical axis has an x value of zero (0 stimuli) so the intercept is the y value (seconds) when x is zero. In this case it is 0.8.

(v) putting the above terms into an equation

y = c + mx

becomes

seconds = 0.8 + .031 number_of_stimuli

The slope and the intercept are psychologically meaningful in this example. The slope can be interpreted as the mean time it takes to scan each stimulus. The intercept can be interpreted as the time it takes to do all the things you need to which are independent of the number of stimuli present e.g., starting to scan and making the response indicating that you have found the 't'.

Figure 3.4 has a negative slope, as one of the variables goes up the other goes down. The less time people have to solve an arithmetic problem the more errors they make (let us say that the participants have a limited amount of time, varying from 1.5 to 4.5 seconds per problem, to solve 100 arithmetic problems).

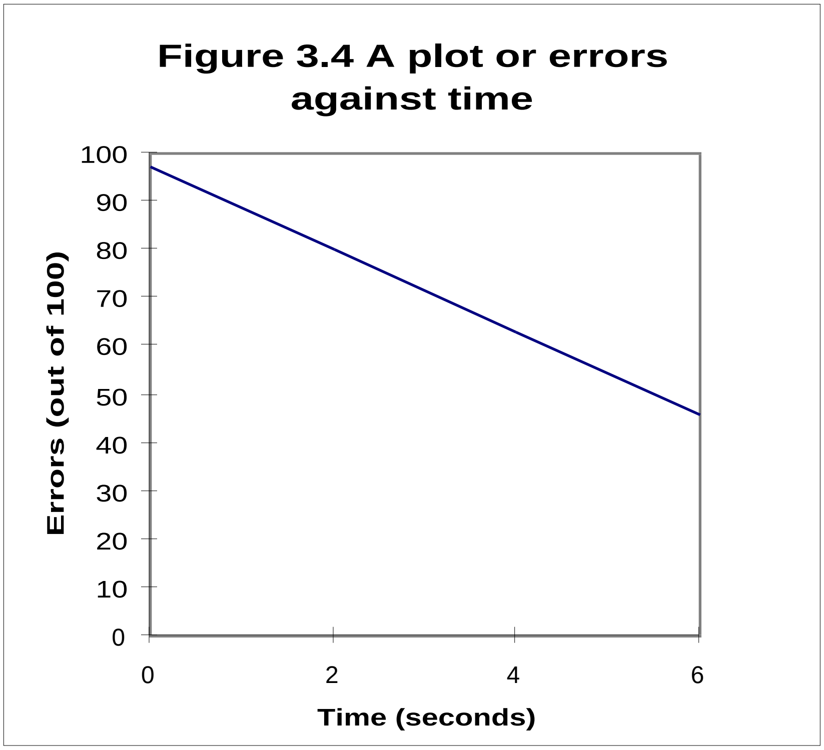


Figure 3.4 A plot or errors against time

Figure 3.4 is transformed into a formula in exactly the same way as Figure 3.3 was. Two arbitrary points are chosen to determine the slope (see dotted lines). As time changes from 2.0 to 4.0 the number of errors changes from 80.7 to 64.1, so

$$m = \frac{64 - 81}{4.0 - 2.0} = \frac{-17}{2.0} = - 8.5$$

The intercept is 97 so the equation for this line is

errors = 97 - 8.5 seconds

The scattergram

The next Chapter considers a rather different kind of plot, the scatter diagram or scattergram. Figure 3.5 contains three plots of this kind. Each point on these graphs represents an individual. For example, Figure 3.5(a) plots one point for each of 12 children according to their score on a measure of arithmetic ability (Scale A) and their mark on an exam at the end of an arithmetic course. As the designers of Scale A hoped children who have high Scale A scores also tend to do well on the exam and children with low Scale A scores tend to do less well on the exam. There are exceptions to this trend however. The point marked P is a child who got a relatively low Scale A score but nonetheless did well on the exam. The point marked Q is a child who got a high Scale A score and did relatively poorly on the exam.

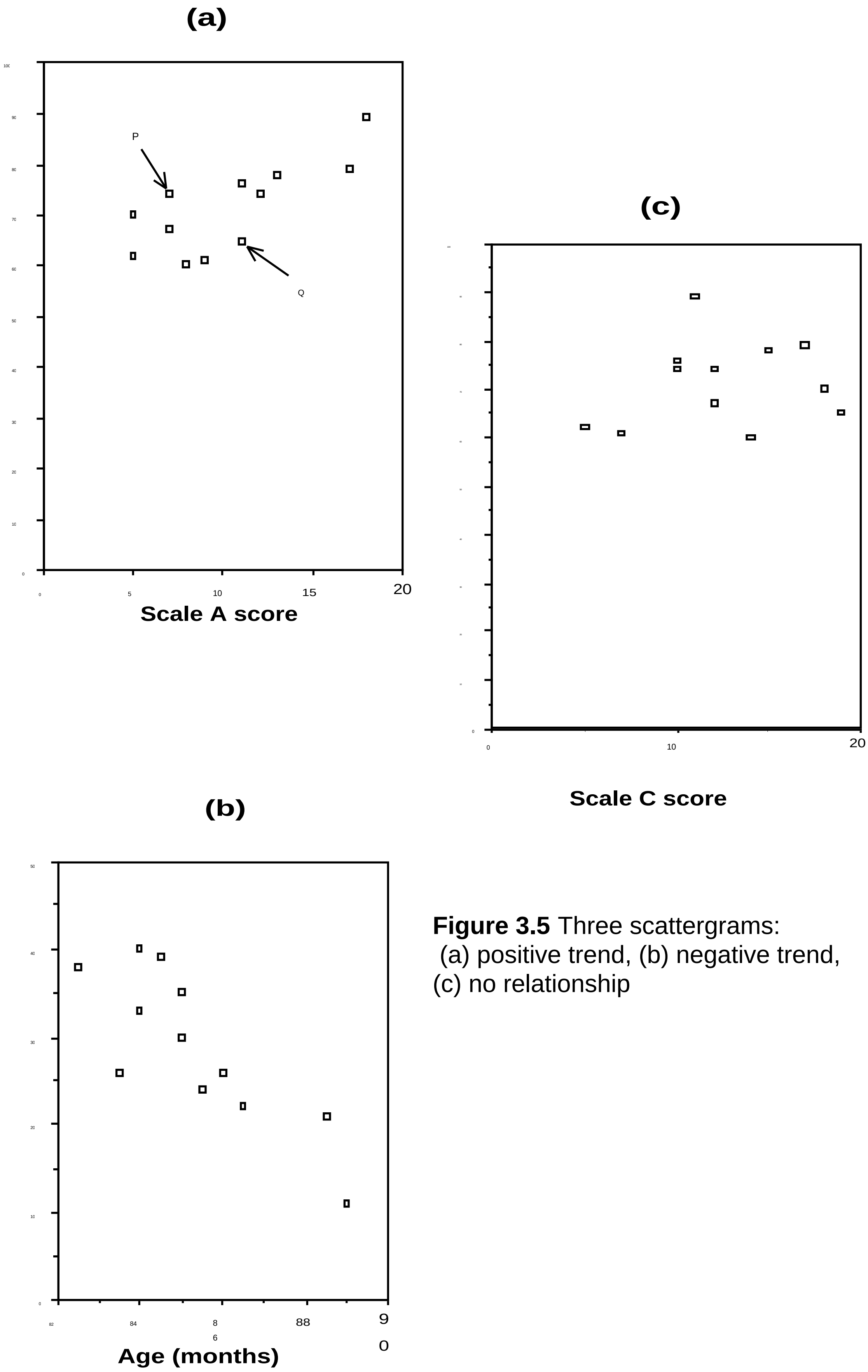


Figure 3.5 Three scattergrams:
(a) positive trend, (b) negative trend,
(c) no relationship

A scattergram differs from the other graphs considered in this Chapter in that there is a 'cloud' of points which cannot be joined up to form a line. In Figure 3.5(a) the cloud of points shows a positive trend. In the next Chapter we shall see how this trend can be described by thinking of it as a positively sloping linear relationship between Scale A and exam score plus some deviation from this idealised function.

Figure 3.5(b) is a scattergram showing a negative relationship. This plots the time it takes to solve a jigsaw puzzle against the age of the child in months. Here the trend is for the older children to solve the problem in less time. This can be thought of as an underlying linear trend, with a negative slope, plus some deviation from the trend.

Sometimes a scattergram will provide no evidence of an underlying trend. Figure 3.5(c) plots another measure of scholastic ability, Scale C, against exam score. Here there is no sign of a positive or a negative trend. There does not seem to be any relationship between Scale C and exam score.

'S' shaped curves

This far all the functions considered have been linear. Even in the case of a scattergram, where no simple function can be plotted through all the points, the trend in the data can be described as an *approximation* to a straight line (this is the topic of Chapter 4). Only one non-linear function will be considered in this book, that is the 'S' shaped curve derived from frequency distributions.

When a test of intelligence or whatever is constructed it is given to a large random sample from the population it will eventually used on. The score of each person in this sample is obtained. When someone later uses the test to assess some individual the score obtained can then be related to the scores obtained by the people in initial large sample to see whether it is good or bad. The procedure of testing a large sample is known as *standardising* a test. The sample is known as the *standardisation sample*.

A psychometric test was given to a standardisation sample of 1000 individuals. The number of individuals with scores between 8 and 11, 12 and 15 and so on are given in the second column of Table 3.3 'Frequency'. This column shows that the commonest scores are the central ones, 32 to 35. Those at the top and bottom ends of the scale are much less frequent. This would be apparent were a frequency histogram to be plotted as in Figure 2.1 in Chapter 2. Instead Figure 3.6 plots the cumulative frequency. This is obtained by summing the frequencies starting at the bottom of the table (see column 3 of Table 3.3). Thus 2 is added to 3 to get 5, then 2, 3 and 8 are added to get 13 and so on.

The cumulative frequency in column 3 of Table 3.3 shows how many people got some score *or less.* Two individuals get scores of 11 or less, five individuals get scores of 15 or less, 13 of 19 or less an so on up to 55 which is the highest score recorded so 1000 individuals (all of the standardisation sample) got 55 or less. The reason for doing this is that, when transformed to percentages, these cumulative frequencies give a score which can be compared across tests.

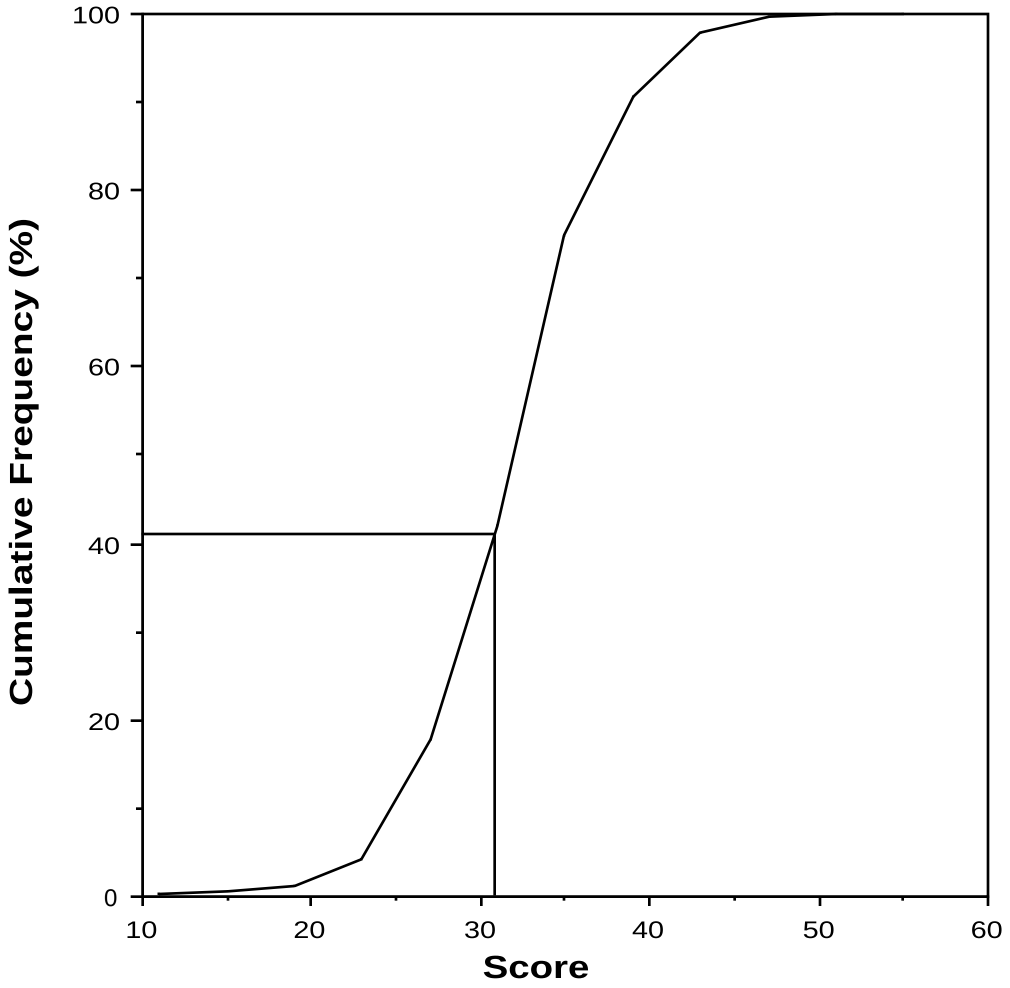


Figure 3.6 Cumulative frequency plot, data from Table 3.3

Class interval	Frequency	Cumulative Frequency	Cumulative Percentage Frequency
52-55	1	1000	100.0
48-51	1	999	99.9
44-47	20	998	99.8
40-43	73	978	97.8
36-39	156	905	90.5
32-35	328	749	74.9
28-31	244	421	42.1
24-27	136	177	17.7
20-23	28	41	4.1
16-19	8	13	1.3
12-15	3	5	0.5
8-11	2	2	0.2

Table 3.3 Frequency data used in Figure 3.6

Rather than reporting a particular individual's score as 42, say, we can report that 97% of the standardisation sample got scores lower or equal to his. Without knowing anything about the test we can tell that this is a good score. If the score was reported as being one which only 27%, say, of the standardisation sample equalled or were lower than, then we would know that it was a moderate to poor score. This is known as reporting 'percentile points'. 42 is a score at the 97th percentile point. The median is the 50th percentile point and the first quartile the 25th percentile point. The 'S' shaped curve in Figure 3.6 can be used to extrapolate between values in the table. The dotted lines show how a raw score of 31 is transformed to percentile points.

Summary

1. Tables, graphs and formulae can all be used to describe functions, that is the relationship between two variables.

2. The simplest functions considered in this Chapter are linear functions. They take the general form

y = c + mx

c is the intercept and m the slope of the line.

3. A scattergram plots relationships between variables when no simple function can be plotted through all the points. The relationship between the two variables plotted in a scattergram may be positive negative or non-existent. If there is a positive relationship then individuals high on one variable will tend to be high on the other. If there is a negative relationship then individuals with high scores on one variable will tend to have low scores on the other.

4. Cumulative frequency plots are usually 'S' shaped curves. They can be used to determine the percentile points corresponding to any raw score. The first quartile is the 25th percentile point, the median is the 50th percentile point and the the third quartile is the 75th percentile point.

Chapter 4

Goodness of fit

Statistical concepts introduced in this Chapter

Regression, deviation, deviance, correlation (r^2 and r), reliability.

Regression - the problem

Chapter 3 contained a scattergram (Figure 3.5(a)) depicting the relationship between Scale A, a psychometric test for predicting a child's ability to learn arithmetic and the child's mark in an examination following an arithmetic course. On the whole children who do well on Scale A also do well on the examination. This trend could be summarised by drawing a straight line through the points as in Figure 4.1. When all the points fall on or very close to a straight line it is simple enough to fit such a line 'by eye'. With a data set such as this, where there is some scatter in the points, this is not advisable as it is possible that you would draw a line different from the one someone else would choose. The solution to this problem is to use a statistical procedure known as regression by which one can compute the best fitting straight line. Excel can always be trusted to do this in a sensible way and you will never need to do the calculations described in the next 4 sections. They are included because they are of theoretical rather than practical importance.

Deviations from a prediction and the deviance of a prediction

The line drawn onto Figure 4.1 can be thought of as a prediction. It allows us to predict exam score from Scale A score. For some individuals, the ones close to the line, this prediction is good, for others it is less good. Table 4.1 gives each individual's score on Scale A and the exam, and the predicted score on the exam using the formula. The difference between the prediction and the observed score is given in the column 'deviation'.

Participant	Scale_A	Exam Score	Predicted exam score	Deviation	Squared deviation
1	11	76	72.45	3.55	12.60
2	8	60	67.80	-7.80	60.84
3	11	65	72.45	-7.45	55.50
4	9	61	69.35	-8.35	69.72
5	7	74	66.25	7.75	60.06
6	17	79	81.75	-2.75	7.56
7	7	67	66.25	0.75	0.56
8	18	89	83.30	5.70	32.49
9	12	74	74.00	0.00	0.00
10	5	62	63.15	-1.15	1.32
11	13	78	75.55	2.45	6.00
12	5	70	63.15	6.85	46.92
				Deviance	353.59

Table 4.1 Data for Figure 4.1

Take participant 1 for example. His score on Scale A was 11 so the formula for the line predicted_exam_score = 55.4 + 1.55 Scale_A_Score

predicts that his exam score should be

55.4 + 1.55x11 = 72.45

His actual 'observed' exam score was 76 so the difference, or deviation from prediction, is 3.55. Some of the differences are positive and some are negative. They are plotted onto Figure 4.2 the positive ones go up from the line the negative ones go down.

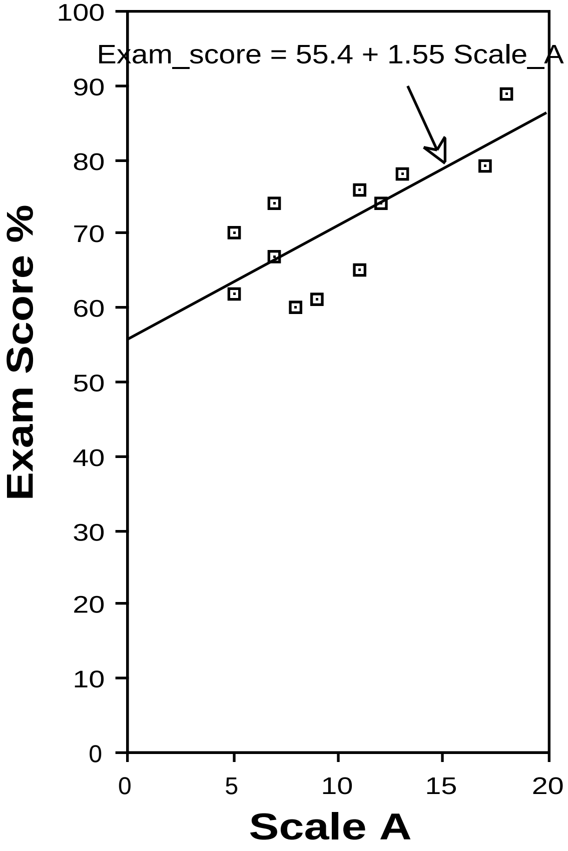


Figure 4.1

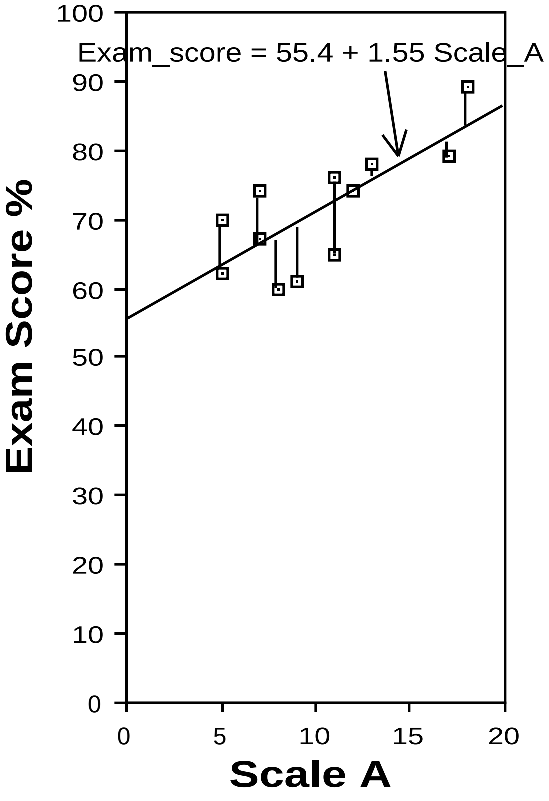


Figure 4.2

The extent to which the observed values deviate from the predictions is summarised by summing the squared deviations. This is the purpose of the last column in Table 4.1. The sum of the squared deviations (353.59) is known as the *deviance*. It is a measure of how bad the fit is. Readers will recognise this calculation as being similar to that made when computing the variance. There the mean was subtracted from the observed score and the deviations squared and summed.

As we shall see later on in this introduction, computing a deviance is the starting point of many of the calculations performed for you by Excel. The procedure is summarised in Box 4.1.

Box 4.1 Summary of the procedure for obtaining the deviance of a prediction

1. Tabulate the observed and predicted scores. In this example we are comparing the predicted_exam_score with the participant's actual exam score. The participant's predicted_exam_score is obtained from his Scale_A_Score using a linear equation (predicted_exam_score = 55.4 + 1.55 Scale_A_Score).

2. Subtract each predicted score from each observed score to get the deviation of each participant from the prediction e.g., the deviation of each participant is obtained by subtracting his predicted_exam_score from his actual exam score.

3. The deviance of the prediction is obtained by squaring all the deviations and then summing the squared deviations.

The regression equation - the best fitting straight line

We are now in a position to define what is meant by the best fitting straight line. It is the one with the least deviance. As the deviance, as defined above, is the sum of the squared deviations, this is sometimes known as the 'least squares solution'. The best fitting straight line could be determined by computing the deviance for a best guess, moving the line slightly, seeing if the deviance is better or worse, moving it again and so on. Fortunately, this kind of 'iterative' procedure is not necessary and the slope and intercept of the best fitting straight line can be computed using simple formulae. You will always use the computer to evaluate these formulae so they are not given here. In point of fact the best fitting straight line for the data in Table 4.1 is the one drawn on Figure 4.1,

predicted_exam_score = 55.4 + 1.55 Scale_A_Score

This is known as the regression equation, more precisely it is the regression equation for predicting exam scores from Scale_A_Scores or technically *the regression of exam score on Scale A*.

Two regression lines: X on Y and Y on X

Notice in order to define what we meant by 'best fitting line' we had to say what was being predicted from what. You may find it surprising to learn that the best fitting straight line for predicting Scale A from exam score is not the same line as the best fitting straight line for predicting exam score from Scale A. The best fitting straight line for predicting Score A from exam score is

predicted_Scale_A_score = -16.2 + 0.372 exam_score

Rewritten with exam score as the y variable

exam_score = 43.55 + 2.69 predicted_Scale_A_Score

which has a steeper slope and a lower intercept than the best fitting straight line for predicting in the other direction. The two lines are plotted in Figure 4.3.

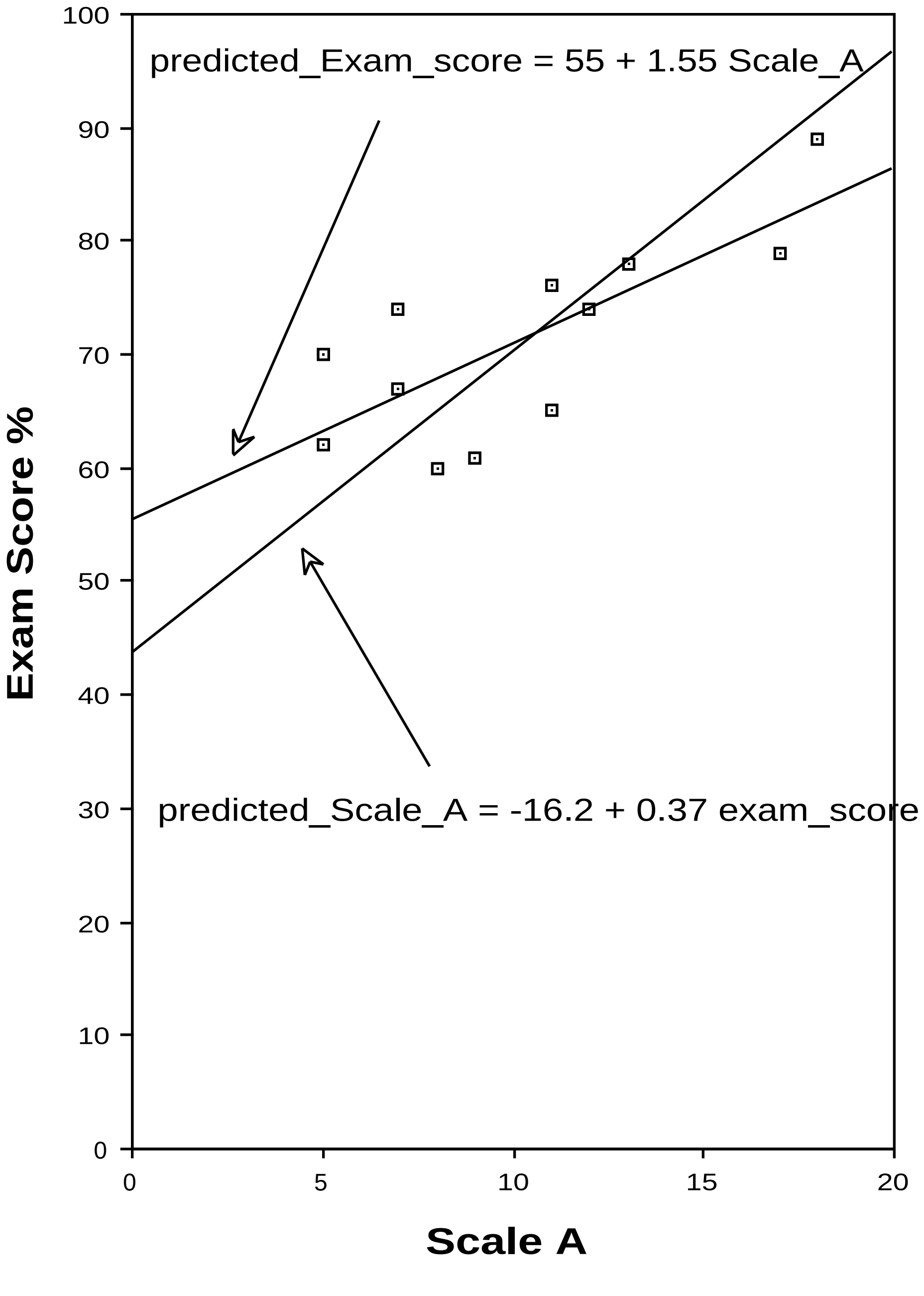


Figure 4.3 Figure 4.1 with both regression lines drawn onto it.

To reiterate, the regression line for predicting exam score from scale A is said to be the regression of exam score on Scale A. The regression line for predicting Scale A from exam score is said to be the regression of Scale A on exam score.

The two regression lines for a set of data will only coincide when all the points fall precisely on a straight line, i.e., when there is a perfect fit. When there is no real relationship between the two variables they will be at right angles to one another. This follows from the form of the regression equation. We are predicting exam score with a formula of the form

predicted_exam_score = c + m Scale_A_Score

If Scale_A_Score has no bearing on exam score then m will be zero i.e., the formula becomes

predicted_exam_score = c

Likewise the formula for predicting Scale A score will be of the general form

predicted_Scale_A_score = c

These two lines will always be at right angles to one another.

Deviance from the mean

The deviance of the best prediction that can be made of the form

predicted_exam_score = c_o

can be thought of as a baseline against which we can evaluate the regression equation

predicted_exam_score = c + m Scale_A_Score

The regression equation is the linear equation that has the smallest deviance. If we compare that with the deviance of the equation

predicted_exam_score = c_o

which has the smallest deviance we can see how much we have gained by adding Scale_A_Score to the equation.

It turns out that the least deviance is obtained if c_o is set to the mean exam score. So the baseline deviance we use to evaluate the deviance of the regression equation is the deviance from the mean. This is computed in Table 4.2.

Exam Score	Deviation from mean	Squared deviation
76	4.75	22.56
60	-11.25	126.56
65	-6.25	39.06
61	-10.25	105.06
74	2.75	7.56
79	7.75	60.06
67	-4.25	18.06
89	17.75	315.06
74	2.75	7.56
62	-9.25	85.56
78	6.75	45.56
70	-1.25	1.56
	deviance	834.25

Table 4.2 Computing the deviance from the mean for the exam scores from Figure 4.1. The deviation from the mean is computed by subtracting the mean exam score (71.25) from each participant's actual exam score.

So we know that the best prediction we can get of the form

predicted_exam_score = c_o

is when c_o = 71.25 and that this prediction has a deviance of 834.25.

The deviations going into this deviance are plotted in Figure 4.4. Compare these with the deviations from the regression equation in Figure 4.2. On the whole they are larger. This can be seen in the deviance computed when the deviations are squared and summed. The best prediction of the form

predicted_exam_score = c + m Scale_A_Score

has c = 55.4 and m = 1.55 and a deviance of 353.59. Adding Scale_A_Score has reduced the deviance from 834.25 to 353.59. This gives us a way of quantifying how strong the relationship between two variables is. This quantification is known as a correlation coefficient.

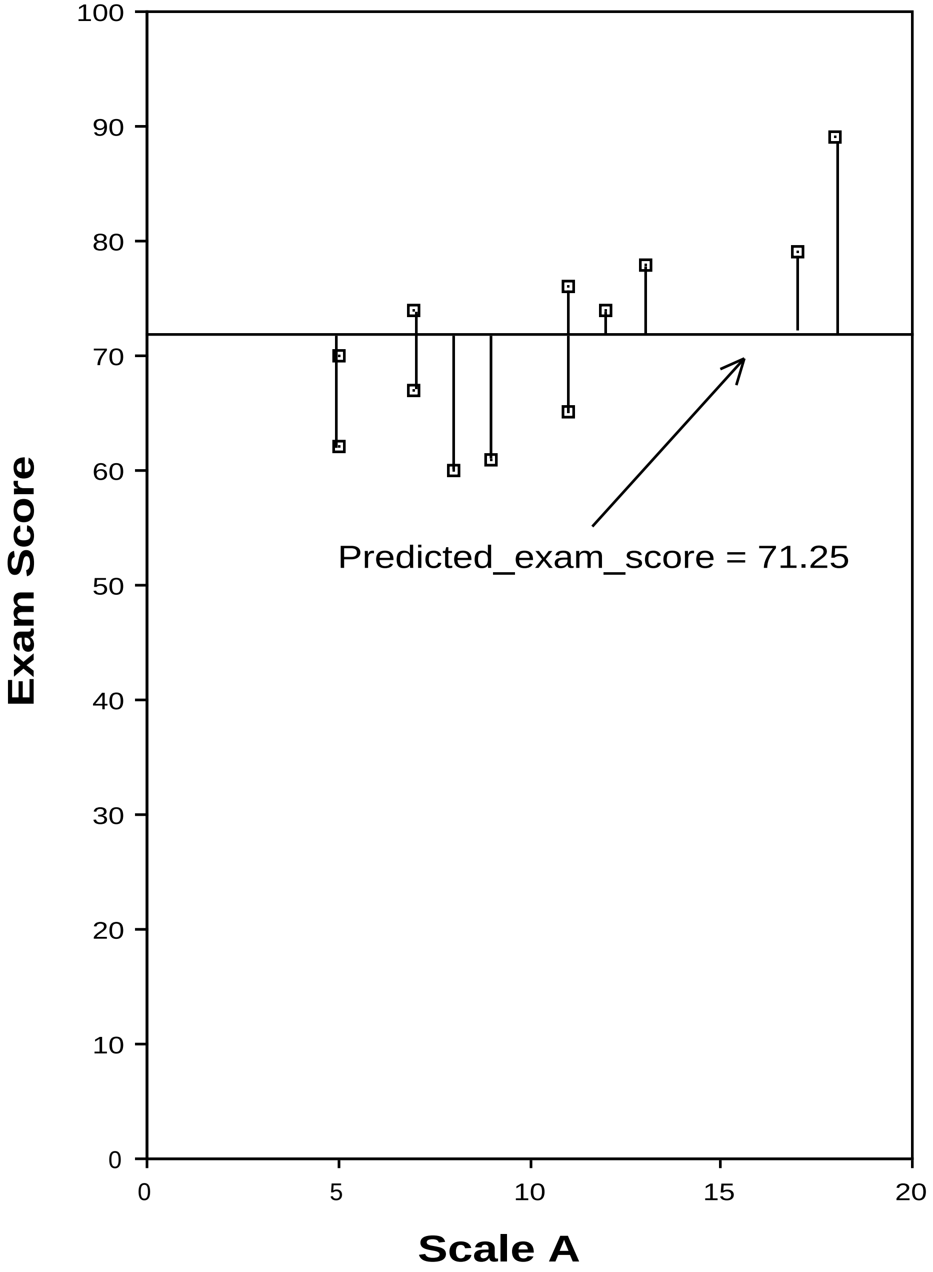


Figure 4.4 Figure 4.1 with deviations from the mean drawn onto it.

Correlation

Regression is a procedure for finding the best fitting straight line. Correlation is a way of quantifying how good this best fit is. Figures 4.5 and 4.6 illustrate two extreme cases. Figure 4.5 illustrates the somewhat unlikely case of a perfect fit. If someone's score on Scale B is known then their exam score is known also.

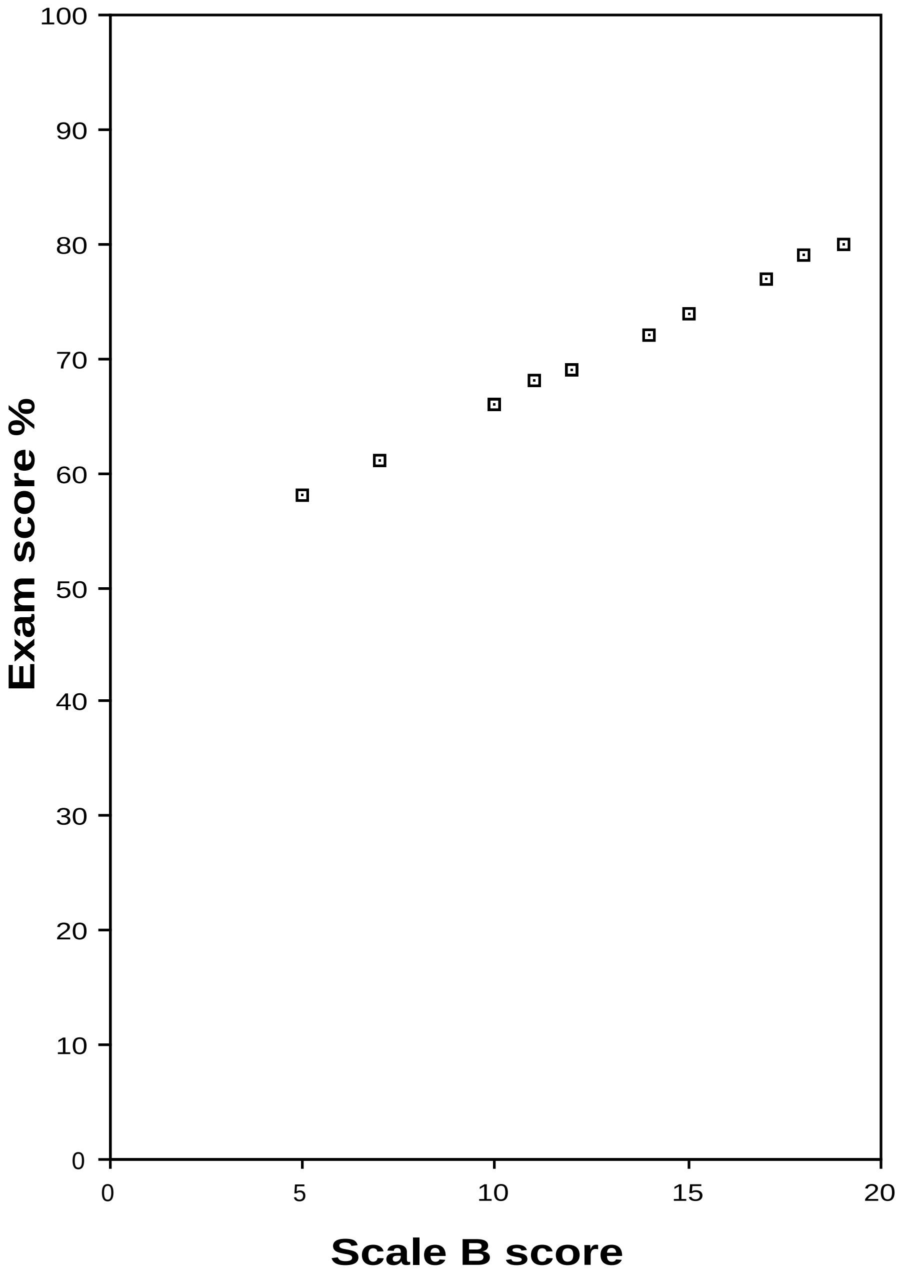


Figure 4.5 Scattergram depicting the relationship between Exam score and Scale A score. These scores fall alm exactly on a straight line.

There is no apparent relationship between the scores plotted in Figure 4.6. Someone with a high Scale C score is just as likely to have a high exam score as a low one. The situation in Figure 4.1 is somewhere in between these two extremes. There is a relationship between Scale A and exam score but there is some deviation from a perfect fit.

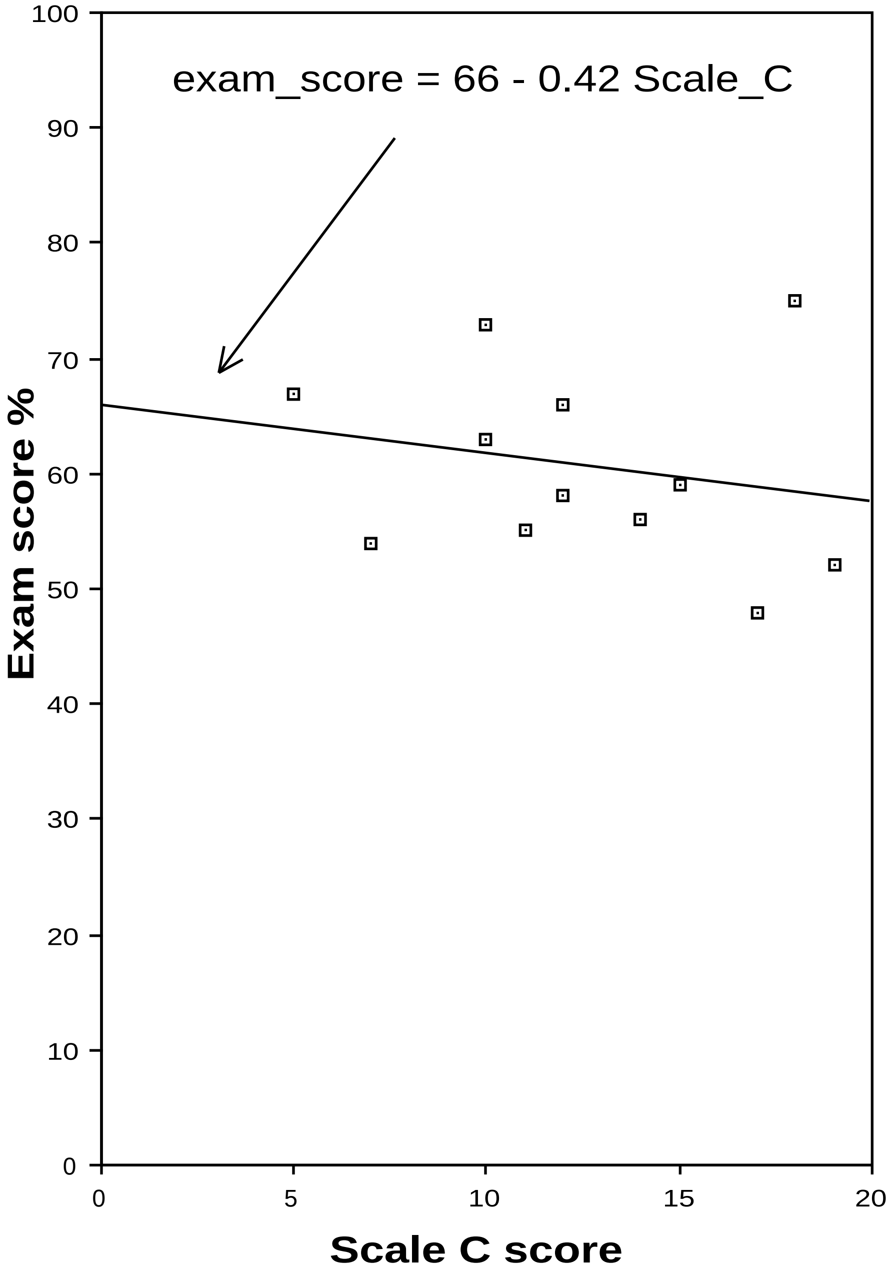


Figure 4.6 Scattergram depicting the relationship between Exam score and Scale C score. There is very little evidence of a relationship here.

The correlation coefficient r summarises the strength of the relationship. r is just one of the possible ways a correlation can be summarised. Its full name is

Pearson's product-moment correlation coefficient. This book concentrates on r to the exclusion of other correlation coefficients because it can be interpreted in terms of deviance from a prediction. This will be important when we come to consider multiple regression in Chapters 8 and 9. We will start off by considering r^2 , the square of r. This is a number between zero and one. The higher r^2 is the stronger the relationship between the variables.

For the data presented in Figure 4.5 $r^2 = 1.00$ (a perfect correlation). For the data presented in Figure 4.6 $r^2 = .047$ (nearly a zero correlation). For the data in Figure 4.1 $r^2 = .58$.

These values for r^2 can be obtained in the following way

$$r^2 = \frac{\text{deviance of the mean - deviance of the regression equation}}{\text{the deviance of the mean}}$$

So for the data presented in Table 4.1

$$r^2 = \frac{834.25 - 353.59}{834.25} = .58$$

Adding Scale_A_Score to the equation reduces the deviance from 834.25 to 353.59 i.e., by 480.66. r^2 is simply this figure expressed as a proportion of the baseline deviance (deviance of the mean).

When there is a perfect correlation between the two variables the deviance of the regression equation will be zero so r^2 will be 1.00. For example, in Figure 4.5 the exam score predicted from Scale B is equal to the actual exam score for all the participants. This means that the deviation from the prediction is zero for all participants. Summing and squaring all these zeros gives zero as the deviance. The deviance from the mean exam score is 796.8 (this is a new set of participants) so

$$r^2 = \frac{796.8 - 0.00}{796.8} = 1.00$$

For the data in Figure 4.6, where there is very little sign of a relationship, the deviance of the regression equation is very close to the deviance of the mean (see Table 4.3). There is very little reduction in deviance achieved by adding Scale C to the equation (36.43) so r^2 is close to zero.

$$r^2 = \frac{775.0 -738.57}{775.0} = .047$$

Exam	Dev. from	Sq. Dev	Scale C	Prediction	Dev.	Sq.Dev
	mean			from Scale C		
63	2.5	6.25	10	61.56	1.44	2.07
56	-4.5	20.25	14	59.86	-3.86	14.93
52	-8.5	72.25	19	57.74	-5.74	32.99
54	-6.5	42.25	7	62.83	-8.83	78.00
73	12.5	156.25	10	61.56	11.44	130.87
48	-12.5	156.25	17	58.59	-10.59	112.19
66	5.5	30.25	12	60.71	5.28	27.96
55	-5.5	30.25	11	61.13	-6.13	37.65
58	-2.5	6.25	12	60.71	-2.71	7.35
67	6.5	42.25	5	63.68	3.32	11.02
59	-1.5	2.25	15	59.44	-0.44	0.19
75	14.5	210.25	18	58.16	16.83	283.31
deviance of mean		775.0		deviance of reg.		738.57

Table 4.3 Data from Figure 4.6

The logic behind these is summarised in Box 4.2.

Box 4.2 r^2 as a measure of how good a linear model is.

(a) we have a hypothesis or model to evaluate, in this case the model is that one score can be predicted from the other with a linear equation (the regression equation);

(b) the model is to be evaluated against a simpler model, in this case we take the baseline model of predicting each score from the mean score;

(c) the deviance of each of these two models is computed by summing the squared deviations of their predictions;

(d) the deviances are compared in a statistic reflecting their difference, in this case r^2;

(e) if the deviances are similar (r^2 close to 0) the first model can be said to be a poor one, it does not do much better than the baseline model;

(g) if the deviance of the first model is much smaller than that of the baseline model (r^2 close to 1) then it can be concluded to be a good one.

Interpreting r^2

r^2 can also be interpreted as the proportion of the variance in one score that can be accounted for by the variance in the other score. So we can say that .58 of the variance in examination scores can be accounted for by the variance in Scale A scores. r^2 is the same whatever direction you are predicting so we can also say that .58 of the variance in Scale A scores can be accounted for by the variance in exam scores.

This proportion is often reported as a percentage i.e., 58% of the variance in examination scores can be accounted for by the variance in Scale A scores.

The correlation coefficient r

r, unsurprisingly, is the square root of r^2. r takes the sign of the slope of the regression equation. A positive r indicates a relationship where high values on one variable are associated with high values on the other as in our example (the r for Scale A and exam score is .76). A negative r indicates that high values on one score are associated with low values on the other. Figure 4.7 (based on Figure 3.11(b)) gives an example of such a relationship. The higher the age of a child in months the less time it takes to solve a jigsaw puzzle.

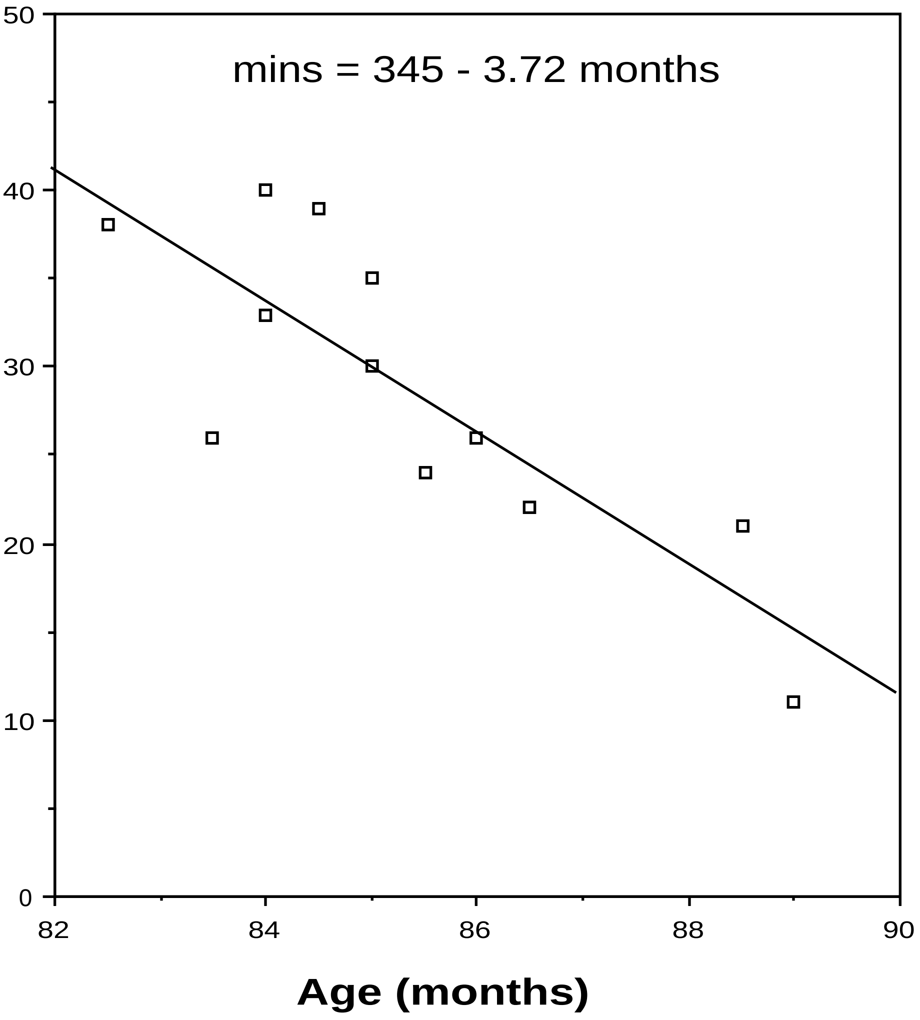


Figure 4.7 Scattergram depicting the relationship between age and time to complete a puzzle.

Using r - the reliability of a psychometric test

Correlation coefficients, particularly r, are used extensively in behavioural science. One important example of this is the concept of the 'reliability' of a psychometric test. Any measuring instrument should give you the same answer if you measure the same thing twice. A tape measure which gave the length of some object as being 12 cm on one occasion and 13 cm on another would not be of much use. Unfortunately behavioural characteristics are difficult to measure with precision and so we need to assess how repeatable the measures obtained are. Technically this is known as assessing the 'reliability' of the test and is performed as part of the standardisation of the test.

The simplest way of assessing the reliability of a test is to apply it twice to the same set of individuals. If the test has a high reliability then, in general, people will get the same score on the second test as they do on the first. We can describe the strength of that relationship by computing r. This is known as the 'test-retest reliability' of the test. Figure 4.8 is a scattergram plotting each participant's score on Day 1 (first testing) with their score on Day 2 (second testing). On the whole the results obtained on the two days are similar, the scattergram approximates to a straight line with slope of 1. r for these data is .95 which is a perfectly adequate reliability. Figure 4.9 depicts a much less satisfactory situation. Although people who do well on Day 1 also do well on Day 2 (as one would hope!) there are some quite large discrepancies. r for these data is .64. In general one is looking for a reliability of .8 or .9 in a psychometric test. This second test obviously needs further development before it can be used as a measuring instrument.

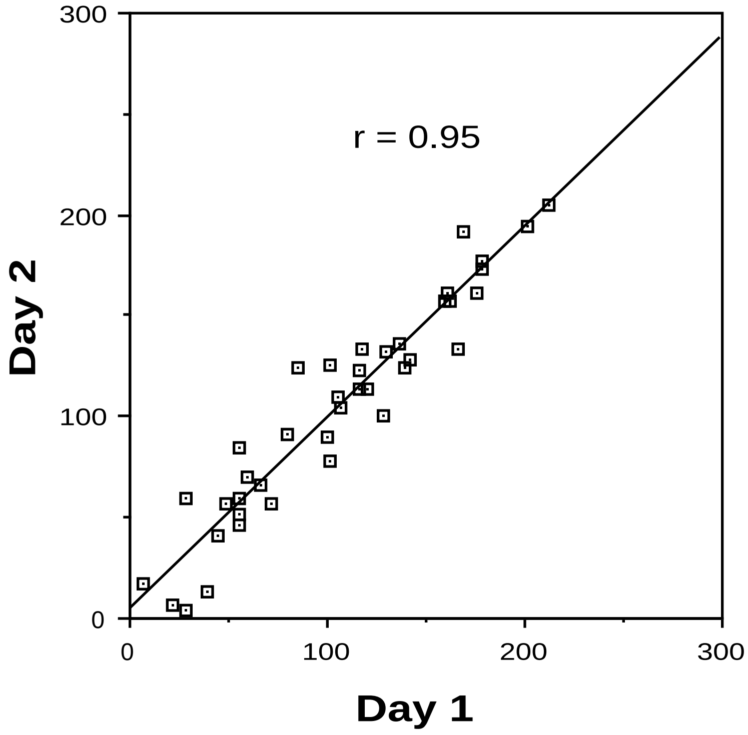


Figure 4.8 Scores obtained from the same test on two different days plotted as a scattergram. Good reliability.

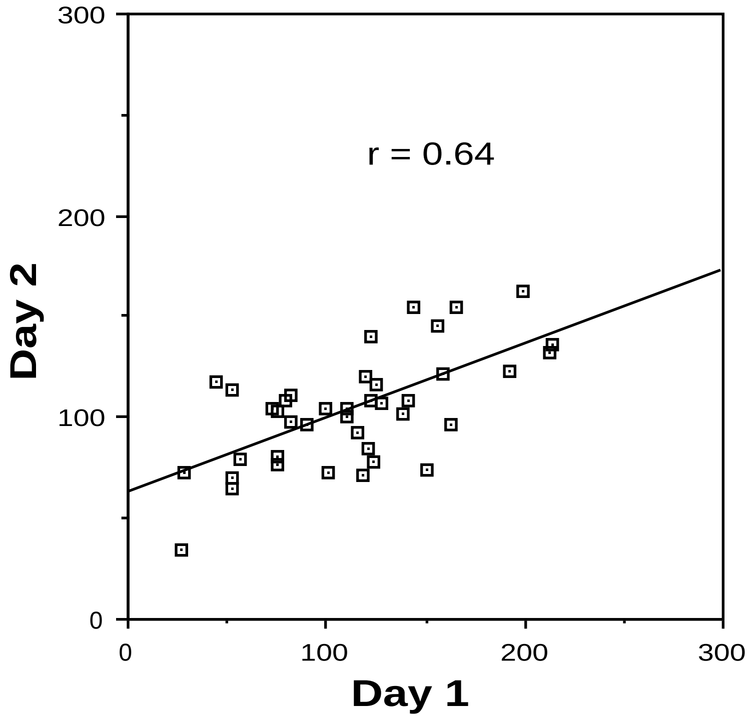


Figure 4.9 Scores obtained from the same test on two different days plotted as a scattergram. Poor reliability.

With many tests there are practical problems in interpreting a test-retest reliability. For example, a personality test may involve reading 40 statements (e.g., 'I enjoy noisy parties'). You have to say whether you agree or disagree with each statement. Let us say that you are given this test one day and then again the next day. On the second day it is quite likely you will be able to remember your responses from the first day. This may result in an artificially high estimate of the reliability of the test because you are trying to be consistent. More generally, the problem of using a test-retest reliability is that testing the sample on the first occasion may somehow affect their results when they are tested on the second occasion.

One solution to this problem is to have two versions of a test. Version 1 can then be used in the first testing session and Version 2 in the second. The correlation between the scores of each individual on two versions of a test is known as the 'alternate forms' reliability of the test. This also has its problems. The alternate forms reliability of a test may be an under-estimate if the two versions are not really equivalent. It is also extremely tedious to have to prepare two tests when one is only for testing reliability.

The most commonly used measure of reliability is 'split-half' reliability. This is similar to alternate forms reliability except that the two versions of the test are obtained by dividing the test into halves. Consider again the example of a personality test consisting of 40 statements. In the normal course of events a score would be obtained for the whole test by examining the participant's response (agree or disagree) for each of the 40 statements. To obtain a split-half reliability the test is administered in the normal way but scored as if it had been two 20-item tests. Perhaps the odd statements are treated as if they were from one test and the even items as it they were from another. We can then correlate the scores for the two halves. All the items in a test should be measuring the same thing so this correlation will reflect the reliability of the two halves of the test. If there is a high correlation the reliability is high. If not it is not.

The only problem with the procedure as stated this far is that the results from two 20-item tests have been correlated, where as the final test contains 40 items. This will make the correlation computed an underestimate. Imagine an arithmetic test base on 3 addition problems. How would you expect its reliability to compare with one based on 100 addition problems? Other things being equal the more items you have the more reliable a test will be. So, we have estimated the reliability of a 40 item test by correlation two 20-item tests. The resulting under-estimation is systematic and we can correct for it using a formula

$$r_{\text{split-half}} = \frac{2r}{1+r}$$

where r is the correlation between the two halves of the test. This is known as the Spearman-Brown formula.

Summary

1. The trend in a scattergram can be summarised by a straight line.

2. The formula for the best fitting such straight line is known as the regression equation.

3. 'Best fitting' is defined as follows:

(a) One variable, the predictor variable, is being used to predict the other, the predicted variable

(b) A linear equation is used to compute a predicted value from the predictor value

(c) The deviation is the difference between the observed value of the predicted variable and its predicted value

(d) The deviance is the sum of the squared deviations

(e) The linear equation with the lowest deviance is the regression equation.

4. The regression equation for predicting A from B (the regression of A on B) is not the same line as that for predicting B from A (the regression of B on A) unless all the points fall on a straight line.

5. Correlation is a measure of how good the best fit is.

6. r^2 is used to compare the deviance of the regression equation with the deviance from the mean. (r is the Pearson product-moment correlation coefficient).

7. $r^2 = \dfrac{\text{deviance of the mean - dev. of the regression equation}}{\text{the deviance of the mean}}$

8. r^2 can be interpreted as the proportion of the variance in one score that can be accounted for by the variation in the other score.

9. r takes the sign of the slope of the regression equation. So a positive relationship has a positive r and a negative relationship has a negative r. No relationship is indicated by an r approaching zero.

10. The reliability of a psychometric test is measured by obtaining two estimates of each participant's score on it and then correlating these scores.

Chapter 5

Inference with Statistics

Statistical concepts introduced in this Chapter:

The null hypothesis, p value, significance level, sign test, significance of a correlation.

The problem

Chapter 1 showed how an initial question is refined into an experiment. This is the business of science. The question considered in Chapter 1 may not be of great scientific importance but it serves to illustrate this process. The general question was 'Do people prefer butter to margarine?' There are various ways that this general question could be translated into an experiment and two are described below. In Chapter 1 and Experiment 1 below, the more specific experimental question asked was 'Do people who taste butter give higher or lower preference ratings compared with people who taste margarine?'

Chapters 2, 3 and 4 described how the results of experiments can be summarised. In this case the summary statistic used would be the mean. The mean preference rating for a group tasting butter is compared with the mean preference rating for a group tasting margarine. The results of other experiments might be summarised using a contingency table or a correlation.

A result from an experiment can be thought of as providing evidence for or against some hypothesis. Say that the butter group gives a higher mean rating than the margarine group. This supports the hypothesis that 'butter gives higher preference ratings'. Let us assume that the experiment was a fair test of this hypothesis. There is still a potential problem when interpreting the result. It may not be trustworthy e.g., there might not be any real difference in preference rating, just by chance, the result has come out this way. If this were the case and we were to repeat the experiment with some new participants at some time in the future we might draw the opposite conclusion.

Scientists need to be sure that if they repeat an experiment they will come to the same conclusion. Three imaginary experiments are described below. In each case two research teams, A and B, both carried out the experiment and came to differing conclusions. Examine these conclusions and the data they are derived from. Which research team do you feel is producing the most trustworthy results i.e., which research team's results are most likely to be repeatable?

Experiment 1 - Preferences for butter and margarine when the taster samples only one fat.

Tasters are given one sample of fat and asked to make a preference rating on the scale 'very nasty' to 'very nice' as described in Chapter 1.

Box 5.1 Results for Experiment 1

Team A results

15 participants were given butter their ratings were:

34 38 60 44 47 31 58 53 37 37 46 43 54 35 44

Mean 44.07, Standard deviation 8.99.

15 participants were given margarine their ratings were:

32 31 19 38 34 41 40 27 28 19 51 27 30 34 29

Mean 32.0, Standard deviation 8.32.

Conclusion: People rate butter more highly.

Team B results

10 participants were given butter their ratings were:

36 5 48 46 44 42 59 44 44 29

Mean 39.70, Standard deviation 14.43.

10 participants were given margarine their ratings a were:

58 62 55 39 63 38 70 43 3 45

Mean 47.6, Standard deviation 19.13.

Conclusion: People rate margarine more highly.

Experiment 2 - Preference for margarine or butter in a forced choice discrimination

Tasters are given two samples of fat. They know that one sample is margarine and the other is butter but not which order they are presented. Half the participants have the order butter-margarine the other half margarine-butter. They are required to state which sample they prefer, the first or the second, after tasting both samples.

Box 5.2 Results for Experiment 2

Team A results

Sixteen participants were tested, 14 chose the sample which was butter.

Conclusion: People prefer butter.

Team B results

Ten participants were tested, 7 chose the sample which was margarine.

Conclusion: People prefer margarine.

Experiment 3 - The relationship between extroversion and the use of extreme categories in a rating scale

People vary in their willingness to use the extremities of a rating a scale such as 'very nice' or 'very nasty'. Some tend only to use more central ratings. The question is whether this can be predicted from personality tests. Participants are given a sample of butter and asked to rate its taste on a 100 mm scale 'very nasty' to 'very nice' as in Experiment 1. Their willingness to use the ends of the scale is established by measuring the distance of the mark they make from the centre of the line. Before they do this they fill in a personality questionnaire which gives an extroversion score. Previous work has established that overall preference for butter cannot be predicted from extroversion.

Box 5.3 Results of Experiment 3

Team A results

Figure 5.1a is a scattergram plotting extroversion against distance from the centre. There is a positive relationship evidenced by a correlation coefficient of .87

Conclusion: The more extroverted participants are more likely to use the extremities of the scale.

Team B results

Figure 5.1b is a scattergram plotting extroversion against distance from the centre. There is a negative relationship evidenced by a correlation coefficient -.33

Conclusion: The more extroverted participants are less likely to use the extremities of the scale.

Team A's results versus Team B's

In all three experiments Team A's results seem more trustworthy. Were the experiments to be repeated one would expect to come to the same conclusions. Team B's results seem to contain much more random fluctuation than Team A's. Perhaps they did not take the trouble to put the participants at ease, perhaps they did not give clear instructions or sufficient practice. Whatever the cause, their conclusions appear to be untrustworthy given the data presented.

For example, in Experiment 1 the difference in means obtained by Team B is small compared to the random variation within groups. It is quite likely that the difference observed in the mean scores could have arisen from these random fluctuations. In contrast, the means obtained by Team A differ more than one would expect from the within group variation. In Experiment 2, Team B find more participants prefer the sample which is margarine but only 7 out of 10. If they had been responding at random one might have got 7 out of 10 responses in one direction rather than the other. Team A, on the other hand, found 14 out of 16 preferred butter. This is very unlikely to have happened if they were responding at random. Finally, in Experiment 3 the scatter in Team B's results is considerable and the correlation is only -.33 (11% of the variance). It would not be surprising if the regression line sloped in the opposite direction when the experiment is repeated. Figure 5.1a shows very much less scatter. The correlation might not be exactly .87 if the experiment were to be repeated but it would be surprising if it forced a change in conclusion because it sloped in the opposite direction.

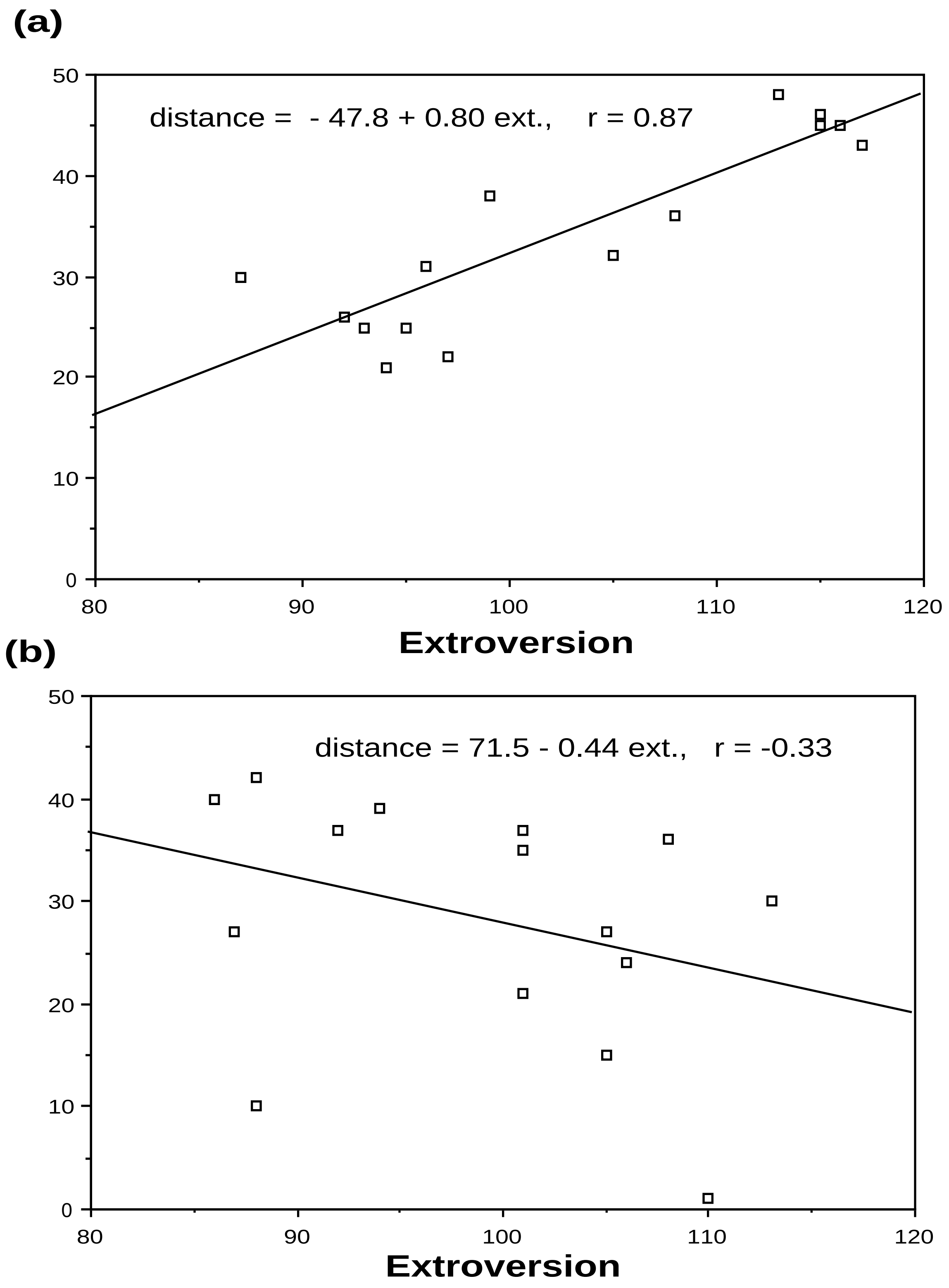


Figure 5.1 Scattergrams depicting the relationship between extraversion score and the distance of the rating made from the centre of the scale, (a) results from Team A, (b) results from Team B.

It is possible to check whether a result is repeatable by doing the experiment again. This would be a test of reliability parallel to the use of correlation coefficients in Chapter 4. Scientists do repeat experiments, indeed important results need to repeated, if possible by different investigators. This is known as replicating an experiment. However, we also need a way of estimating how trustworthy an individual result is without actually having to repeat it. Statistical inference has a different logic to reliability testing but the aim is the same, that is, to test how much one can trust a given result.

Statistical inference

A chance result is inherently untrustworthy, on one occasion we may draw one conclusion and on another the opposite. The purpose of tests based on statistical inference is to reject the possibility that a result has arisen by chance. This is done by computing the probability of getting a result as extreme as the one obtained, by chance. If that probability is very small the possibility that the result was due to chance is rejected. Technically we say that the result is *significant*.

The logic of statistical inference is as follows:

(a) We have a result to evaluate. We wish to draw some conclusion from this result (e.g., X is greater than Y).

(b) We want to reject the possibility that the result arose by chance, (e.g., X is really the same as Y).

(d) If we can do this then it is probably a trustworthy result.

As explained above, statistical tests apply this logic by computing a probability. This is the probability that the result arose by chance. If this is small we say the result is 'statistically significant' and reject the possibility that the result occurred by chance. To do this we have to define precisely what is meant by chance and, as with the computation of any a priori probability, make certain assumptions. The different statistical tests you will learn in this book all follow this logic. They differ in the way 'chance' is defined and the assumptions made to compute the probability. The definition of chance used by the test is known as the null hypothesis.

This process can be illustrated by considering Experiment 2.

1. The experimental hypothesis is that people can discriminate butter from margarine and that the probability of choosing butter is not the same as the probability of choosing margarine.

2. The null hypothesis is that each individual is equally likely to choose each of the fats i.e., the probability of choosing butter is .5 and the probability of choosing margarine is .5.

3. One further assumption is necessary to compute a probability using this null hypothesis. This is that the choices made are independent of one another i.e., one participant's choice had no effect on any other participant's choice.

4. The probability of getting a result as extreme as 14 out of 16 participants choosing butter purely by chance is .0042. This is known as the p value and is usually written 'p = .0042'

5. .0042 is a very small probability. An event with probability .0042 will only occur 42 times in every 10,000 i.e., 4.2 times in every 1000 or .42 percent. We

would expect to have to repeat this experiment on average about 240 times before getting a result as extreme as this by chance. This could be that occasion but that is most unlikely. It would seem quite safe to reject the null hypothesis and conclude there is a real effect. People do prefer butter.

Now consider Team B's results. The probability of getting a results as extreme as 7 out of 10 is .3438 (p = .3438). This is not a small probability. An event with probability .3438 will occur 34 percent of the time. If we performed the experiment 10 times we would expect about three of those experiments to produce results as extreme as this. It would seem unwise to reject the null hypothesis in this case. The result could well have occurred by chance. If so, were the experiment to be repeated, the conclusion drawn might be quite different.

Significance levels

The null hypothesis can be rejected when the probability of getting a result, assuming the null hypothesis is true, is very small. But how small does the probability have to be? The answer is set by convention as something smaller than .05. When the null hypothesis can be rejected the result is said to be significant and the maximum probability which is acceptable is said to be the significance level. Experience has shown that .05 (1 in 20) is small enough to prevent us from building theories on chance results but not so strict that experiments become extremely expensive to run because very large quantities of data have to be collected in order to rule out the possibility of a chance result.

Some authors refer to the .05 level of significance as the '5% level of significance'. This is simply a matter of style. It is more common to refer to a significance level as a probability than a percentage and this is the convention which will be followed in this book.

Box 5.4 summarises the last two sections.

Box 5.4 the steps taken when a statistical test is applied to some data

There is a result to be evaluated (in our example, 14 out of 16 choose butter). We wish to distinguish between a chance result and a real result. A chance result is by definition untrustworthy.

(a) Precisely what is meant by 'chance' must be formulated. This is the null hypothesis. (In our example this is that the probability of one individual choosing butter is .5).

(b) A p value, the probability of getting the result being evaluated, or something even more extreme, is computed under the assumption that the null hypothesis is true. This involves making certain further assumptions about the data. (In our example that choices are independent).

(c) If p is very small i.e., less than the significance level of .05, the null hypothesis is rejected and the result is said to be significant. It is very unlikely that this result, or something better, could have arisen by chance so it probably is a real result. (In our example p = .0042 so the result is significant at the .05 level).

The binomial test (sign test)

The computation used to illustrate statistical inference above is known as a binomial test, or sometimes the sign test. This statistical test was considered first because the null hypothesis used is easy to understand. This is that each individual had an equal probability of preferring butter and margarine. The computation uses something called the binomial expansion and the probabilities generated are said to form a binomial distribution. There is no need to understand the precise nature of the computation. Excel does most of the work of computing a p value for you.

This statistical test is also known as the sign test because it can be used in situations where one counts the number of participants whose results show differences in some predicted direction. For example, in Experiment 2 one might have had participants rate each of the fats by making a mark on the preference scale used in Experiment 1. The prediction is that butter will receive a higher rating. For each participant, it is possible to subtract the rating made for margarine from the rating made for butter. This gives a difference score the sign of which indicates whether the prediction is supported. A positive difference supports the prediction a negative difference goes against it. Counting the number of positive signs indicates how many participants prefer butter to margarine.

In Experiment 2, as originally described, there was no difference score to compute and so no signs to count but the principle is the same in that the participants supporting a hypothesis are counted. There are better tests for use with difference scores (see Chapter 7) and so the binomial test is rarely used by counting positive or negative signs. It is nearly always used with data like those in Experiment 2 where the participant makes some response which is classified as being for or against the hypothesis.

Technical note - what 'a result as extreme as' means in this context

The probability to be computed is the probability of getting 'a result as extreme as 14 out of 16 choosing butter' not the probability of precisely '14 out of 16 choosing butter'. The probability of precisely 14 out of 16 participants choosing butter, under the null hypothesis is .0018 (this is computed using something called the binomial expansion). The p value computed in a binomial test is .0042. This discrepancy arises because when we say that a result of 14 out of 16 is significant we are really setting a criterion. Clearly if 14 out of 16 is acceptable then so is 15 or 16 out of 16. So if the criterion for significance is 14 out of 16 the probability of getting 14, 15 or 16 out of 16 must be less than .05. The probability of getting 14, 15 or 16 out of 16 is .0021.

Now imagine that the result was actually 2 out of 16 choosing butter (14 out of 16 choosing margarine). Even though the experimental hypothesis was that butter is most likely to be chosen we would not discard such a result. This means that 0, 1 or 2 out of 16 choosing butter must also be included in our class of 'interesting' results. Thus the probability we want is the probability of getting 0, 1, 2, 14, 15, or 16 out of 16. This probability, the p value, is .0042.

The t-test and other tests for comparing means

Statistical tests for evaluating the significance of differences between means are discussed more fully in Chapters 6 and 7 but the results of Experiment 1 will be considered briefly here for completeness. The mean preference rating for butter could be compared with the mean preference rating for margarine by using one of these tests. Here a real result is a true difference in the mean rating response. The null hypothesis is that there is really no difference but that the apparent difference arose from the random fluctuations observable in the variance within the two groups. The t-test, to be introduced in Chapter 6, is one way of computing probabilities under this null hypothesis. Having done so the procedure is the same as for the sign test. If the p value is less than or equal to .05 the result is said to be significant. If it is not we cannot reject the null hypothesis.

A t-test performed on the results of Experiment 1, for Team A, shows that the probability of getting a difference as big or bigger than 12.07 (44.07 - 32.0), given the fluctuations apparent within the two groups, is .0007 (p = .0007). This is smaller than .05, the result is significant at the .05 level. We can be confident that if Team A were to repeat the experiment they would come to the same conclusion.

The same cannot be said of Team B's results. A t-test shows that a difference of 7.9 (47.6 - 39.7) or more, given the larger variation within groups for Team B, has a probability of .31 of occurring (p = .31). This is not significant at the .05 level. The difference between these two means may well be a chance difference. If Team B were to do the experiment again they might well come to the opposite conclusion.

The significance of r

With a correlation the conclusion to be drawn concerns the direction of the relationship. Is X positively related to Y so that X increases when Y increases, or are they negatively related so that as one increases the other decreases?

A real result in Experiment 3 is a true trend, either a positive trend such that the higher a participant's extroversion score is the more extreme the ratings they make, or a negative trend such that the lower a participant's extroversion score the more extreme the ratings. The null hypothesis is that the true regression line has no slope and so the 'real' correlation is zero but, because of random scatter (i.e., chance fluctuations), the observed regression has some slope. The Excel regress command computes the probability under this null hypothesis.

Team A obtained a correlation, r = .87. The conclusion to be drawn is that there is a positive trend. The probability that a correlation as strong as this or stronger could occur by chance is .0001 (p = .0001). This is less than .05 so the correlation is significant. If Team A were to repeat Experiment 3 they might not get a correlation of exactly .87 but they are likely to come to the same conclusion.

Team B's correlation is much weaker, r = -.33. The probability that a trend as strong as this or stronger could occur by chance is .227 (p = .227). This is not significant at the .05 level and may well be a chance result. If Team B were to repeat the result they might well come to the opposite conclusion.

Critical values of r - Table A.1

Excel computes a p value for a correlation as part of the regress command (see Box 5.7 in the work sheets for a step by step procedure to follow). An alternative way of assessing the significance of a correlation is to use Table A.1. This has been constructed by working out the minimum correlation that would be significant for a sample of a particular size. The larger the number of participants in the sample (N) the smaller the correlation coefficient can be and still be significant. These values are known as critical values of r. If the r you have just computed is greater than the critical value (ignoring the sign of the correlation) then it is significant at the .05 level.

Both teams tested 15 participants (N = 15). The critical value from Table A.1 is thus .514. .87 is larger than .514 so the correlation obtained by Team A is significant at the .05 level. We ignore the sign of the correlation when using Table A.1 and .33 is smaller than .514 so the correlation obtained by Team B is not significant at the .05 level.

Table A.1 may be particularly useful when determining which correlations are significant in a correlation matrix. N will be the same for all the correlations and so it is only necessary to look up the critical value of r for that N and then see which correlations are larger than or equal to that value. A step by step procedure to compute a correlation matrix and then evaluate it using Table A.1 is specified in Box 5.6 in the Work sheets.

Note that quite small correlations can be significant if there are large numbers of participants. For example, Table A.1 shows that a correlation of .2 is significant for a sample of 102 participants. Such a correlation only accounts for 4% of the variance (r^2 = .04). Significance indicates that the sign of the slope of the regression is likely to be the same if the experiment is repeated. It does not necessarily say anything about the amount of variance that regression accounts for. So to indicate the strength of a correlation quote r^2. In general, the significance of any effect should not be taken as an indication of its strength.

Advanced note - how Excel computes the p value for a particular r^2

Chapter 4 introduced the correlation coefficient through the concept of the deviance of a regression equation. r^2 was defined as

$$\frac{\text{deviance of the mean - deviance of the regression equation}}{\text{the deviance of the mean}}$$

The deviance of the mean was viewed as a baseline against which the regression equation was evaluated. The example used was the correlation between exam scores and a measure of arithmetical ability, Scale A. The deviance of the regression equation for predicting exam scores from Scale A was 353.59 and the deviance of the mean was 834.25 so adding Scale A to the equation reduces the deviance by 480.66.

When Excel computes the p value for a correlation coefficient as part of the regress command it does so by computing a statistic called an 'F ratio' from these deviances. The deviances computed 'by hand' in Chapter 4 can be found in the display generated by the regress command. Box 5.5 contains part of the display generated for these data. The deviances can be found in this table. They are printed in bold in Box 5.5. This table also included the F ratio and the p value computed from the F ratio. The F ratio evaluates the significance of the reduction in deviance. The null hypothesis is that there really is no correlation and the reduction in deviance is due to chance. p = .004 and so the F ratio is significant at the .05 level.

Box 5.5 Regression of Exam score on Scale A score from Chapter 4. Emphasis (bold face) added.

```
SOURCE        DF          SS          MS         F        p
Regression     1      480.67      480.67     13.59    0.004
Error         10      353.58       35.36
Total         11      834.25
```

One- and two-tailed tests

Some statistics books give procedures for performing one- and two-tailed tests. In this book we consider only two-tailed tests. The distinction has to do with the form the experimental hypothesis takes. A one-tailed hypothesis is strongly 'directional'. In Experiment 2 for example, we might have predicted that butter would be preferred to margarine. If the prediction was so strong that a result in the opposing direction would simply not be considered then we have a one-tailed hypothesis. Thus, under a one-tailed hypothesis, finding that more people prefer margarine is equivalent to finding that exactly half prefer margarine. There is no point in further analysing either result.

In Psychology hypotheses are very rarely so strongly directional. We might *expect* butter to be preferred, but if it turns out the other way around we are unlikely to discard the result. A one-tailed binomial test for Team A, Experiment 2, would be to compute the probability of getting 14, 15 or 16 out of 16. The two-tailed test, used here, computed the probability of getting 0, 1, 2, 14, 15 or 16 out of 16. For binomial and t-tests the p value for a one-tailed hypothesis is half that for the two-tailed test.

How to report a correlation

When writing up the results of statistical tests in a paper or some other form of report it is important to give the right amount of detail. The reader of the report will not want to be swamped with unnecessary information such as part calculations or the explanation of statistical concepts such as significance. On the other hand sufficient detail is required to show that the appropriate test has been performed. A statistical test is used to evaluate a result so the first step should be to report that result. Here the result to be evaluated is a correlation and the first thing to be reported is that correlation and the conclusion it suggests. Having done this the significance of the correlation is reported.

When interpreting a correlation it is often useful to know the mean and standard deviation of each of the scores. This allows one to judge whether the range of scores obtained is typical of the population concerned. One might also plot a scattergram to illustrate the relationship graphically. The correlation coefficient r indicates the strength of the linear relationship between the variables. A strong non-linear relationship would show up in a scattergram but not as a high r.

Team A's results for Experiment 3 might be written up as follows.

> Results:-
>
> There is a positive correlation (r = .87) between extroversion score and the extremity of the rating made (distance from the central point of the rating scale). This strong positive relationship is in line with the predictions made. The correlation is significant at the .05 level.
>
> Table 5. 1 gives the mean and standard deviation for each of the scores. Figure 5.1a is a scattergram using these data. There is no evidence of any additional non-linear relationship.

	Mean	Std.Dev.
Extroversion score	102.8	10.4
Extremity of rating	34.2	9.5

Table 5.1 Means and standard deviations, Experiment 3, Team A.

How to report a binomial test

Similar principles apply to the reporting of the results of a binomial test. Here it is sufficient to report the number of individuals who support the hypothesis and the total number of individuals considered. Since binomial tests are often (incorrectly) reported as one-tailed probabilities it is worth saying that you have used a two-tailed test.

Team A's results for Experiment 2 might be written up as follows.

Results:-

Sixteen participants were tested of whom 14 chose the sample which was butter. A two-tailed binomial test shows this to be significant at the .05 level (p = .0042).

Summary

1. Inferential statistics is the name given to a set of procedures for assessing whether a result is trustworthy i.e., whether one will come to the same conclusion if the experiment is performed again.

2. All these procedures work as follows

(a) A chance result is inherently untrustworthy. If we can reject the possibility that a result is due to chance then it is probably a trustworthy result.

(b) 'Chance' is defined as a null hypothesis, a 'no effect' hypothesis which can be stated with sufficient precision to make possible the calculation of probabilities.

(c) The probability of getting our particular result, or one which is more extreme, under the null hypothesis is computed. This is the p value.

(d) If that probability is less than or equal to the significance level the null hypothesis is rejected and the result is said to be significant.

(e) The most common significance level, and the one used consistently through this book, is .05. If the probability of getting the result, under the null hypothesis, is less than or equal to .05 it is probably repeatable.

(e) If it is greater than the significance level it is not possible to reject the null hypothesis.

3. The binomial test works as follows (See Box 5.8 for the detailed procedure to use)

(a) Each participant is classified as supporting the hypothesis (+) or not supporting it (-). The number of + participants is counted, let us say that it is M out of N.

(b) The null hypothesis is that the probability of an individual participant being + is equal to the probability they are - (.5).

(c) Excel uses the binomial expansion to compute a p value. The underlying assumption is that each participant's result (+ or -) is independent of the result of every other participant.

(d) If that probability is less than .05 the result is said to be significant at the .05 level.

(e) If it is greater than .05 we cannot reject the null hypothesis.

4. The Excel command regress computes the p value of r by first computing an F ratio (see Box 5.7 in the work sheets for the procedure to follow).

5. As an alternative, a critical value of r can be looked up in Table A.1 in the appendix. If the correlation to be evaluated is larger than this critical value then it is significant at the .05 level (see Box 5.6 for detailed procedure).

(Tests for comparing means are discussed in more detail in Chapters 6 and 7)

Chapter 6

Comparing means

Statistical concepts introduced in this Chapter:

The t-test, the Mann-Whitney U test.

This Chapter is concerned with the problem of evaluating a difference between means. The example developed in Chapter 5 was the comparison of two mean preference ratings, one from a group tasting butter and the other from a group tasting margarine. Team A obtained a mean rating of 44.1 from the butter group and 32.0 from the margarine group. The conclusion is that butter gives higher ratings. Team B got a mean rating of 39.7 for their butter group and 47.6 for their margarine group which would lead one to the opposite conclusion. In Chapter 5, Team B's results were dismissed because the difference between the two means was small compared with the large variation of scores within groups. Given this large 'error' variation the difference between the means could easily have arisen by chance. The purpose of this Chapter is to show how this intuition is quantified in statistical tests.

71	61	48	40	50	53	35	62	19	39	42	59	40	63	61	36	39	69	37	40	27	63
52	61	20	76	40	42	38	46	54	41	54	38	38	23	47	28	54	44	62	53	39	48
68	49	37	51	55	52	46	60	71	36	69	41	34	56	43	22	44	59	55	27	7	49
35	49	71	27	34	26	89	61	71	23	51	58	76	57	39	51	67	49	43	24	45	61
38	52	49	71	58	61	48	57	42	53	51	70	28	32	57	62	60	71	47	52	39	42
23	67	65	45	91	46	14	67	48	46	37	41	23	58	56	64	54	60	52	88	50	50
64	66	58	57	37	43	47	71	37	64	35	46	62	69	28	50	41	60	60	76	49	41
57	51	61	58	53	40	67	50	47	41	46	21	49	72	26	58	64	43	67	92	40	53
45	60	36	37	31	54	46	52	72	60	54	26	50	42	69	52	74	42	61	43	51	36
45	64																				

Table 6.1 A population of scores

The null hypothesis

Table 6.1 contains 200 randomly generated numbers. Imagine they were obtained by getting a rating from all the students on a particular statistics course. Each student is given a sample of butter and asked to make a rating, from 'very nasty' to 'very nice' as in Chapter 1. Ten values are randomly sampled from this population of scores.

71, 27, 34, 26, 89, 61, 71, 23, 51, 58, (Mean = 51.1)

Sampling another ten scores at random

46, 52, 72, 60, 54, 26, 50, 42, 69, 52, (Mean= 52.3)

gives a slightly different mean. You may want to repeat this process. Chose ten numbers from the table using a pin. It is very unlikely that your sample will have the same mean as either of the other two samples. This is because each sample mean is only an estimate of the population mean (49.92). The above exercise illustrates the null hypothesis when comparing means. This is that the two samples to be compared in fact come from the same population and the means differ only through chance.

So, the statement that Team B's results "could have arisen by chance" is reformulated more precisely as the null hypothesis that "the two sets of scores come from the same population of scores". Giving one group one fat and the other another has had no effect. If we could reject this null hypothesis because the probability of getting a difference of this magnitude or greater is very small (less than .05) then we could say that the result was significant i.e., the result is likely to be replicable. As we saw in Chapter 5, we cannot reject the null hypothesis when evaluating Team B's results and so there is the possibility that the means do differ only by chance. If the experiment were to be repeated we might come to the opposite conclusion that butter is preferred to margarine.

The problem then is to compute the probability that the two sets of scores to be compared have been drawn from the same population of scores. As with the computation of any a priori probability it is necessary to make assumptions. There are two ways of doing this, they are described in the sections that follow.

The normal distribution

Statisticians have shown that many distributions observed in 'nature' have a characteristic shape which can be described by a mathematical function known as the normal or Gaussian distribution function. A normally distributed variable arises when that variable is determined by a large number of independent factors. Height is the classic example. The height of an individual is determined by many different genes in the individual's genotype and also by many different environmental factors. Figure 6.1 is a histogram drawn from the 200 points in Table 6.1. It conforms to this general shape, being approximately symmetrical about the mean, with the most common values being central. As values deviate further from the mean they become less frequent giving the distribution a characteristic symmetrical bell shape.

The normal distribution is convenient for statistical calculations as it can be described completely by two 'parameters', its mean and its standard deviation. Thus it is known, for example, that 68.26% of all the scores in a normal distribution will fall within one standard deviation either side of the mean and 95.44% within two standard deviations either side of the mean. This is of considerable practical importance. Imagine you are screening infants for abnormalities in a clinic. One thing of interest is the circumference of the infant's head. An abnormally large or abnormally small head may be indicative of a clinical problem. The nurse taking the measurements has to be given criteria, a lower and an upper limit. If the infant's head is outside of these limits they should be referred for more specialist tests. National statistics are collected to determine the mean and standard deviation of head circumference for a range of ages. Let us say that at a particular age the mean head circumference is 50 cm and the standard deviation is 5 cm. Setting a lower limit of 45 cm and an upper limit of 55 cm would result in 31.7% (100-68.26) of the infants measured being referred for further tests. Setting a lower limit of 40 cm and an upper limit of 60 cm would result in 4.56% (100-95.44) being referred. In this way the normal distribution function can be used to set criteria which will result in the desired proportion of the population being referred for further testing.

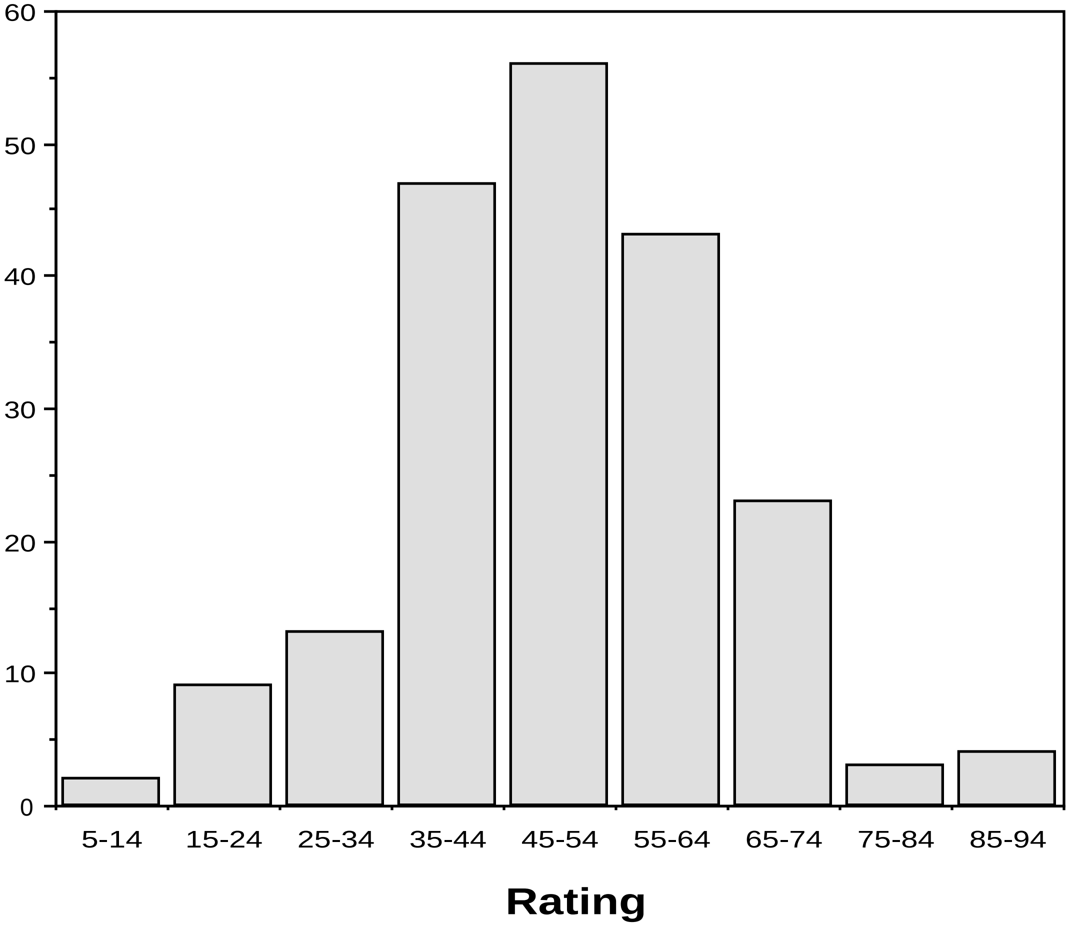


Figure 6.1 Frequency distribution based on the data in Table 6.1. Even though it is slightly asymetric, this approximates to a normal distribution.

Parametric tests

To recap, the objective is to evaluate the difference between two means with respect to the inherent 'error' variation in the scores. The null hypothesis is that the two sets of scores that the means were computed from both came from the same population of scores, i.e., the manipulation had no effect. We need to compute the probability that the two means could differ as much as they do, or more, assuming that null hypothesis is true. We hope to be able to reject the null hypothesis because that probability is very small. To compute a probability or 'p value' further assumptions have to be made. Statistical tests that assume the scores all come from the same normal distribution are known, for somewhat obscure reasons, as parametric tests. The F ratio computed to evaluate the significance of a regression equation in Chapter 5 was a parametric test. The parametric test for comparing means used in this Chapter is the t-test.

The t-test makes the assumption that all the scores come from the same normal distribution. Another theoretical distribution, known as Student's t distribution, is then used to compute a p value. Student's t distribution is derived from the normal distribution. Applying this test to Team A's results, Excel gives a p value of .0007 This is less than .05 so the difference is significant, the mean rating for the butter group is significantly higher than that for the margarine group. Were the experiment to be repeated, they might not get precisely the same means, but they are likely to come to the same conclusion. The same is not true of Team B's results. The p value computed by Excel for these data is .31 which is greater than .05. We cannot reject the null hypothesis. On the basis of Team B's data we cannot tell whether tasting butter or margarine gives the highest ratings. It is very likely that were the experiment to be repeated we could come to the opposite conclusion to that suggested by the present data.

When the assumptions are violated

With some kinds of data the assumption that the scores come from a normal distribution may be hard to justify. If this is the case then the validity of the p value computed using a parametric test may not be trustworthy. Reaction time is a classic example. With small measurements of time the underlying distribution is not normal. This is because it is generally easier to take a relatively long time to make a response than to take a relatively short time. In the limit it is impossible to respond in anything much less than about 200 ms.

Figure 6.2 plots some real data from a reaction time task, the two graphs are for two experimental conditions, the first results in generally faster reaction times. In both cases there is a clearly defined modal value but the distributions are not symmetrical about this value. The distribution extends further to the right of the modal value than to the left. As is often the case the decrease in modal reaction time is accompanied by a corresponding decrease in variance, that is the spread of scores in Figure 6.2a is much less than in Figure 6.2b. This is another sign that the underlying distribution is not normal because it is generally easier to increase than to decrease reaction time.

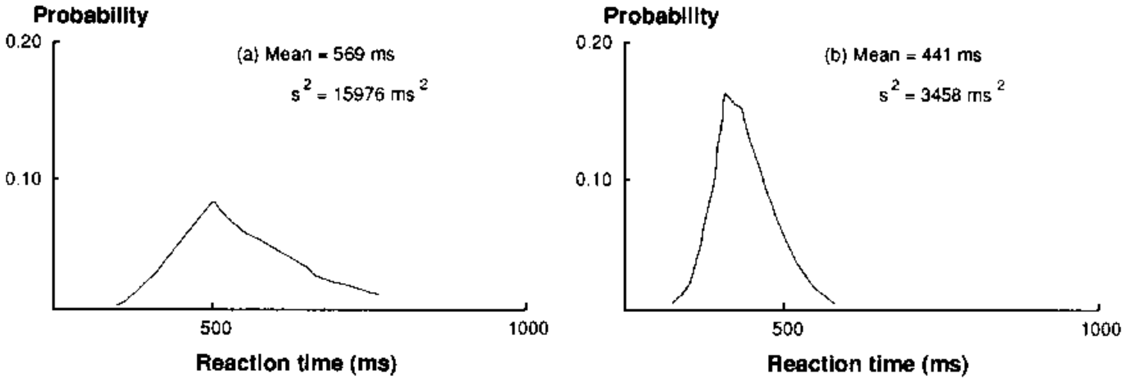


Figure 6.2 Distributions of reaction time under two conditions giving different means (real data—each graph is based on about 600 observations)

Were smaller samples to be drawn from these two distributions and their means compared in a t-test one might be sceptical about the validity of the p value computed. The null hypothesis for a t-test is that the two samples were drawn from the same normal distribution. The asymmetrical shapes of these distributions and the change in variance which accompanies the change in mean suggest this is not the case. The solution to this problem is to 'transform' the data, i.e., to translate each score into some other more suitable scale. With small reaction times it is common to take the logarithm of the time as the dependent variable rather than the time itself. Log reaction time has been found to be reasonably normally distributed. The approach adopted in this book is transform the scores to ranks. The ranks of ten scores will always be the numbers one to ten (assuming there are no tied ranks) and so fewer assumptions have to be made when computing a p value.

Statistics using ranks

Consider the following data from an experiment in which the dependent variable is the number of typing errors made in creating a document. The participants are instructed to avoid errors at all costs and so the number of errors made is in general small. The experiment is to determine whether music, played through headphones, can improve performance by cutting out the normal auditory distractions present in an open plan office.

Experimental Group (music) 5, 2, 23, 35, 7, 15, 10, 11 (Median = 10.5)

Control Group (no music) 8, 65, 83, 21, 18, 20, 43, 27, 12 (Median = 21)

Table 6.2 Error scores on a typing task

Like small times, small frequencies will also generally violate the assumptions of parametric tests and so the scores are transformed to ranks.

Score	2	5	7	8	10	11	12	15	18	20	21	23	27	35	43	65	83
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Group	E	E	E	C	E	E	E	C	C	C	C	E	C	E	C	C	C

Table 6.3 Ranking the scores, E indicates a score from the Experimental Group, C a score from the Control Group

The smallest score in the whole data set comes from the experimental group so that is assigned a rank of 1. The next smallest is 5 and so that is assigned a rank of 2, 7 gets a rank of 3 and 8, from the control group, a rank of 4. This continues up to the score 83 which is the largest in the data set and so that has a rank of 17, there being 17 participants altogether. The transformed scores can be gathered together again as follows:

Experimental Group (music) 2, 1, 12, 14, 3, 8, 5, 6 (Mean rank = 6.375)

Control Group (no music) 4, 16, 17, 11, 9, 10, 15, 13, 7 (Mean rank = 11.333)

Table 6.4 Ranks and mean ranks for each group

A difference between mean ranks can be evaluated with the Mann-Whitney U test which tests the degree to which the two samples overlap. In these data the higher ranks tend to be from the control group and the lower ones from the experimental group but there is some overlap. Minitab computes a statistic U which reflects the degree of overlap and then computes the probability that an overlap as small as that could have occurred by chance i.e., if the participants had been randomly assigned ranks. This requires making assumptions about the data but they are generally less demanding assumptions than those required for parametric tests. For the above data p is .0485 and so we can reject the null hypothesis. The difference in mean rank is significant at the .05 level.

When to use the t-test and when the Mann-Whitney U test

You have just collected some data from two groups of participants. There is an apparent difference in the mean scores for these two sets of scores and you need a statistical test to determine whether it is significant. How do you decide whether to use the t-test or the Mann-Whitney U test?

The 'power efficiency' of a statistical test refers to the number of participants which have to be tested in order to gather sufficient evidence to reject the null hypothesis. The statistical theory is that your two sets of scores are samples from two populations of scores. There will be a minimum sample size which is capable of demonstrating some given difference in the means of these two populations of scores. Intuitively one can see that a sample of three individuals from each population is unlikely to provide sufficient evidence to reject the null hypothesis that the scores came from the same distribution. At the other extreme, a sample of 100 individuals from each population ought to be sufficient to demonstrate any difference of practical importance. The minimum needed for a particular statistical test to demonstrate a real effect of some particular size can be estimated and will lie between these two extremes. In general, parametric tests have a greater power efficiency than statistics based on ranks, that is they require less data to reject the null hypothesis if there really is an effect. The reason for this is that in transforming the scores to ranks some information is lost. In Table 6.3 the lowest three scores all come from the experimental group. The lowest control group score is 8 so the mean ranks would not be changed if the number of errors made by these participants were 0, 1 and 2 or 5, 6 and 7 instead of 2, 5 and 7. The ranks do not distinguish between these different scores, the information they provide is lost.

The disadvantage of the t-test is that it makes more stringent assumptions. If the assumptions used to compute a p value are violated then we can no longer believe that p value. For the t-test the basic assumption is that all the scores come from the same normal distribution. It has already been noted that small measurements of time and small frequencies, such as error counts, may violate this assumption. Proportions suffer the same problems. It is likely to be harder to improve one's score from say 95% to 99% than from say 45% to 49%. As with small times this distorts the distribution. Proportions of one kind and another are the most common form of data collected by behavioural scientists. Any test or performance measure, where each participant gets some score between zero and a known maximum can be expressed as a proportion.

So, statistical theory recommends the use of the t-test because it has a greater power efficiency than the Mann-Whitney U but cautions against it because much of the data one is likely to collect will not be normally distributed. In practice both tests almost always give the same answer. That is, if one test shows that a difference is significant so will the other and if one test does not give a p value sufficiently small to reject the null hypothesis, neither will the other. There are two reasons for this. Firstly, it turns out that the power efficiency of the Mann-Whitney U is only slightly less than that of the t-test. Secondly, the t-test has been shown to be 'robust to violations of its assumptions'. If there are equal numbers of participants in each of the two groups and the distribution is not seriously skewed then the p value obtained can be trusted even if the data are to some extent non-normal.

The reader who is looking for practical advice about when to use the t-test and when to use the Mann-Whitney U test may be getting impatient with this section by now. The data collected in a single small scale experiment are unlikely to be sufficient to distinguish between a normal and a non-normal distribution. Anyway, 'small' deviations from normality are known not to be a problem. Deciding which test to use on the basis of the statistical theory is thus not easy to do. It is possible to suggest guidelines or rules of thumb which will be acceptable to most statistical authorities. In general parametric tests, such as the t-test, should not be used

1. If there is evidence for serious skew in the distribution (see Chapter 2). The signs of this are

(a) extreme outliers, perhaps scores greater than three standard deviations from the group mean;

(b) the range of score in one group is very much larger than (say more than twice) that in the other;

(c) there are ceiling or floor effects;

2. If there is not the same number of participants in one group as the other.

A floor effect will occur when a test is too hard. If many (say more than 20%) of the participants score close to zero then there is probably a floor effect. This will make the range of scores in whatever condition raises scores larger than that in the other. For this reason 1(a) and 1(c) usually go together. A ceiling effect is the same problem at the other end of the scale. A ceiling effect will occur when the test is too easy. Ceiling and floor effects may also make it harder to demonstrate any effect of the manipulation. There should be an equal number of participants in each group if you are using a t-test because the robustness of the test against violations of its assumptions depends crucially on this and very few variables used in the behavioural sciences can be trusted to be truly normal.

Given the high power efficiency of the Mann-Whitney U test a practical and very reasonable strategy is to always use it in place of the t-test. However, the t-test has a venerable history, it would seem churlish to abandon it except in the above circumstances!

How to report the results of the t-test and the Mann-Whitney U test

As stated in Chapter 5, when writing up the results of statistical tests it is important to give the right amount of detail. Part calculations should not be presented nor should basic concepts such as significance be explained. On the other hand sufficient detail is required to show that the appropriate test has been performed. A statistical test is used to evaluate a result so the first step should be to report that result. In this Chapter the results to be evaluated are differences between means so the first thing to be reported should always be the means and the conclusion they suggest. Having done this the results of the test used to assess the significance of the effect are reported.

Some statistics books discuss the use of one- or two-tailed t-tests. In the behavioural sciences the use of one-tailed tests can rarely be justified (see Chapter 5). For this reason Excel computes two-tailed probabilities. The t-test we have considered here is a 'two-tailed between subjects t-test'. The meaning of 'between subjects' will be explained in the next Chapter.

Team A's results might be written up as follows.

> Results:-
>
> The marks made by the participants on the rating scale were turned into scores between 0 and 100 by measuring the distance from the 'very nasty' end. Table 6.5 gives the mean rating and standard deviation for each group. The mean rating for the butter group is higher than that for the margarine group indicating that butter tends to result in higher ratings. This difference can be shown to be significant at the .05 level using a two-tailed between subjects t-test (t = 3.82, p = .0007).

	Mean	Std.Dev.
Butter Group	44.07	8.99
Margarine Group	32.00	8.32

Table 6.5 Results of tasting experiment, rating.

Note that summary statistics are presented in the report and not the raw scores. Large tables of numbers are hard for a reader to digest and best separated from the text, perhaps in an appendix.

A Mann-Whitney U test compares ranks not raw scores. The dependent variable being examined is now the rank and not the raw score so it is important that the mean rank for each group is reported along with the results of the test. Also, if the guidelines suggested above are followed then data which require a Mann-Whitney U test, rather than a t-test, are likely to come from skewed distributions. This being the case the median will be a more representative measure of central tendency than the mean (see Chapter 2). In the example which follows the mean, median and mean rank are all reported but the median is foregrounded as the 'prize result'.

Results:-

The number of errors made in typing the test document was determined for each participant. Table 6.6 gives the median, mean and standard deviation for the experimental group (with music) and the control group (no music). The median error score is higher for the control group suggesting that music may have had a beneficial effect. Raw error scores were transformed to ranks in order to perform a Mann-Whitney U test. The mean ranks differ in the same direction as the medians (see Table 6.6) and the effect was found to be significant at the .05 level (p = .0485).

	Median	Mean	Std.Dev.	Mean Rank
Experimental Group (music)	10.5	13.50	10.82	6.375
Control Group	21	33.00	25.65	11.333

Table 6.6 Results of the typing test, number of errors in document.

Both the above examples were significant effects. When reporting a non-significant result a similar format can be used. Note that not being able to reject the null hypothesis is not the same thing as concluding that there is no effect. The independent variable really may not have any effect but it is also possible that the experiment was not sufficiently sensitive. Team B's results might be reported as follows.

Results:-

The marks made by the participants on the rating scale were turned into scores between 0 and 100 by measuring the distance from the 'very nasty' end. Table 6.7 gives the mean rating and standard deviation for each group. Although there is an apparent difference in favour of the margarine group this was not found to be significant at the .05 level when evaluated with a two-tailed between subjects t-test (t = 1.04, p = .31).

	Mean	Std.Dev.
Butter Group	39.7	14.4
Margarine Group	47.6	19.1

Table 6.7 Results of the tasting experiment, ratings (Team B)

Summary:

1. This Chapter is concerned with evaluating experiments where there are two groups of participants. Each participant gives us one score and the conclusions drawn depend on the difference in mean score for each group (other types of experiment will be discussed in the next Chapter).

2. The scores vary within each group due to random error of measurement. There is thus the possibility that the difference in mean score, between the groups, is due to chance. If the difference between the means is due to chance then the result is probably not repeatable i.e., if the experiment were to be repeated we might draw the opposite conclusion.

3. The null hypothesis (what we mean by 'chance') is that all the scores were drawn from the same population. Their division into groups is essentially random.

4. By making assumptions about this population of scores we can compute the probability of getting two groups which differ in their measure of central tendency as much as was observed in the experiment. If this probability, or 'p value', is small (less than or equal to .05) then the null hypothesis is rejected and the result is said to be significant.

5. Two statistical tests are considered, each uses different assumptions in order to compute a p value.

(a) the t-test assumes that the scores come from a normal distribution (parametric test)

(b) the Mann-Whitney U test makes less demanding assumptions (non-parametric test). The scores are ranked and the computation is based on the degree to which the rank of participants in the two groups overlap.

6. The parametric test (t-test) will generally require less data in order to demonstrate the significance of an effect of some given size, however, calculations show that this difference in 'statistical power efficiency' is small. The t-test is robust to violations of its assumptions so long as there is an equal number of participants in each group and the underlying distribution is not seriously skewed. As a rule of thumb the Mann-Whitney U test should be used if:

(a) there are more participants in one group than the other

(b) there are extreme outliers (scores greater than three standard deviations from the group mean)

(c) the range of scores in one group is very much larger than, say more than twice, that in the other

(d) there are large ceiling or floor effects.

7. When reporting the results of one of these tests, first describe the result being assessed e.g., the means to be compared. A t-test is reported as a t value and a p value. Mean ranks should be reported with the p value for the Mann-Whitney U test. If the Mann-Whitney U test is being used because it is suspected that the data distribution is skewed then medians may be more representative measures of central tendency than the group means.

Chapter 7

Two experimental designs

Statistical concepts introduced in this Chapter:

Within subjects designs (alias: repeated measures, correlated or matched samples), between subjects designs (alias: independent samples, completely random), within subjects t-test, Wilcoxon matched-pairs signed-ranks test, Chi Square, choosing the appropriate test.

Within and between subjects designs

The previous Chapter described two statistical tests for evaluating a difference in central tendency between two groups. Each participant experiences only one of the levels of the independent variable and so contributes just one score to the analysis. This way of arranging an experiment will be described in this book as a *between subjects design*. It is so called because the experimental manipulation distinguishes between participants, either they are in the butter group or they are in the margarine group. The alternative to a between subjects design is a *within subjects design*. Here each participant experiences both levels of the independent variable and so contributes two scores to the analysis. It is called a within subjects design because the manipulation is made within the experience of each participant. Some examples of these two experimental designs will make this distinction clear. The first (Box 7.1) was used in the last two Chapters. Compare that with the within subjects design described in Box 7.2.

Box 7.1 A between subjects design

Hypothesis: Participants tasting butter will rate it more highly on a scale from 'very nasty' to 'very nice' than participants tasting margarine.

Experimental design: Between subjects, 15 participants taste a sample of butter and another 15 a sample of margarine. Each participant rates the taste of the one sample. The independent variable is 'Fat Tasted' and the dependent variable the rating.

Data:

Butter group		Margarine Group	
S1	34	S2	32
S3	38	S4	31
S5	60	S6	19
S7	44	S8	38
S9	47	S10	34
S11	31	S12	41
S13	58	S14	40
S15	53	S16	27
S17	37	S18	28
S19	37	S20	19
S21	46	S22	51
S23	43	S24	27
S25	54	S26	30
S27	35	S28	34
S29	44	S30	29
Mean	44.07	Mean	32.0

Table 7.1 Rating data

Conclusion: The mean rating for the butter group is higher than the mean rating for the margarine group as predicted.

Null hypothesis for significance testing: the two groups are random samples from the same population (the independent variable 'type of fat' had no effect on the scores).

Box 7.2 A within subjects design

Hypothesis: Participants will respond faster to a red light than to a green light in a simple reaction time experiment.

Experimental design: within subjects, 10 participants press a response button when ever a light comes on. There are 200 trials. On half of these trials the light is red and on half it is green. The sequence of red and green lights is odd trials red and even green for half the participants and even trials red and odd green for the other half. The mean reaction time to the red light and the mean reaction time to the green light is computed for each participant. The independent variable is 'Colour of Stimulus' the dependent variable is mean reaction time.

Data:

	Red	Green	Difference
S1	632	644	12
S2	571	631	60
S3	472	514	42
S4	848	926	78
S5	567	681	114
S6	505	511	6
S7	572	638	66
S8	729	713	-16
S9	577	597	20
S10	770	752	-18
Mean	624.3	660.7	36.4

Table 7.2 Reaction time data

Conclusion: The red light is responded to faster than the green light as predicted.

Null hypothesis: The differences are sampled from a distribution with a mean of zero (the independent variable 'colour' has no effect on reaction time).

Note that in the within subjects design the scores are paired up and a difference can be computed. While the scores in the between subjects design are tabulated with the score for participant 1 next to the score for participant 2 and so on, this arrangement is completely arbitrary. There is no reason why the score for participant 1 should not have been tabulated next to the score for participant 16, say. It would make no sense to compare individual scores in these data.

Because it is possible to pair up the scores in a within subjects design and then compute a difference score, a different null hypothesis is called for and hence different statistical tests. For a within subjects design the null hypothesis is that the 'real' difference between the pairs of scores is zero, but, because of the random error present in all behavioural data this is not what is observed. More precisely the null hypothesis is that the differences are sampled from a hypothetical distribution which has a mean of zero. The test computes the probability that the mean difference for the sample could deviate this much from zero by chance. If this p value is less than .05 the null hypothesis is rejected and the difference is said to be significant.

Again there are parametric and non-parametric tests for computing p. The parametric test is a *within subjects t-test*. The non-parametric test, the equivalent of the Mann-Whitney U test, is the Wilcoxon matched-pairs signed-ranks test. These are described below. First the other names by which these two experimental designs are known will be discussed.

Other names

Unfortunately the authors of statistical text books and computer programs cannot agree on labels for these two experimental designs. In Minitab they are distinguished as *one-sample* or *two-sample* tests. The command **ttest** takes a single set of numbers (one sample) and tests whether their mean is significantly different from zero (the null hypothesis). If you start off with pairs of scores these have to be converted to differences before the test can be applied. In contrast the **twosample** command (**twos** for short), introduced in the last Chapter, is used with between subjects designs. As it is not possible to compute a difference score in this case, this command takes two columns of data (two samples). In this case the null hypothesis is that both samples are drawn from the same population. As these two t-tests use different null hypotheses they will give very different answers. The test used must be appropriate for the experimental design.

Between subjects designs are sometimes referred to as *independent-samples designs*, within subjects designs as *correlated-* or *matched-samples* designs. The labels within and between subjects are used here because they are useful for thinking about more complex experimental designs in behavioural science which you will encounter if you go on to consider other experimental designs outside of the scope of this book.

Matched samples designs

This term is used to describe within subjects designs where the sampling unit or 'subject' is a pair of matched individuals. For example, you might be interested in some rather rare disease of the nervous system. You can scrape together a group of 12 such individuals but they are a very heterogeneous group as they vary widely in age, IQ, and so on. The hypothesis is that they have a deficit in verbal reasoning. Simply comparing them with a sample of normal individuals is unsatisfactory as matching the mean age, IQ and so on may be difficult and the wide variation in reasoning ability due to age and IQ differences within the group will mean that the experiment will only be sensitive to very large differences between groups. One solution is to match each patient to a normal control of the same age and IQ. This will guarantee that the mean age and mean IQ for the normal participants are well matched to those for the patients. In addition, the data can be taken in pairs and difference scores computed. This could be thought of as a 'within pairs' design and could be evaluated with a within subjects t-test. The new sampling unit (normally subjects) is the pair.

t-test for within subjects designs

This t-test uses the null hypothesis that the differences come from a normal distribution with mean zero. This is a parametric test like the between subjects t-test considered in the previous Chapter. If this assumption of normality is seriously compromised because the distribution of differences is very skewed then the alternative Wilcoxon matched-pairs signed-ranks test described in the next section should be used. In a small sample outliers are the main evidence for skew. If one or two of the differences of one sign are very large (say more than three standard deviations from the mean difference) then the t-test should not be used.

The Wilcoxon matched-pairs signed-ranks test

To perform this test the differences are ranked, irrespective of sign. For example take the differences from the reaction time experiment above. Table 7.3 tabulates the absolute size of the differences from Table 7.2. The smallest difference is 6, this is given a rank of 1. The next smallest is 12 and the next 18 and so on. The largest absolute difference is 114. That is given a rank of 10.

Participant	1	2	3	4	5	6	7	8	9	10
Difference (ignoring size)	12	60	42	78	114	6	66	16	20	18
Rank	2	7	6	9	10	1	8	3	5	4
Sign	+	+	+	+	+	+	+	–	+	–

Table 7.3 Ranked absolute differences for the reaction time data

If there was no difference between the red and green conditions one would expect the sum of the ranks for the differences with a positive sign to be the same as the sum of the ranks of the differences with negative ranks. In these data, when sign is taken into account, there is a net positive difference (mean difference 36.4) and this can be seen in the rankings. There are more positive differences than negative and the negative differences tend to be smaller than the positive ones. The Wilcoxon test uses the null hypothesis that the ranks of the negative and positive differences are assigned at random to compute a p value. In this case it is .041 so the difference is significant at the .05 level.

The same arguments, developed in Chapter 6, that apply when choosing between the between subjects t-test and the Mann-Whitney U test also apply in choosing between a within subjects t-test and the Wilcoxon test. Any test based on ranks will require more evidence to reject the null hypothesis than an equivalent test using untransformed scores but the Wilcoxon matched-pairs signed-ranks test has a high power efficiency. It should always be used when there are signs of serious skew in the distribution of difference scores such as is signalled by the presence of extreme outliers.

Designing your own experiments

There are advantages and disadvantages for both of these experimental designs which you should be aware of when designing your own experiments or reading about those of others.

The advantage of a within subjects design is that a difference score may be a more sensitive dependent variable than either of the raw scores that are used to compute it. Reaction time for example is a variable with large individual differences. Participants differ from one another a great deal, although within a given participant's performance reaction time may be relatively stable.

Looking at the reaction time data in Table 7.2 it is clear that S2 is considerably better at this kind of task than S4. Nevertheless the difference between red and green reaction times is similar for the two participants. By computing a difference score we remove the error variance due to individual differences. In a between subjects design it is not possible to do this and so very many more participants may need to be tested in order to provide sufficient evidence to reject the null hypothesis.

The disadvantage of a within subjects design is that experiencing one experimental condition (one level of the independent variable) may affect the participant in some way so that the second experimental condition has a different effect than it would have had this not been the case. Take the following example. It has been found that instructions to form mental images linking the words in a list makes that list easier to learn. An experiment might be performed to demonstrate this. In one of the experimental conditions instructions are given to form visual images, in the other no such instructions are given. Consider a participant who has imagery instructions for the first list learnt and no instructions with the second. If imagery really makes the task easier then he is likely to use it with both lists. Even if he is instructed not to use imagery with second list the experience of learning the first list is likely to affect the way he approaches the task of learning the second. This carry over effect will decrease the size of the difference score. Alternatively he may perform the control condition (no instructions) first and then the experimental condition (instructions). This does not suffer from the above problem but if only participants who experience the conditions in this order are tested then any advantage observed could be due to practice effects. Practice would be expected to increase the recall of the second list learned anyway. However, it is conceivable that giving the instruction to image words changes the nature of the task so radically that any practice effect is cancelled. All in all it would be very difficult to interpret the results of either of these experiments.

Randomising the order in which different participants get the different conditions will control unwanted effects like practice but only if the effects are unchanged by the experimental manipulation. The reaction time example was chosen because it is practical to have a long sequence of red and green stimuli and to interleave red and green trials. A less satisfactory design would have been to perform the experiment as two blocks of trials, half the participants responding to 100 red trials and then 100 green and half doing it the other way round. By doing it this way it is hoped that averaging across the two groups of participants any practice effects will be cancelled. This depends on the assumption that the practice effect is the same for red and green stimuli. In this case there is no reason to believe this is not true but there might be some subtle effect of this kind. If the conditions are interleaved then there will be red and green trials at the beginning, middle and end of the experiment so differential order effects which could change the result would have to be extremely subtle.

The possibility of order effects of one kind and another has been used to argue that within subjects designs should not be used under any circumstances. A more pragmatic approach is to use them with dependent variables which make between subjects designs impractical because of the large number of participants required. Reaction time is nearly always used in within subjects designs. Every so often someone has to replicate an important result obtained in a within subjects design using a between subjects design. That way the investigators in a particular area of research can check that the assumptions they are all making in order to interpret the results of their within subjects experiments are justified.

Sampling bias and self selection

The major source of potential bias in a between subjects design comes in choosing the groups for the experiment. Groups must be randomly sampled from the same population. This requirement is generally upheld if participants are assigned arbitrarily to groups before they appear at the experiment. A good procedure is to use a rule of the form 'The first participant who turns up will be put into Group A the second into Group B and so on'. Since neither you or they know in advance whether they will turn out to be odd or even numbered participants this is an effective way of randomly assigning people to groups.

If participants choose the condition they perform in there may well be bias. Had the participants in the tasting experiment chosen whether they would like to be in the Butter or Margarine Groups the result might have been to select rather different types of people for the two groups. For example, the people who particularly dislike margarine might insist on being in the butter group.

A more subtle form of self selection may occur when participants have to be rejected after or while they are performing the experiment. Consider the following rather extreme example. It is proposed that punishment is a more effective motivation than reward. In a prior experiment the monetary reward required to get people to suffer a large electrical shock of a given size has been accurately determined so that the reward and the punishment for the present experiment are matched in subjective strength. All the participants have to learn a set of nonsense syllables. The reward group of ten participants are given an appropriate sum of money every time they get more than half the list correct. The punishment group, originally ten participants, are given the shock every time they get less than half the list correct. Five of the ten participants in the punishment group refuse to go on after they have received a few shocks, however, averaging the results from the remaining five participants gives a mean which is significantly better than that for the ten reward group participants.

The problem when interpreting this result is that participants have effectively selected themselves. Only the best participants will avoid the shocks and carry on to the conclusion of the experiment. It is not surprising that the mean for the five best participants in one sample is better than the mean of all of another sample. This is an extreme example but the same potential problem arises whenever participants are lost during or after an experiment and the loss could be attributed to the experimental condition experienced. Losses due to equipment failure, where this is not the case, can be replaced by running more participants without compromising the interpretation of the results.

A test for use with nominal data - Chi Square ($\chi^{2)}$

Most of this book is concerned with data in the form of scores. Each participant is assigned a number or numbers which quantify some aspect of behaviour. However, readers will remember that in Chapter 2 we considered summarising data in the form of a contingency table. Here participants do not provide scores to the analysis, rather the participants are put into categories. As an example consider a survey where the sample is classified according to whether they drive or not and whether they suffer from lower back pain. Table 7.4 shows how many people fell into the four categories: drivers with L.B.P, drivers without L.B.P., non-drivers with L.B.P. and non-drivers without L.B.P. In addition the table gives the total number of drivers and non-drivers (the column totals) and the total number of L.B.P. sufferers and non-sufferers (row totals).

	Drivers	Non-drivers	
Lower back pain	172	58	230
Not having L.B.P.	248	489	737
	420	547	967

Table 7.4 The association between lower back pain (L.B.P.) and driving (fictional data)

This is known as a two by two contingency table (see Chapter 2). The variables are both dichotomies, lower back pain or not and driving or not. There is clearly an association between these variables 41% of the drivers have lower back pain where as only 11% of the non-drivers do (these are fictional data!) If the incidence of lower back pain was independent of whether you drive or not then these proportions should be the same.

To illustrate this possibility Table 7.5 gives some data where there is no association between the variables. There are fewer males than females and there are fewer people with lower back pain than with it but the proportion of males suffering lower back pain is very similar to the proportion of females (35% and 37% respectively).

	Male	Female	
Lower back pain	35	64	99
Not having L.B.P.	86	148	234
	121	212	333

Table 7.5 The association between lower back pain (L.B.P.) and sex (fictional data)

The Chi square (χ^2) test evaluates the null hypothesis that the two dichotomies are independent by comparing the observed frequencies with what would be expected if this null hypothesis were true. If the null hypothesis were true then one would expect the frequencies in the cells of Table 7.4 (172, 58, 248, 489) to be determined by the row and column totals (230, 737, 420, 547). Thus as the proportion of drivers in the sample is 420/967 and the proportion of people having lower back pain 230/967 then the proportion expected to drive *and* have lower back pain is

$$\frac{420}{967} \text{ x } \frac{230}{967}$$

As there were 967 people in the sample then the expected frequency of such people is

$$967 \text{ x} \frac{420}{967} \text{ x } \frac{230}{967} = 99.90$$

The observed frequency, 172, is much higher than this expected frequency. More people who drive have lower back pain than would be expected (by chance). Table 7.6 gives the expected frequencies for all four cells of Table 7.4. As one would expect given the above finding for drivers with lower back pain, the observed frequency for lower back pain in non-drivers is lower than expected, for drivers not having lower back pain it is lower than expected and for non-drivers not having lower back pain it is more than expected.

	Drivers	Non-drivers
Lower back pain	99.9	130.1
Not having L.B.P.	320.1	416.9

Table 7.6 Expected frequencies for Table 7.4

	Drivers	Non-drivers
Lower back pain	36.0	63.0
Not having L.B.P.	85.0	149.0

Table 7.7 Expected frequencies for Table 7.5

Table 7.7 gives the expected frequencies for the observed frequencies in Table 7.5. For these data the observed and expected frequencies are very similar. There is no sign of an association between the sex of the participant and whether they suffer from lower back pain or not. The data conform closely to the null hypothesis for the Chi square test which is that the two dichotomies are independent.

Excel computes a Chi Square statistic by comparing expected and observed frequencies. For the data in Table 7.4 Excel gives a Chi Square of 120.714. For any 2 by 2 contingency table Chi Square has to be greater than 3.84 to reject the null hypothesis at the .05 level. This association between driving and lower back pain is thus significant at the .05 level. On the other hand the Chi Square statistic computed by Excel for the data in Table 7.5 is 0.059. This is less than the critical value of 3.84 so there is no evidence for a significant association in this case.

Chi square can be computed for nominal variables where people are put into more than two categories (e.g., experienced drivers, novice drivers and non-drivers). It is much more common for the variables concerned to be dichotomies as in the examples above. There are also problems interpreting a significant Chi Square when the variables are not dichotomies. This book only considers 'two by two' (or '2 x 2') Chi Squares tests i.e., Chi Square when applied to a contingency table based on two dichotomies.

N.B. for technical reasons a two by two Chi square test can only be applied when all of the expected frequencies are five or more. This means there must be *at least* twenty participants.

Reporting the results of a within subjects t-test and the Wilcoxon test

The within subjects t-test is used to evaluate a mean difference and so the means for the two experimental conditions being compared should be reported first. The difference between the means is equal to the mean difference and so it is not necessary to report the latter statistic. Having described the means you can then quote the t statistic and the p value. You should always make it clear what kind of t-test, i.e., within or between subjects, has been performed and that you have used a two-tailed test (see Chapter 5). The reaction time experiment might be presented as follows.

> Results:-
>
> The mean reaction time to the red lights and the mean reaction time to the green lights was computed for each participant. The means and standard deviations of these scores are given in Table 7.8. Averaging across participants, the mean reaction time to the red light is less than that to the green light. This difference can be shown to be significant at the .05 level using a two-tailed within subjects t-test (t= 2.67, p= .026). (While reaction time can produce skewed distributions there is no evidence of outliers in the difference scores and so a parametric test can be justified for these data).

	Mean	Std. Dev.
Red	624.3	120.6
Green	660.7	120.9

Table 7.8 Reaction times to red and green lights.

The Wilcoxon test is similarly reported. If the guidelines suggested above are followed then the data that require the use of a Wilcoxon test, as opposed to a within subjects t-test, are likely to come from a skewed distribution. In such cases the median may be a more representative measure of central tendency than the mean. It is good practice to quote the means and standard deviations as well as the medians for the two sets of scores. The Wilcoxon Statistic is written as 'T' (note that this is a capital, 'T', where as the t-test uses a lower case 't'). This is the sum of the ranks for the differences with the less frequent sign. It is conventional to quote this statistic along with the p value. The full name 'Wilcoxon matched-pair signed-rank test' should be used. Wilcoxon had to do with developing the Mann-Whitney U test and so some people refer to that test as a Wilcoxon test also.

If the reaction time experiment had been evaluated with a Wilcoxon matched-pair signed-rank test the results might be reported so.

> Results:-
>
> The mean reaction time to the red lights and the mean reaction time to the green lights was computed for each participant. The means, medians and standard deviations of these scores are given in Table 7.9. Averaging across participants, the median reaction time to the red light is less than that to the green light. This difference can be shown to be significant at the .05 level using a Wilcoxon matched-pairs signed-ranks test (T = 48, p = .041).

	Mean	Median	Std. Dev.
Red	624.3	575	120.6
Green	660.7	641	120.9

Table 7.9 Reaction times to red and green lights.

Reporting the results of a Chi Square test

The Chi Square test is used to evaluate an association in a contingency table. The contingency table should be reported first. It is often useful to turn the frequencies into percentages. For example, in Table 7.4 there are more non-drivers than drivers and many more people without lower back pain than with it. This makes it difficult to judge the extent of the association. In many cases one can think of one of the two dichotomies as being a dependent variable and the other an independent variable. In this example having or not having lower back pain can be thought of as a dependent variable (the result) and driving or not driving as the independent variable as it forms the contrast of interest. In such cases it makes sense to report proportions within each of the categories corresponding to the independent variable. Thus here the proportion of drivers suffering lower back pain and the proportion of non-drivers suffering lower back pain are reported. Having described the contingency table Chi square and significance can be quoted. An example is given below.

> Results:-
>
> The total number of participants falling into the four categories, drivers with lower back pain, drivers without lower back pain, non-drivers with lower back pain and non-drivers without lower back pain are given in Table 7.4 (see above). The proportion of the drivers having lower back pain and the proportion of the non-drivers having back pain are 44 and 11% respectively. This association between lower back pain and driving can be shown to be significant at the .05 level by a Chi Square test, Chi Square being 120.714.

Choosing a test - Figure 7.1

This far in this book we have considered eight statistical tests. Figure 7.1 should allow you do determine which one is applicable to a particular set of data. To use it first determine what sort of data you are dealing with. Data of the type labeled A in Figure 7.1 have an independent variable with two levels (conditions or groups), and a dependent variable which is a score of some kind. The alternatives are Type B, where a correlation is to be computed (no levels), and Type C, where a Chi square or binomial test is to be applied. In the latter case the participants are classified rather than given a score as such.

A common mistake is try to apply a Chi square test to other data where counts are obtained (e.g., the number of errors made by a participant). These are scores, so the data are of type A.

One then follows lines in the figure to determine the precise test to be used. In the case of Type A this involves deciding whether the design is a within or between subjects and then whether a t-test is appropriate. The box containing a general step by step procedure to be used when applying the test is referenced in the figure.

It should be noted that the tests described in this book, and thus covered in Figure 7.1, do not exhaust the potential experimental designs and experimental hypotheses which can be statistical evaluated. In particular we have not considered experimental designs where there are more than one independent variable (factorial designs), independent variables with more than two levels or nominal variables which are not dichotomies. These tests will however be sufficient for the majority of experiments. The experimental designs to which they are applicable have the considerable advantage of being straightforward to interpret. When designing your own experiments there is a lot to be said for restricting yourself to these designs, even if you do know how to analyse the alternatives.

Summary:

1. In the previous Chapter we considered a design for experiments where each participant contributes one score to the data. Two groups are compared on their mean scores. An alternative design for an experiment is to have each participant provide two scores, one for each experimental condition. The former arrangement is called a between subjects design, the latter a within subjects design.

2. Between subjects designs are also sometimes called independent or unrelated samples designs. Within subjects designs are sometimes known as repeated measures, correlated samples or matched samples designs.

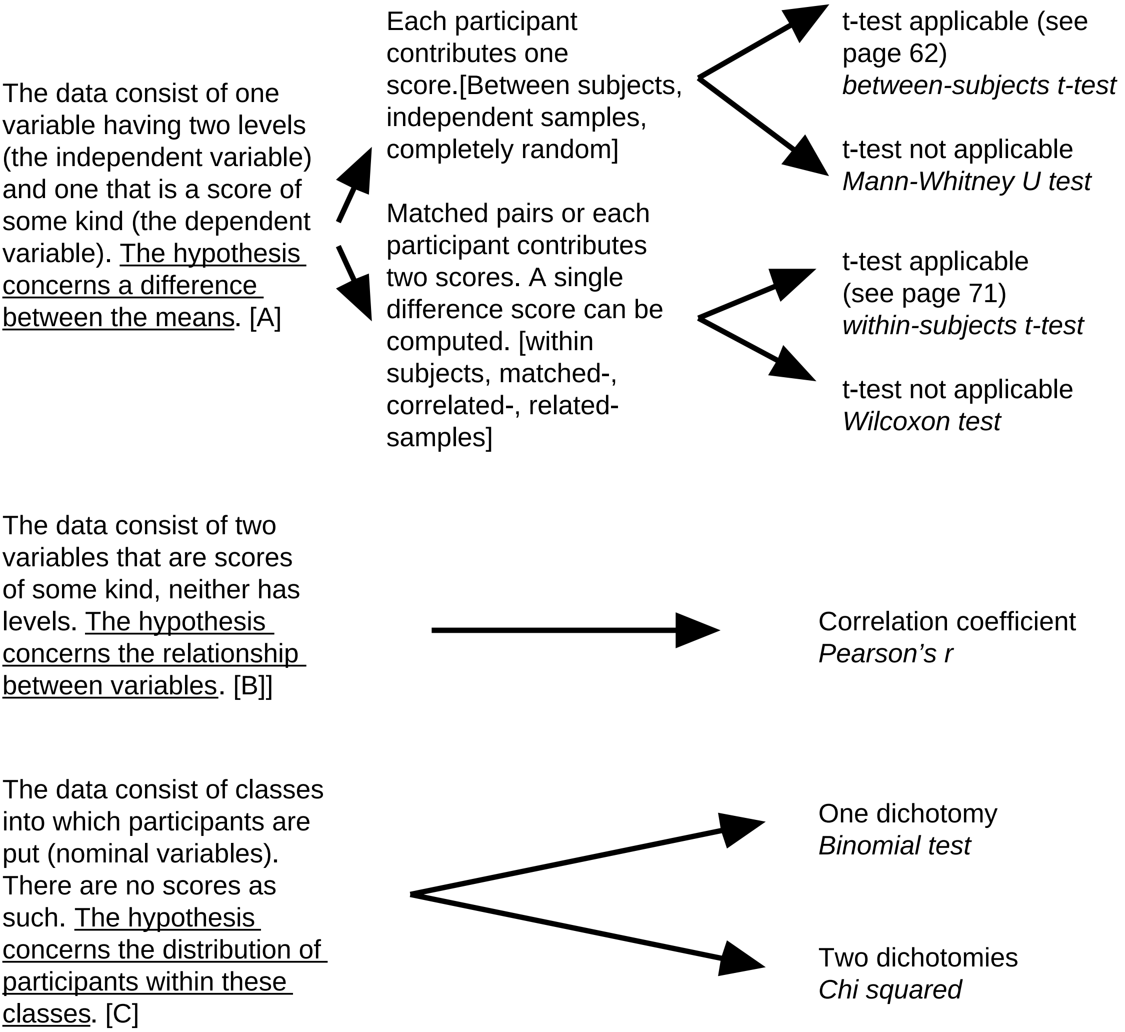
3. Different experimental designs require different statistical tests. For testing the significance of the difference between means in a between subjects design the null hypothesis is that all the scores come from the same distribution. In the case of a within subjects design it is that the difference between the scores has a true mean value of zero.

4. The independent samples t-test and the Mann-Whitney U test considered in the last Chapter are used for between subjects designs, the former when the assumptions underlying parametric tests are met, the latter when they are not. The within subjects t-test and the Wilcoxon matched-pairs signed-ranks test is used with within subjects designs, again the former when the assumptions underlying parametric tests are met, the latter when they are not.

5. The third test to be introduced in this Chapter was the Chi Square test for association in a two by two contingency table. The null hypothesis for this test is that the marginal totals, i.e., the distribution of values in each dichotomy, independently determine the cell totals.

6. Figure 7.1 summarises how to recognise data for which one of the eight statistical tests considered in this book is applicable. The first step is to

determine whether one is: (A) comparing means i.e., there is an independent variable with 2 levels and a dependent variable which is a score; (B) testing a relationship between continuous variables or (C) looking at nominal data i.e., participants who have been categorised in some way but not assigned scores. Further distinctions are made in order to arrive at the particular test required.



The data consist of one variable having two levels (the independent variable) and one that is a score of some kind (the dependent variable). The hypothesis concerns a difference between the means. [A]

Each participant contributes one score.[Between subjects, independent samples, completely random]

t-test applicable (see page 62) *between-subjects t-test*

t-test not applicable *Mann-Whitney U test*

Matched pairs or each participant contributes two scores. A single difference score can be computed. [within subjects, matched-, correlated-, related-samples]

t-test applicable (see page 71) *within-subjects t-test*

t-test not applicable *Wilcoxon test*

The data consist of two variables that are scores of some kind, neither has levels. The hypothesis concerns the relationship between variables. [B]]

Correlation coefficient *Pearson's r*

The data consist of classes into which participants are put (nominal variables). There are no scores as such. The hypothesis concerns the distribution of participants within these classes. [C]

One dichotomy *Binomial test*

Two dichotomies *Chi squared*

Figure 7.1 How to choose a test

Appendix A

Table A.1 Critical values of r

If the observed correlation is equal to or greater than the critical value it is significant at the .05 level (two-tailed).

N is the number of participants, subtract 1 from N when assessing the significance of a partial correlation.

If there is no value of N that corresponds exactly take the next *lower* N. Since this gives a critical r larger than the one you would have used had the table been more detailed this has to result in a safe decision. e.g.

r = .41, N = 65; next lower N in table is 62; $r_{critical}$ = .250.

N	$r_{critical}$
3	.997
4	.950
5	.878
6	.811
7	.754
8	.707
9	.666
10	.632
11	.602
12	.576
13	.553
14	.532
15	.514
16	.497
17	.482
18	.468
19	.456
20	.444
21	.433
22	.423
23	.413
24	.404
25	.396
26	.388
27	.381
28	.374
29	.367
30	.361
31	.355
32	.349
37	.325
42	.304
47	.288
52	.273
62	.250
72	.232
82	.217
92	.205
102	.195

Appendix B

Multiple and stepwise regression

Statistical concepts introduced in this chapter:

Multiple regression, R, stepwise regression, partial correlation

Introduction

Multiple Regression - the problem

Chapter 4 introduced simple regression and the correlation coefficient, r, to describe the relationship between two variables. Very often there are more than two variables to correlate with one another. In previous chapters we have dealt with this by computing a correlation matrix. Every variable is correlated with every other variable. Table 8.1 for example, shows high correlations between exam score and three other variables. There are also moderate correlations between Scale A and IQ and Scale A and reading age.

	exam	Scale A	IQ
Scale A	.76		
IQ	.77	.59	
Reading age (RA)	.74	.64	.39

Table 8.1 A correlation matrix

Correlation matrices are not always easy to interpret and various statistical techniques for characterising them have evolved. This chapter describes one such technique. Multiple regression is used to ask questions about the relationship between some single variable and some combination of variables. For example, one might use multiple regression to ask whether combining Scale A, IQ and Reading age gives a better prediction of exam scores than Scale A alone, or, if one were to choose just two of those three in combination which would be the best two to choose.

We shall describe the single variable (exam scores) as the predicted variable and the variables used in combination (Scale A, IQ and Reading age) as the predictor variables. As with simple regression, it is very unusual for anyone to want predict individual scores using a regression equation. We ask what the best prediction might be in order to determine the strength of the relationship. In that sense the predicted variable is like the dependent variable in an experiment where two groups or condition are compared. The dependent variable is only predicted in that it reflects the result of the experiment. In the same way the predictor variables are equivalent to the independent variable in such an experiment. An independent variable is a predictor in that it is expected to affect the dependent variable.

Simple regression

Before describing multiple regression, simple regression will be quickly reviewed. Simple regression is based on the linear equation y = c + mx

In Chapter 4 the example used was,

predicted_exam_score = 55.4 + 1.55 Scale_A_Score

This is the best fitting linear equation of this form, where best fitting is defined as the equation with the smallest deviance. The deviance of an equation is computed by finding the squared difference between the predicted and observed exam score for each subject and then summing these squared differences. Deviance was also used to define the correlation coefficient r.

$$r^2 = \frac{\text{deviance of the mean - deviance of the regression equation}}{\text{the deviance of the mean}}$$

One can think of the deviance of the mean as a baseline against which the deviance of the regression equation is measured. In moving from the equation

predicted_exam_score = 71.25

to the equation

predicted_exam_score = 55.4 + 1.55 Scale_A_Score

we reduce the deviance of the prediction from 834.25 to 353.59. r^2 is that improvement, expressed as a ratio of the baseline deviance.

$$r^2 = \frac{834.25 - 353.59}{834.25} = .58$$

Extending the linear equation

In multiple regression the equation is extended to include more than one predictor variable. For example, the best linear combination of IQ and Scale A for predicting exam scores might be determined. This is given by the formula

predicted_exam_score = 10.3 + 0.963 Scale_A_Score + 0.505 IQ

'Best' is again defined with regard to the deviance of the prediction. The first student's Scale A score is 11 and his IQ 104 so his predicted exam score is

10.3 + 0.963x11 + 0.505x104 = 73.413

His observed exam score is 76 so the deviation between observed and predicted is 2.587. As before, squaring the deviations for all the subjects and summing gives the deviance. Table 8.2 contains: Scale A and IQ scores; predicted and observed exam scores; deviations and squared deviations. The deviance is 222.21 which is better than the deviance when predicting with Scale A alone (353.58).

Again, as with simple regression, we can compute a correlation coefficient.

$$R^2 = \frac{\text{deviance of the mean - deviance of the regression equation}}{\text{the deviance of the mean}}$$

R is the multiple correlation coefficient and it is computed in exactly the same way as r.

$$R^2 = \frac{834.25 - 222.21}{834.25} = .734$$

Simple regression with Scale A alone gave an r^2 of .576 so the correlation is considerably improved by adding IQ to the regression equation.

The regr command in SPSS will compute multiple as well as single regression giving the best fitting equation, R^2 and the deviance of the regression. You will never have to do the computations shown in Table 8.2 except in artificial exercises of the kind found in this statistics book.

Subj.	Scale_A	IQ	Pred. exam score	Observed ex. sc.	Deviation	Sq.Dev
1	11	104	73.413	76	2.587	6.6926
2	8	99	67.999	60	-7.999	63.9840
3	11	98	70.383	65	-5.383	28.9767
4	9	93	65.932	61	-4.932	24.3246
5	7	113	74.106	74	-0.106	0.0112
6	17	100	77.171	79	1.828	3.3452
7	7	98	66.531	67	0.469	0.2200
8	18	119	87.729	89	1.271	1.6154
9	12	100	72.356	74	1.644	2.7027
10	5	93	62.080	62	-0.080	0.0064
11	13	107	76.854	78	1.146	1.3133
12	5	90	60.565	70	9.435	89.0192
			Sum of squared deviations or Deviance			222.21

Table 8.2 Computing the deviance of a multiple regression.

Stepwise regression - linear equations as models

The three linear equations considered for predicting exam scores can be considered as different models that are fitted to the data. The deviance of the equation is the 'error' left after the model is fitted. The simplest, let us call it Model 0, is that exam score can be predicted from a single constant. The mean exam score turned out to be the constant which gives the least deviance so Model 0 is

predicted_exam_score = 71.25

this has a deviance of 834.25. The next model to be fitted, Model 1, was that exam score can be predicted from Scale A. The best linear equation for doing this was

predicted_exam_score = 55.4 + 1.55 Scale_A_Score

and this has a deviance of 353.58. Finally Model 2 predicts exam score from Scale A and IQ

predicted_exam_score = 10.3 + 0.963 Scale_A_Score + 0.505 IQ

and this model has a deviance of 222.21. This is summarised in Table 8.3.

	Deviance
Model 2 (Regression of exam score on Scale A score and IQ)	222.21
Model 1 (Regression of exam score on Scale A alone)	353.58
Model 0 (Deviance of the mean exam score)	834.25

Table 8.3 Three linear models

Moving from Model 0 to Model 1 can be thought of as a 'step' in which Scale A is added to the equation. Similarly, moving from Model 1 to Model 2 can be thought of as a step in which IQ is added to the equation. R^2 evaluates Model 2 against Model 0 i.e., it says how good the prediction using Scale A and IQ together is compared with no predictors at all. It is also possible to evaluate Model 2 against Model 1. This says how much one has gained by adding IQ to an equation already containing Scale A. This is known as stepwise regression.

There are two basic ways of doing this stepwise regression. One is used when there is a hypothesis to test and the other as a way of describing the relative strengths of predictors when used in combination. We shall consider these two uses of stepwise regression after a discussion how the significance of a correlation is assessed. This will involve introducing the concept of degrees of freedom.

Degrees of freedom

Given that we are always looking for the best possible prediction it is not surprising that predicting with two things is better than predicting with one. Indeed it can be shown that a regression equation with N terms (N-1 predictors and a constant) can always be found to predict the scores of N subjects with zero deviance.

Each time a new term is added to the model the opportunity for prediction to fail is reduced. This notion is expressed in a parameter known as the 'degrees of freedom' (d.f.). The deviances computed for this example and their associated degrees of freedom are retabulated in Table 8.4.

	Deviance	d.f.
Model 2 (Regression of exam score on Scale A score and IQ)	222.21	9
Model 1 (Regression of exam score on Scale A alone)	353.58	10
Model 0 (Deviance of the mean exam score)	834.25	11

Table 8.4 Deviances and degrees of freedom (d.f.) for three models

Model 0, deviance of the mean, has 11 degrees of freedom, there being 12 subjects. Model 1 (Scale A only) has 10 degrees of freedom and Model 2 (Scale A and IQ) has 9. To make the deviances comparable one can express them as deviance per degree of freedom. This statistic is known as the mean square and can be thought of as a kind of variance (in the case of Model 0 it *is* the variance of the exam scores).

The SPSS regress command computes these deviances, degrees of freedom and mean squares and displays them as an 'Analysis of Variance' table. The tables produced by SPSS for the simple and multiple regressions respectively are given in Tables 8.5 and 8.6.

SOURCE	DF	SS	MS	F	p
Regression	1	480.67	480.67	13.59	0.004
Error	10	353.58	35.36		
Total	11	834.25			

Table 8.5 Analysis of variance, predictor Scale A (Model 1).

SOURCE	DF	SS	MS	F	p
Regression	2	612.04	306.02	12.39	0.003
Error	9	222.21	24.69		
Total	11	834.25			

Table 8.6 Analysis of variance, predictors Scale A and IQ (Model 2).

The column DF gives the degrees of freedom. Deviance (sum of the squared deviations) is abbreviated as SS for Sum of Squares and mean square is abbreviated as MS. The 'Error' SS is the deviance of the regression equation. This is because the deviance of the regression equation is the 'error' left when one has fitted the model. The 'Total' SS is the deviance of the mean. The SS labeled 'Regression' is the difference between the other two, that is the reduction in deviance. This may be slightly confusing. The authors of SPSS have labeled this row in the table 'Regression' as it may be thought of as the SS explained by the regression equation. It may be better to think of it as the 'reduction in deviance'.

Comparing the error MS we see that the deviance of Model 2 is smaller that that of Model 1, even when the smaller degrees of freedom of the former is accounted for (24.69 versus 35.36). However, comparing the Regression mean squares (306.02 versus 480.67) we see that the reduction in deviance gained by fitting a model which uses up 2 degrees of freedom is not twice as much as that gained by fitting a model which uses up one degree of freedom.

The F ratio

Tables 8.5 and 8.6 contain p values obtained from the F ratios given to their left. The F ratio is a statistic like t or Chi square with a known distribution from which a p value can be computed. F is in turn computed from two mean squares.

Where there is a p value there must be a null hypothesis! For the simple regression evaluated in Table 8.5 this is that Scale A score does not in fact predict exam scores. The scores are randomly scattered on the scattergram and the variability represented by the regression line is the same as the general variability in the scores.

For convenience the 'Error' MS will be written MS_{Error} the 'Regression' MS, $MS_{Regression}$. The two mean squares which go into the F ratio are viewed as independent estimates of the true variance due to the regression line ($MS_{Regression}$) and the true variance about the regression line (MS_{Error}). So the null hypothesis is that they are both independent estimates of the same random fluctuation. Now if the true $MS_{Regression}$ equals the true MS_{Error} then

$\frac{\text{true } MS_{\text{Regression}}}{\text{true } MS_{\text{Error}}}$ = 1 so F should be 1. However, they are only estimates of the true state of affairs. MS_{Error} may be a chance under estimate and $MS_{\text{Regression}}$ may be a chance over estimate. These possibilities are important because if Scale A did predict exam scores then $MS_{\text{Regression}}$ would be larger than MS_{Error} i.e., the alternative to the null hypothesis is that F is greater than 1.

By making assumptions about how the data arose statisticians, notably Fisher who gives his name to this statistic, have worked out how to compute the probability that an F greater than some value could occur by chance from two independent estimates of the same true variance. If that probability is small (less than .05) then we can reject the null hypothesis and accept that r is significant. The logic of using F to compute the significance of a regression equation is the same logic as was introduced in Chapter 5 for all statistical tests. It is summarised in Box 8.1.

Box 8.1 Statistical inference with the F ratio

(a) A chance result is inherently unreliable. If we can reject the possibility that a result is due to chance then it is probably a repeatable result.

(b) 'Chance' is defined as a null hypothesis, in this case that MS_{Error} and $MS_{\text{Regression}}$ are independent estimates of the same random variation.

(c) The probability of getting our particular result, or better, under the null hypothesis is computed. In this case the probability of getting an F ratio greater than or equal to 13.59 when the larger MS has d.f. 1 and the smaller d.f. 10.

(d) If that probability is less than or equal to the significance level the null hypothesis is rejected and the result is said to be significant. .004 is less than .05 so the correlation is significant.

In Chapter 5 the significance of r was interpreted in terms of the reliability of the slope of the regression. In the case of R this is not possible. The significance of R can only be interpreted by reference to deviance (SS). A significant R indicates that the deviance explained by the regression equation is unlikely to be due to chance.

Stepwise regression - statistical control

Having considered degrees of freedom and F ratios we can return to the practical analysis problems mentioned earlier. The first is the problem of assessing the effect of a 'nuisance variable' on a correlation. What this means is best explained by an example.

Table 8.7 is a correlation matrix derived from an experiment which measured the ability of children to recall lists of words (see Appendix B, Table B.4). These data were collected to see how recall changes with age.

	Recall	Age	Time
Age	.87		
Time	.05	.50	
IQ	-.11	-.15	-.33

Table 8.7 Correlation matrix for recall experiment

The experimental procedure used to collect these data was not entirely satisfactory and subjects vary in the amount of time they spent learning the list. There is a moderate correlation (.50) between age and time spent learning the list which indicates that the latter variable does not vary randomly with respect to age. Time spent learning the list can be described as a nuisance variable which covaries with the predictor variable we are interested (age). As older subjects tend to spend more time learning the list this alone could account for any apparent effect of age on recall. It would have been best to control this variable experimentally but failing this the effect of time spent learning can be 'statistically controlled'. The problem then is to assess the effect of age on recall after partialling out the effect of time spent learning on recall. In stepwise regression this is done as follows.

1. Compute the regression equation and deviance for the nuisance variable to be partialled out, in this case the regression equation for predicting recall from time spent learning the list (Model 1). This can be thought of as the new baseline deviance.

2. Add the predictor variable we are interested in to the equation i.e., compute the deviance for the regression equation for the nuisance variable plus the predictor variable. In this case this is the regression equation for predicting recall from time spent learning and age (Model 2).

3. The decrease in deviance in making the step between Models 1 and 2 indicates the strength of the relationship between the predictor and predicted variables when the nuisance variable is statistically controlled. When converted to a mean square, this reduction in deviance is the variance in the recall scores left for the predictor to account for when the variance attributable to the nuisance variable has been removed.

Predictor(s)	Deviance	d.f.	Dev. reduced
None (Model 0)	102.4	9	
Time (New baseline, Model 1)	102.18	8	0.22
Time & Age (Model 2)	3.505	7	98.675

Table 8.8 Stepwise regression (statistical control)

Table 8.8 recorded these computations including, for completeness, the step from no predictors (Model 0) to Model 1. Adding age to the equation considerably reduces the deviance. It is apparent that age still has plenty of predictive power even when the variance due to time taken to learn the list has been statistically controlled.

Partial correlation

Stepwise regression, when used in this way is equivalent to partial correlation. A partial correlation can be computed from the correlation matrix as follows. The correlation between recall and age will be written r_{ra} and that between time to learn the list and recall r_{tr} and so on. The partial correlation

$r_{ra.t}$ is the correlation between recall and age after partialling out the effect of time spent learning the list. This can be computed from the other correlations using the formula

$$r_{ra.t} = \frac{r_{ra} - r_{tr}\, r_{ta}}{\sqrt{(1-r_{tr}^2)(1-r_{ta}^2)}}$$

$$= \frac{.873 - .046x.501}{\sqrt{(1-.046^2)(1-.501^2)}}$$

$$= .983$$

$$r_{ra.t}^2 = .9657$$

$r_{ra.t}$ is in fact stronger than r_{ra} (.87). Controlling for time spent learning the list has improved not worsened the correlation between recall and age.

It is not immediately obvious how this formula works. However, the same figure can be computed from the stepwise regression depicted in Table 8.8 where the motivation behind the computation may be clearer. It will be recalled that to compute a correlation (r^2) the reduction in deviance in moving from Model 0 (no predictors) to Model 1 (one predictor) is expressed as a proportion of the deviance for Model 0. In terms of our example

$$r_{ra}^2 = \frac{102.4 - 24.381}{102.4} = .762$$

The partial correlation is computed similarly but using the new baseline deviance (Model 1) in place of Model 0. In terms of our example

$$r_{ra.t}^2 = \frac{\text{dev. of reg. on time - dev. of reg on time \& age}}{\text{dev. of regression on time}}$$

$$= \frac{102.18 - 3.501}{102.18}$$

$$= .9657$$

Significance testing in stepwise regression - statistical control

The significance of a partial correlation can be looked up in the Table A.1 if the N is reduced by one. It is preferable to test the significance of a partial correlation in an analysis of variance table. Table 8.9 does this for our example. Notice first that the partial correlation has a very high F ratio and so a very low p value. The correlation between recall and age is significant at the .05 level when time to learn the list is statistically controlled.

Table 8.9 is constructed in the same way as the analysis of variance table for the step from Model 0 to Model 1 except that the new baseline deviance is used instead of the deviance of Model 0. For comparison Table 8.10 gives the analysis of variance table for the correlation between recall and age.

.	df	SS	MS	F	p
Reduction in deviance	1	98.675	98.675	197.07	<.0001
Time & Age	7	3.505	.5007		
Time	8	102.18			

Table 8.9 An analysis of variance table, stepwise regression for statistical control

.	df	SS	MS	F	p
Reduction in deviance	1	78.019	78.019	25.60	<.0001
Age	8	24.381	3.048		
Model 0 (no predictors)	9	102.4			

Table 8.10 An analysis of variance table, correlation

In both tables the reduction in deviance SS and d.f. are computed by subtraction. Mean squares (MS) are computed by dividing SS by d.f. and F is the reduction in deviance MS divided by the other MS as in Table 8.5 above. Box 8.3 in the work sheets gives a step by step procedure for computing a partial correlation and constructing an analysis of variance table.

More complex hypotheses

In the above calculations stepwise regression was used to test a reasonably sophisticated psychological model of how recall scores relate to other variables. It is straightforward to generalise these calculations to more complex models with more than one nuisance variable and more than one predictor. For example, had there had also been a correlation between IQ and age indicating a bias in the sampling for say older children to be more intelligent then IQ might have been used as a second nuisance variable. The significance of the model is tested by evaluating the change in deviance from a baseline deviance as above. The only difference is that the baseline model contains more than one nuisance variable, in this case time and IQ.

Stepwise regression for statistical control is sometimes described as stepwise regression when one or more variables are 'forced' into the equation. This follows from the way the nuisance variables are entered into the equation first (forced in) before the real predictors are considered. This is in contrast to the procedure to be described below where variables enter the equation in an order which is statistically rather than psychologically determined.

Stepwise regression - descriptive use

Let us return to the example of predicting exam scores and expand it by considering a further potential predictor, Reading Age (RA). Examination of the correlation matrix in Table 8.1 shows that all three predictors, Scale A, IQ and Reading Age correlate with exam score but they also correlate with one another. Stepwise regression gives us a way of describing these correlations as they relate to the predicted variable exam score. In the above sections variables have been entered into the equation in an arbitrary order. Here the order they are entered depends on the amount of deviance reduced. Thus the order in which they enter is a way of describing their importance in predicting exam scores. The procedure is as follows.

1. Choose the predictor which correlates best with the predicted variable, this is said to be 'the first variable to enter the equation'. In this case it is IQ.

2. Compute the regression equation and deviance for the first variable to enter the equation in combination with each of the other predictors in turn. IQ and Scale A gives a deviance of 222.21, IQ and Reading Age 155.82.

3. Choose the combination giving the least deviance. The variable added is the second variable to enter the equation, in this case Reading Age.

This procedure is repeated until all the predictors are used up. In this case there is only Scale A left. It is said to enter the equation last. The order in which the variables enter reflects their importance as predictors when taken in combination with one another.

Predictor(s)	Deviance	d.f.	
None (Model 0)	834.25	11	Deviance from the mean
Scale A	353.58	10	
Reading Age	381.14	10	
IQ	**344.19**	**10**	**IQ enters first**
IQ & Scale A	222.21	9	
IQ & RA	**155.82**	**9**	**RA enters second**
IQ & RA & ScaleA	**138.30**	**8**	**Scale A enters last**

Table 8.11 Stepwise regression

The procedure, as applied to the example, is summarised in Table 8.11. The lines in bold are the 'steps' in the procedure. The initial baseline is no predictors (Model 0). Then three single predictor models are fitted. IQ gives the lowest deviance so the first step is to enter IQ into the equation (in the procedure described above this was done by selecting the highest correlation with exam scores from the correlation matrix, the two procedures are equivalent). Next, the two models with two predictors are considered. One of the predictors must be IQ as it has already 'entered' the equation. IQ with Reading Age gives the lowest deviance so Reading Age is entered next. There is only one model with three predictors so Scale A enters as step 3. We conclude that, when used in combination, the importance of the predictors can be ordered: IQ, Reading age then Scale A.

It is interesting to note that, although Scale A correlates with exam score better than Reading Age does (see Table 8.1), it enters the equation last. This is because of the strong correlation between Scale A and IQ. IQ enters first and 'uses up' all the variance which Scale A could explain.

As each variable entered into the equation the improvement in prediction (decrease in deviance) diminishes. The step None to IQ reduces the deviance by 490, the step from IQ to IQ & Reading Age 188 and the step from IQ &

Reading Age to using all three predictors only 17.5. The next section discusses the use of significance testing as a 'stopping rule' to allow us to determine the point where adding more terms to the equation no longer results in a real gain in predictive power.

The significance of a step in stepwise regression - descriptive use

As well as the deviance of each model from Table 8.11, Table 8.12 gives the reduction in deviance achieved by taking each step and the p value for it. Thus moving from no predictors to IQ only (step 1) reduces the deviance from 834.25 to 344.19. This has a probability of .004 of occurring under the null hypothesis. This step is significant at the .05 level.

Predictor(s)	Deviance	d.f.	Dev. reduced	p
None (Model 0)	834.25	11		
IQ	344.19	10	490.06	.004
IQ & Reading Age	155.82	9	188.37	.009
IQ & Reading Age & Scale A	138.30	8	17.52	.344

Table 8.12 Stepwise regression.

The reduction in moving from IQ to IQ and Reading Age (step 2) is 344.19 - 155.82 = 188.37 which has a p value of .009. So this step is significant at the .05 level too. Moving from IQ and Reading Age to predicting with all three variables (step 3) only reduces the deviance by 17.52. The p value for this step is considerably greater than .05. This step is not significant at the .05 level. It can be concluded that adding Scale A to the equation does not buy you significantly better predictive power and the 'best' prediction is obtained with IQ and Reading Age.

The analysis tables in which these p values are computed are constructed in the same way as the significance of any step in stepwise regression. Table 8.13 is the analysis of variance table for the first step from no predictors to the best single predictor, IQ. The F ratio computed evaluates the reduction in deviance, 490.06 against the deviance left after adding IQ to the model, 344.19.

	df	SS	MS	F	p
Reduction in deviance	1	490.06	490.06	14.24	.004
IQ (Model 1)	10	344.19	34.42		
None (Model 0)	11	834.25			

Table 8.13 Step 1 in predicting exam score, IQ enters first.

The table we construct to evaluate step 2 has the same form. The bottom line of Table 8.14 is the baseline deviance, this time the deviance of the one predictor model, IQ. The next line up in Table 8.14 gives the deviance of the current model, this time the deviance of the two predictor model, IQ with Reading Age. The reduction in deviance and its d.f. are computed by subtraction and the MS and F ratio in the normal way.

	df	SS	MS	F	p
Reduction in deviance	1	188.37	188.37	10.88	.009
IQ & RA (Model 2)	9	155.82	17.313		
IQ (Model 1)	10	344.19			

Table 8.14 Step 2 in predicting exam score, Reading Age enters second.

Table 8.15 is the analysis of variance table for step 3. Scale A enters last. Its baseline deviance is the deviance of Model 2, the deviance of the current model is the deviance of Model 3.

	df	SS	MS	F	p
Reduction in deviance	1	17.52	17.52	1.01	.344
IQ & RA & Scale A	8	138.30	17.288		
IQ & RA	9	155.82			

Table 8.15 Step 3 in predicting exam scores, Scale A enters last.

For steps 1 and 2 the F ratios are large and the p value shows them to be significant at the .05 level. For step 3, F is very close to one. As the null hypothesis is that the two MS are estimates of the same variance i.e., they are equal and so F = 1, it is not surprising that the p value is high and not significant.

Box 8.2 summarises the procedure followed. A general step by step procedure for descriptive stepwise regression is provided in the work sheets as Box 8.5.

Box 8.2 The procedure for stepwise regression (descriptive use)

1. The first predictor to be considered is the one that correlates best with the variable to be predicted, here exam score. This variable is said to enter the equation first. The significance of this simple correlation is assessed in the normal way. If it is significant then step 1 is said to be significant. In this case IQ enters first and this first step is significant at the .05 level.

2. Each of the other predictors is tried in combination with the first variable to enter the equation in turn. IQ with Reading Age gives the lowest deviance so Reading Age is the next variable to enter the equation. An analysis of variance table is specially constructed to test the significance of this step. Step 2 is shown to be significant at the .05 level.

3. In the general case this procedure is repeated, adding each predictor left to the current model to find the next variable to enter the equation and then testing the significance of the step. In the example here there were only 3 predictors so Scale A must enter next. The analysis of variance table constructed for this step showed it not to be significant. We conclude that a model with IQ and Reading Age accounts for all the important variance in exam scores.

Other procedures for stepwise regression as a descriptive technique

The procedure described in Box 8.2 is to start with the the simplest model and add variables. At each step the variable which gives the biggest reduction in deviance is added to those already entered into the equation. This is probably the most commonly used procedure. A similar, but not always exactly equivalent procedure, is to start with the full model (all predictors) and progressively remove variables. The variable removed at each step is the one which has least effect on the deviance of the regression equation. There are other variations on this theme. All can be interpreted in the same way as the more common procedure described here.

Reporting the results of multiple and stepwise regression

The multiple correlation coefficient R can be reported along with the regression equation. Its significance can then be reported along with the analysis of variance summary table. Multiple regression is often a way of summarising a correlation matrix and it is good practice with all the techniques described in this chapter to provide a correlation matrix for all the variables being considered. Similarly, a reader will want to verify that the variables are within the expected range for the population tested so a table of means and standard deviations for the variables should also be included. The multiple regression of exam score on Scale A and IQ might be reported as follows.

> Results:-
>
> The regression of recall on Scale A and IQ was computed. R^2 = .734 and the regression equation is
>
> exam_score = 10.3 + 0.963 Scale_A + 0.505 IQ
>
> R is significant at the .05 level (see Table 8.6 above for analysis of variance summary table). The correlation matrix describing the correlations between exam score, Scale A, IQ and Reading Age are given as Table 8.1 (above) and Table 8.16 includes the means and standard deviations for each of these three variables.
>
> | Variable | Mean | Standard deviation |
> |---|---|---|
> | Exam score | 71.25 | 8.71 |
> | Scale A | 10.25 | 4.27 |
> | IQ | 101.17 | 8.45 |
> | Reading Age (months) | 102.83 | 3.64 |
>
> **Table 8.16** Descriptive data for exam scores, Scale A and IQ.

In the case of stepwise regression for statistical control the result is the partial correlation. That should be reported and described along with the correlation between predicted and predictor variables for comparison. It is probably most straightforward to quote the squared partial correlation as this is can be determined directly from the stepwise regression. If the sign of the partial correlation is critical then the formula given in the summary for computing a partial correlation from the correlation matrix should be used.

To demonstrate the significance (or otherwise) of the partial correlations give the analysis of variance summary table for the stepwise regression. The experiment looking at the effect of age on recall could be reported as follows.

> Results:-
>
> The correlation matrix including the correlations between the three variables is given as Table 8.7 (above). There is a high correlation between recall and age but also a moderate positive correlation between age and time taken to learn the list. To statistically control for this possible confounding variable the squared partial correlation between recall and age, controlling for time spent learning the list, was computed. It is .97. This very high correlation is higher than the correlation between recall and age when time spent learning the list is not statistically controlled (r^2 = .76). The significance of the partial correlation was assessed by stepwise regression. The analysis of variance table for this is given as Table 8.9 (above). This shows that the partial correlation is significant at the .05 level. Table 8.17 gives the mean and standard deviation of each of these three variables.
>
> | Variable | Mean | Standard deviation |
> |---|---|---|
> | Recall | 24.6 | 3.37 |
> | Age (months) | 67.5 | 7.25 |
> | Time to learn list (mins.) | 16.0 | 5.70 |
>
> **Table 8.17** Descriptive data for the recall experiment

When reporting a descriptive stepwise regression a correlation matrix and table of means should also be reported. The 'result' in this case is the order in which the variables enter the equation and whether the step in which they entered is significant. One could provide a complete analysis of variance table for each step or just the F ratio, degrees of freedom and significance. The latter option is illustrated below for the exam scores example.

> Results:-
>
> The correlation matrix describing the correlations between exam score, Scale A, IQ and Reading Age are given as Table 8.1. All four variables are intercorrelated and exam score correlates highly with the other three variables. A descriptive stepwise regression was performed to determine the relative predictive power of these variables to predict exam score when used in combination. The first variable to enter was IQ (F(1,10) = 14.24, p < .05). The next variable to enter was Reading Age (F(1,9) = 10.88, p < .05). Scale A entered last. This step was not significant (F(1,8) = 1.01, n.s.). Table 8.16 gives the mean and standard deviation of each of these variables.

Notice the convention used to report an F ratio. For example, ' (F(1,8) = 1.01, n.s.)' summarises Table 8.15. The degrees of freedom for the two mean squares are given in brackets, 'F(1,8) ='. The significance of F is either 'p <.05', meaning significant at the .05 level, or as here 'n.s.', meaning non significant.

Summary

1. Multiple regression is an extension of simple regression. One variable is predicted from more than one predictor. The regression equation is the linear

combination of predictors giving the least deviance ('linear combination' means multiplying by constants and then adding as in a linear equation).

2. Different combinations of predictor variables in multiple regression can be thought of as different models to be fitted to the data and the deviance of the equation as the 'error' of the model. Stepwise regression involves the progressive addition of terms to a model.

3. Stepwise regression can be used to statistically control one or more nuisance variables. Here the procedure is to use the regression equation containing the nuisance variable(s) as the baseline deviance rather than the deviance from the mean. The reduction in deviance when the other predictors are added to this baseline model shows how much variance is left to explain when the variance attributable to the nuisance variable(s) has been removed. With one predictor and one nuisance variable this is known as partial correlation.

4. A partial correlation can also be computed from a correlation matrix using the following formula

$$r_{xy.z} = \frac{r_{xy} - r_{zx}\, r_{zy}}{\sqrt{(1-r_{zx}^2)(1-r_{zy}^2)}}$$

Where $r_{xy.z}$ is the partial correlation. This gives the correlation between x and y after partialling out the effect of the nuisance variable z. r_{xy} is the correlation between x and y, r_{zx} is the correlation between z and x, and r_{zy} is the correlation between z and y.

5. Stepwise regression can be used to describe how a set of intercorrelated predictor variables relate to one another in the prediction of some other variable. This descriptive technique is most usually structured as follows.

(a) the predictor with the best correlation with the predicted variable enters the equation first.

(b) the predictor, which in combination with the first predictor, gives the least deviance is the second predictor to enter the equation.

(c) the predictor, which in combination with the first and second predictors, gives the least deviance is the third predictor to enter the equation.

(d) and so on.

This is repeated until all the predictor variables are in the equation. The order in which the predictors enter the equation give an indication of their importance, when used in combination to predict the predicted variable.

6. An analysis of variance table can be constructed to test the significance of r, R or a step in stepwise regression.

(a) In each case there is a current model to be evaluated with respect to a baseline model. For r and R the current model is the regression equation and the baseline the deviance of the mean (Model 0). For stepwise regression the current model is the regression equation after the step has been taken and the baseline model is the regression equation before the step was taken.

(b) The change in deviance in moving from the baseline model to the current model is computed and compared with the deviance of the current model in an F ratio. To do this MS (Mean Squares) are computed by dividing the deviances by their degrees of freedom.

(c) The null hypothesis is that the MS for the change in deviance and the MS for the deviance of the current model are independent estimates of the same random variation and F is only greater than 1 because these estimates are approximate.

(d) By making certain assumptions, statisticians have computed a theoretical distribution for F. SPSS computes the probability of getting an F of the observed size or greater using the F distribution.

(e) If that probability is less than or equal to .05 the change in deviance is said to be significant at the .05 level.

(f) If it is not we are not able to reject the null hypothesis.