10 Statistical analysis

Statistical analysis provides a powerful range of techniques for assessing the significance of data which has been collected. Statistical techniques are regularly used in industry, commerce and government to answer a wide range of questions, from the quality of goods produced in a factory, to the most effective medical treatments, or the likelihood of air pollution or traffic congestion in different parts of a city.

For statistical analysis to be carried out, the required data must be in a numerical format. Data collected in the social sciences is often **qualitative**, such as opinions expressed in interviews or in questionnaires. A quantitative treatment can, however, provide a clear view of the structure of the data and can be used in comparisons between data sets. We begin by examining this issue:

Converting qualitative opinions to quantitative data

Education students might carry out a project to compare the success of different methods of teaching mathematics to school children. One approach would be to collect the examination and test results of the children, perhaps at the start and end of the year. This would provide sets of quantitative data which could be directly analysed. However, this would only give a partial picture. We might also want to know the attitudes of the children towards mathematics; whether they enjoy and value the subject, and are enthusiastic to learn more mathematics. This is important if we hope to encourage children to choose careers in the fields of science, technology, engineering or mathematics.

The researchers might interview the children and ask for their opinions about learning mathematics. Responses would be recorded as a series of statements, for example:

'I think maths is an important subject'
'Learning maths is a waste of time'
'I think maths is OK' …..

If we are to make a comparison between two groups of pupils, who are perhaps being taught mathematics by different methods, we will need a way of quantifying the opinion data. This is commonly carried out by means of a Likert scale. A series of numerical values correspond to a range of feelings from very negative, through neutral, to very positive. We might use a five point scale:

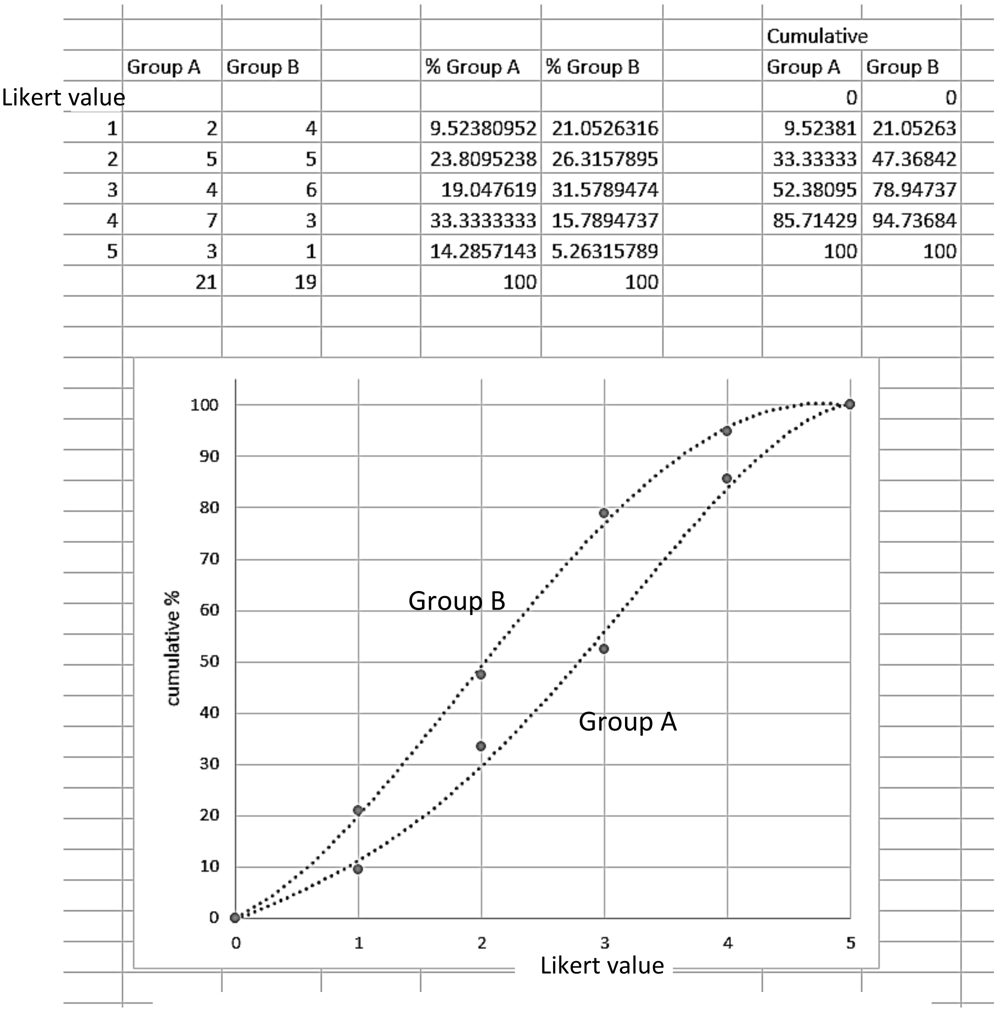
Very negative	Slightly negative	Neutral	Slightly positive	Very positive
1	2	3	4	5

Figure 298: Five point Likert scale

Rather than ask the children to directly allocate scale values themselves, it is often better for the researchers to assess the statements and carry out the allocation. It is good practice for two researchers to work together in making the allocation of Likert values, so that there is opportunity for discussion of the exact significance of the statements made by individual children. Examples of agreed values might be:

'I like maths'	4	quite positive
'I think maths is boring'	2	quite negative
'I wish I did not have to do maths'	2	quite negative
'Maths is my favourite subject'	5	very positive
'I hate maths'	1	very negative
'I usually enjoy maths lessons'	4	quite positive
'I don't understand maths'	2	quite negative
'I don't think maths should be compulsory'	3	neutral

Once the sets of Likert scale values for each pupil group have been decided, the data can then be processed quantitatively. This may involve a statistical test for significant difference between the groups. We might also wish to display the data graphically, so that a visual comparison can be made. Example data for two groups of pupils are shown in figure 299 below. Counts for the two groups A and B are tabulated for each Likert scale value. These results are then converted to percentages, and displayed as cumulative curves.



							Cumulative	
	Group A	Group B		% Group A	% Group B		Group A	Group B
Likert value							0	0
1	2	4		9.52380952	21.0526316		9.52381	21.05263
2	5	5		23.8095238	26.3157895		33.33333	47.36842
3	4	6		19.047619	31.5789474		52.38095	78.94737
4	7	3		33.3333333	15.7894737		85.71429	94.73684
5	3	1		14.2857143	5.26315789		100	100
	21	19		100	100			

Figure 299: Cumulative curves for pupil opinions

From the results in figure 299, group A seems to have developed a more positive attitude towards mathematics. 80% of group B recorded neutral or negative opinions of mathematics, whilst this drops to 50% in group A.

When handling results from Likert scale surveys, whether recorded directly by the participants or coded by the researchers, care must be taken not to claim excessive accuracy for the results. It is tempting to quote averages of sets of quantitative data to one or more decimal places, when only an estimate to the nearest whole number or less is justified. *Opinions* are much less precise than *physical measurements* such as times or distances.

Survey instruments

A special type of opinion survey which is intended to measure specific characteristics of participants is known as a ***survey instrument***. We again convert qualitative statements of opinion into numerical data, but this is used directly to calculate an individual result rather than being summarised for a group. We will look at two examples:

Big five personality characteristics

Many psychologists (for example: Barrick and Mount, 1991; Judge et al., 1999) believe that there are five basic aspects of personality, often referred to as the 'big five' personality characteristics. These are: extraversion, agreeableness, openness, conscientiousness and neuroticism. You can gain an interesting insight into your personality by answering a questionnaire about your attitudes and behaviour. Calculations can then be carried out to obtain scores for each of the five personality characteristics:

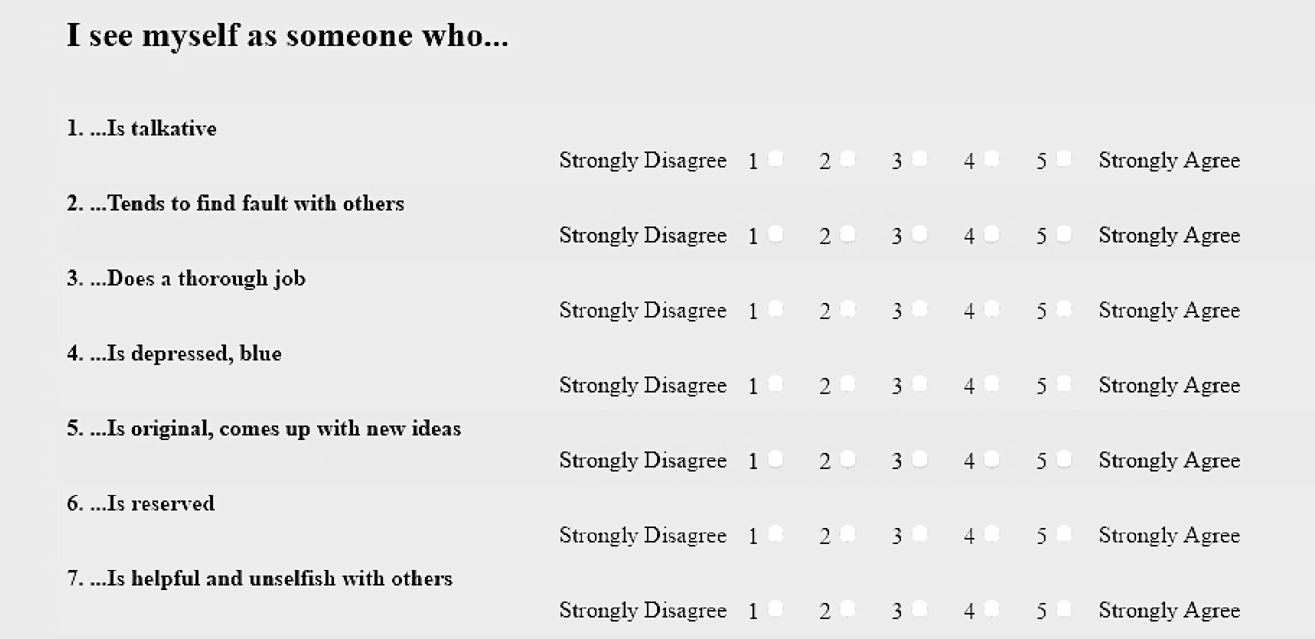
> ***Openness to experience/intellect***
>
> People with a high score tend to be original, creative, curious, complex. Those with a low score tend to be conventional, down to earth, have narrow interests, and be uncreative.
>
> ***Conscientiousness***
>
> High scorers tend to be reliable, well-organised, self-disciplined, and careful. Low scorers tend to be disorganised, undependable, and negligent.
>
> ***Extraversion***
>
> High scorers tend to be sociable, friendly, fun loving, and talkative. Low scorers tend to be introverted, reserved, inhibited, and quiet.
>
> ***Agreeableness***
>
> High scorers tend to be good natured, sympathetic, forgiving, and courteous. Low scorers tend to be critical, rude, harsh, and callous.

Neuroticism

High scorers tend to be nervous, highly-strung and insecure, and tend to worry. Low scorers tend to be calm, relaxed, secure, and hardy.

An online questionnaire to evaluate these personality characteristics can be found at:
www.outofservice.com/bigfive/

To complete the questionnaire, a series of questions are answered using a Likert scale, examples of which are shown below. Summary scores are then displayed for each personality characteristic:



I see myself as someone who...

1. ...Is talkative
Strongly Disagree 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Strongly Agree

2. ...Tends to find fault with others
Strongly Disagree 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Strongly Agree

3. ...Does a thorough job
Strongly Disagree 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Strongly Agree

4. ...Is depressed, blue
Strongly Disagree 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Strongly Agree

5. ...Is original, comes up with new ideas
Strongly Disagree 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Strongly Agree

6. ...Is reserved
Strongly Disagree 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Strongly Agree

7. ...Is helpful and unselfish with others
Strongly Disagree 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Strongly Agree

Figure 300: Example questions from the personality questionnaire

In the second example, we will look at another survey instrument related to psychology:

Belbin role models

R. M. Belbin has made an extensive study of the ways in which different members of teams work together on tasks (Fisher et al., 1998). Belbin suggests that team members seek out particular roles which fit their personalities and abilities, and where they can contribute most effectively to the overall task. By identifying the strengths and limitations of particular individuals, it is possible to build a balanced team which will work well together and have all the necessary skills available to achieve a successful outcome.

Belbin identified eight specific roles that members of a team may select. It is quite possible for an individual to undertake more than one of these roles:

Fixer

The fixer uses their inquisitive nature to find ideas to bring back to the team. They are outgoing, and enthusiastic, exploring opportunities and developing contacts.

Team Player

Helps the team to work together effectively. A co-operative, perceptive and diplomatic person who listens and averts conflict between team members.

Co-ordinator

Focuses on the team's objectives, drawing out team members and delegating work appropriately. They have a mature, confident personality, clarifying goals and identifying team members with particular talents.

Ideas Person

Highly creative and good at solving problems in unconventional ways. Creative, imaginative, free-thinking, and able to generate ideas to solve difficult problems.

Evaluator Judge

Maintains a logical overview of the work, making impartial judgements where required and weighing up the team's options in an objective manner.

Energiser

Provides the necessary enthusiasm to ensure that the team keeps moving and does not lose focus or momentum. They have a dynamic personality, thriving on pressure and having the courage to overcome obstacles.

Doer

Needed to plan a workable strategy and carry it out as efficiently as possible. The doer is practical and reliable in turning ideas into actions, and efficient in organising the work that needs to be done.

Quality Finisher

This role is particularly important at the end of a task, to check the work for errors and subject it to high standards of quality control. The quality finisher is painstaking and conscientious, valuing perfection.

In addition, Belbin identified the role of **Specialist**, who may be called in when particular knowledge and skills are needed which are not currently available within the team.

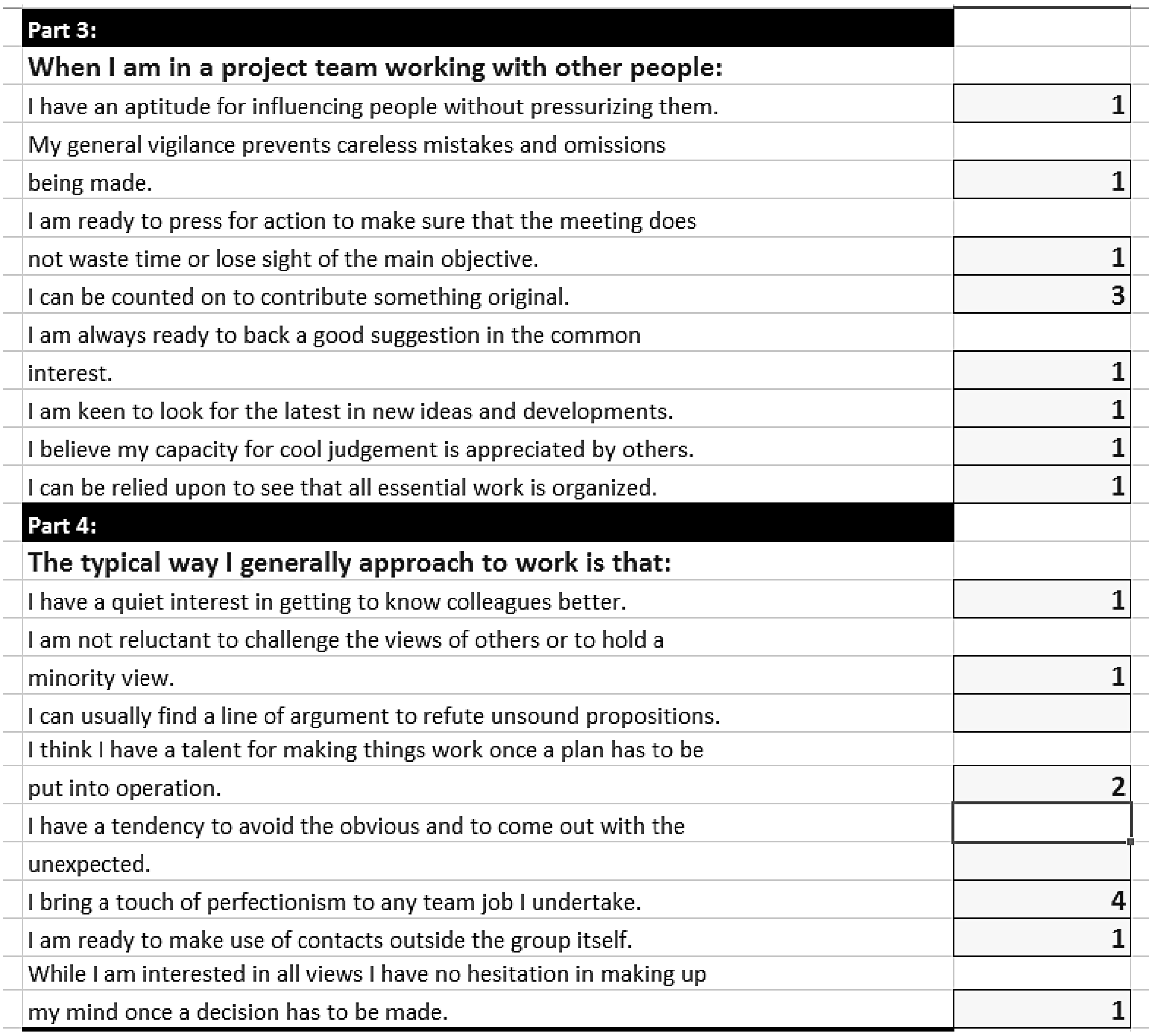
A spreadsheet survey instrument is available which can be used to assess the team roles which most suit the abilities and personalities of particular individuals. This may be downloaded from:

www.grahamhall.org/FEnumeracy/belbin.xls

The spreadsheet is divided into a number of sections, as shown in figure 301. In each section, the subject should read the set of statements and then allocate 10 points in whatever way they wish between these. A larger number of points can be allocated to a statement if it is considered to be particularly appropriate.

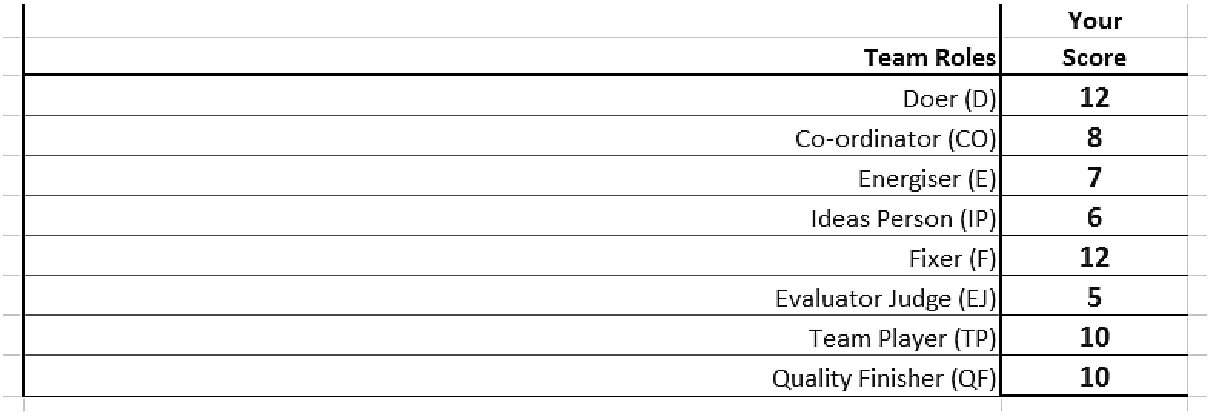
After completion of the sections of the spreadsheet, the subject's relative preferences for the different team roles will be displayed in a table at the bottom of the sheet (figure 302).

It may be the case that one of the roles appears dominant, or the subject may find that they have a balanced preference for several of the team roles.



Part 3:	
When I am in a project team working with other people:	
I have an aptitude for influencing people without pressurizing them.	1
My general vigilance prevents careless mistakes and omissions being made.	1
I am ready to press for action to make sure that the meeting does not waste time or lose sight of the main objective.	1
I can be counted on to contribute something original.	3
I am always ready to back a good suggestion in the common interest.	1
I am keen to look for the latest in new ideas and developments.	1
I believe my capacity for cool judgement is appreciated by others.	1
I can be relied upon to see that all essential work is organized.	1
Part 4:	
The typical way I generally approach to work is that:	
I have a quiet interest in getting to know colleagues better.	1
I am not reluctant to challenge the views of others or to hold a minority view.	1
I can usually find a line of argument to refute unsound propositions.	
I think I have a talent for making things work once a plan has to be put into operation.	2
I have a tendency to avoid the obvious and to come out with the unexpected.	
I bring a touch of perfectionism to any team job I undertake.	4
I am ready to make use of contacts outside the group itself.	1
While I am interested in all views I have no hesitation in making up my mind once a decision has to be made.	1

Figure 301: Example statement sections from the Belbin Role Models survey



Team Roles	**Your Score**
Doer (D)	**12**
Co-ordinator (CO)	**8**
Energiser (E)	**7**
Ideas Person (IP)	**6**
Fixer (F)	**12**
Evaluator Judge (EJ)	**5**
Team Player (TP)	**10**
Quality Finisher (QF)	**10**

Figure 302: Example results produced by the Belbin Role Models survey

Distributions

Many statistical methods depend on an analysis of the distribution of values within a data set. It is useful to carry out some practical investigations of distributions, as a way of gaining a clearer understanding of these statistical methods.

Binomial distribution

The simplest distribution to understand is the binomial distribution. This is a representation of the probabilities of different outcomes when a series of events occur, each having just two possible results. For example, consider the pin board game in figure 303.

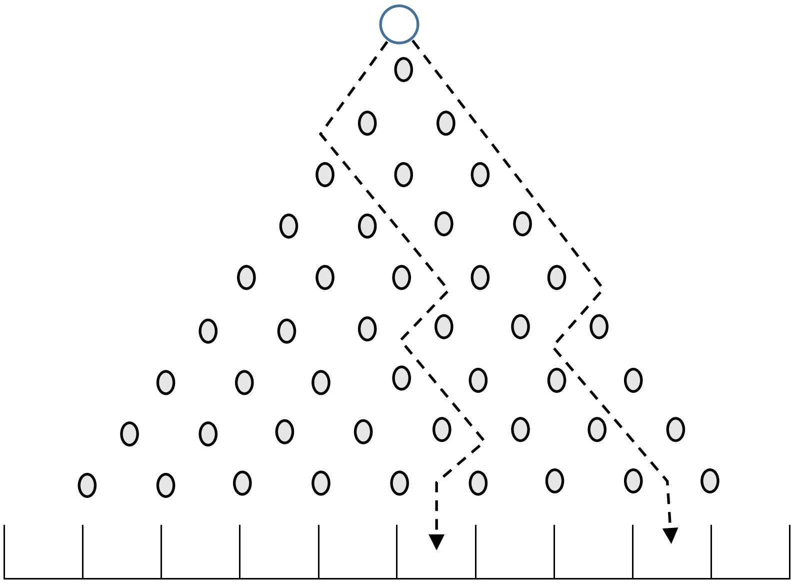
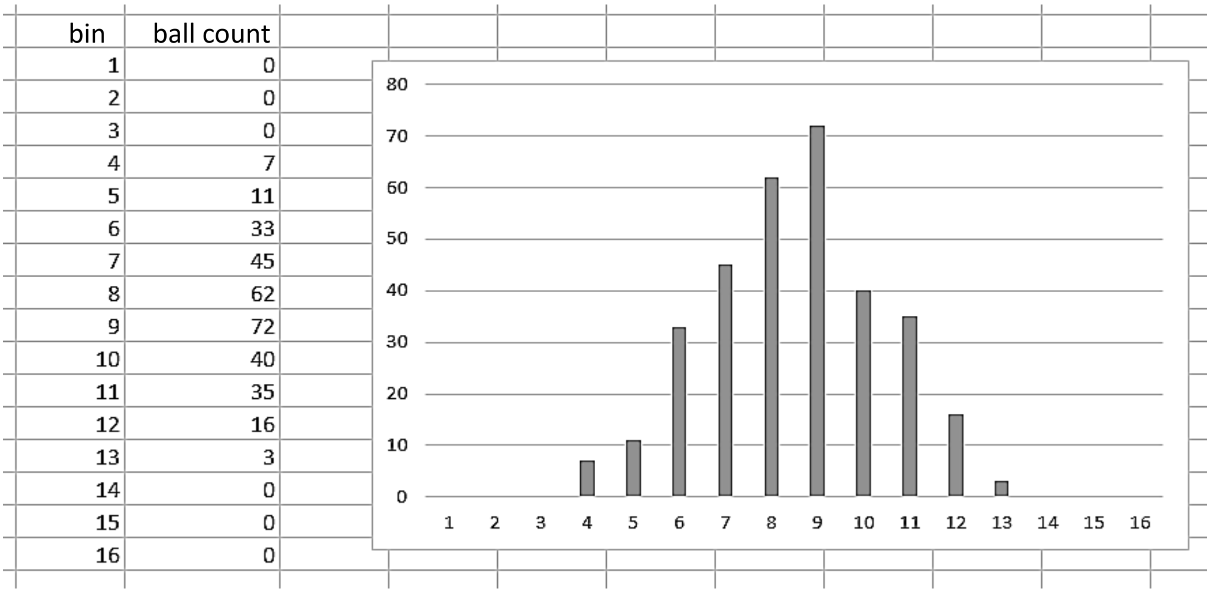


Figure 303: Pin board game to demonstrate the binomial distribution

A ball is released, and can fall to either the left or right of the first pin. The ball reaches a pin on the second layer, and can again pass to either the left or right. This process is repeated until the ball arrives at one of the collecting bins at the bottom of the board. We assume that there are equal probabilities of the ball falling to the left or right at each pin.

If a large number of balls were to be released, say 200, it is interesting to consider whether we could predict the numbers of balls which would land in each of the bins. We might first carry out the practical experiment, either using a real pin board or by means of a spreadsheet simulation as in figure 304. Each outcome of the game is slightly different due to random chance, but we always find that the majority of the balls land in bins near the centre of the board. Few or no balls are found in the bins on the extreme edges of the pin board.

The mathematical explanation for this result lies in the array of possible paths through the board, each or which is equally probable. If we choose a bin at the centre of the board, we can find a large number of possible paths which lead to this point, so there is a high overall chance of a ball arriving there. However, there is only one possible pathway to the bin at the far edge of the board, so this will be a much less likely destination.



bin	ball count
1	0
2	0
3	0
4	7
5	11
6	33
7	45
8	62
9	72
10	40
11	35
12	16
13	3
14	0
15	0
16	0

Figure 304: Example results from the pin board game simulation

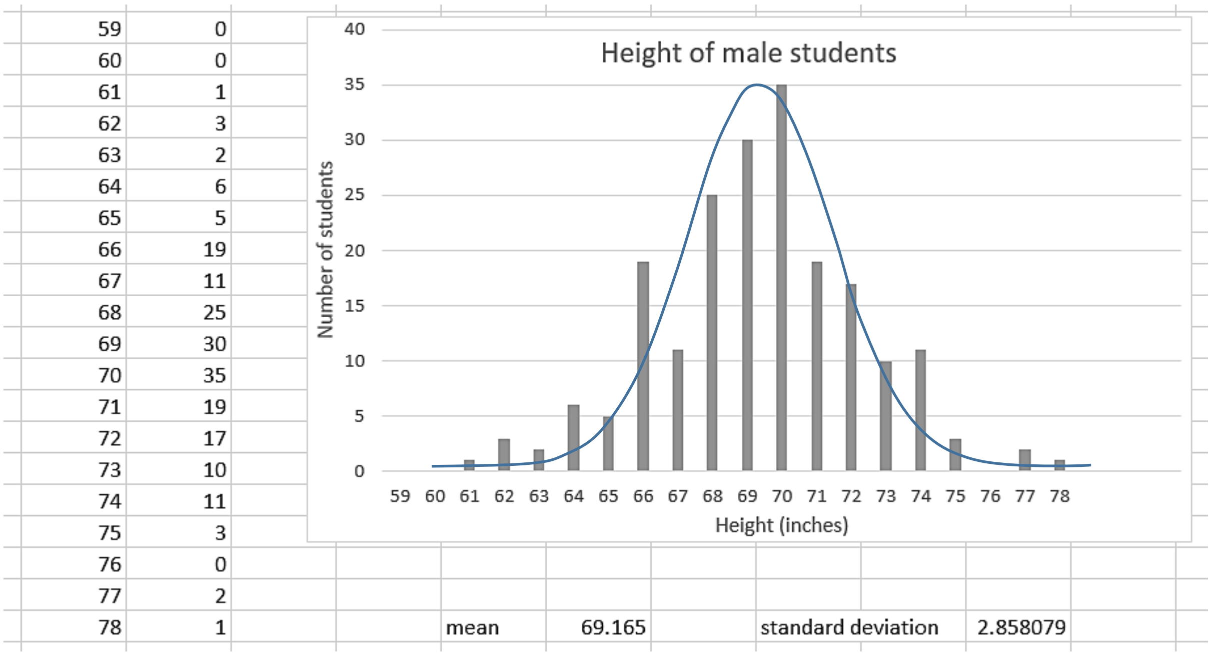
The bell-shaped curve seen in the binomial distribution, with most results grouped around the mean and relatively few results far away from the mean, is a very common distribution pattern found in statistics. We will now look at a similar distribution which occurs in many areas of everyday life:

Normal distribution

A normal distribution again has a bell-shape, but this time the measurements recorded are part of a continuous range, rather than being restricted to counts at specific locations such as the bins of the pin ball game.

As an example, we might make measurements of the heights of 200 male students. Heights can be measured to whatever degree of accuracy is available. In order to plot the data, we might select height intervals representing the nearest inch, then plot the number of persons whose heights fall within each interval.

Typical results are shown in figure 305 below. We find that most of the group have heights close to the mean, with relatively few very tall or very short individuals. We can make sense of this pattern by analogy with the pin board game. A person's height is the result of a large number of factors, some perhaps related to nutrition as a child, but others linked to genetic factors extending back over many generations through the randomness of their family tree. Each of these factors could lead to a slightly taller or shorter height. For any individual, it is most likely that a mixture of taller and shorter influences have balanced out to produce a height close to the mean. It is much less likely that every factor has had a taller height influence, or that every factor has had a shorter height influence.



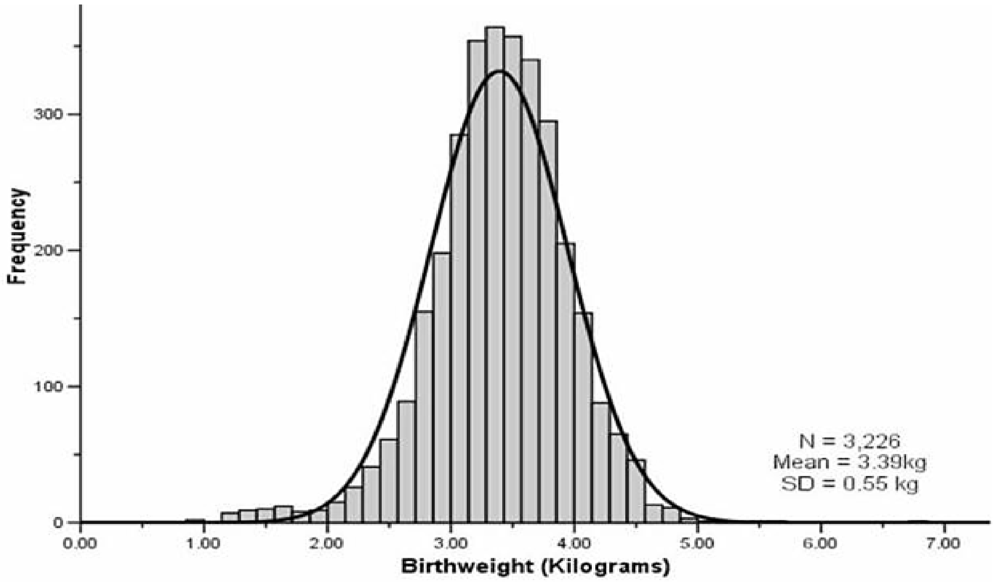
59	0
60	0
61	1
62	3
63	2
64	6
65	5
66	19
67	11
68	25
69	30
70	35
71	19
72	17
73	10
74	11
75	3
76	0
77	2
78	1

mean	69.165	standard deviation	2.858079

Figure 305: Plot of student heights

The underlying bell-shape of the frequency curve for the normal distribution has been sketched on figure 305. If larger numbers of students were included in the height survey, it is likely that the randomness would average out and the data points would move ever closer to the ideal curve.

Normal distribution patterns are frequently used in health and social care to monitor patients, and identify measurements outside the expected range for the general population. Birth weight, for example, is found to be normally distributed.



www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/statistical-distributions

Figure 306: Normal distribution of birth weights

It can be shown theoretically that the formula for the normal distribution curve is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \, . \, e^{-(x-\mu)^2/2\sigma^2}$$

This formula involves two variables which control the shape of the bell curve:

- the mean **μ** which determines the horizontal position for the peak of the curve
- the standard deviation **σ** which determines how narrow or wide the curve should be

If a set of data values such as the heights of our 200 male students are entered into an Excel spreadsheet, functions are available to calculate the mean and standard deviation. For example, the values obtained from the student height data in figure 305 were:

mean	69.2 inches
standard deviation	2.9 inches

These values can be very useful when analysing a data set. It is known that:

- 68% of values will be expected to lie within 1 standard deviation of the mean. We can therefore predict that two-thirds of all male students will have a height between 66.3 and 72.1 inches, or roughly between 5 feet 6 inches and 6 feet.
- 95% of values will be expected to lie within 2 standard deviations of the mean. We can therefore predict that nearly all male students will have a height between 63.4 and 75.0 inches, or roughly between 5 feet 3 inches and 6 feet 3 inches.

Often we wish to make a specific prediction based on a normal distribution. As an example, we will estimate the percentage of babies with a birth weight or 4.0 kg or greater, using the data given in figure 306:

mean	3.39 kg
standard deviation	0.55 kg

A simple starting point is to consider a normal distribution in which the mean has a value of 0 and the standard deviation has a value of 1. This is known as a **standard normal distribution**. The equation can then be simplified to:

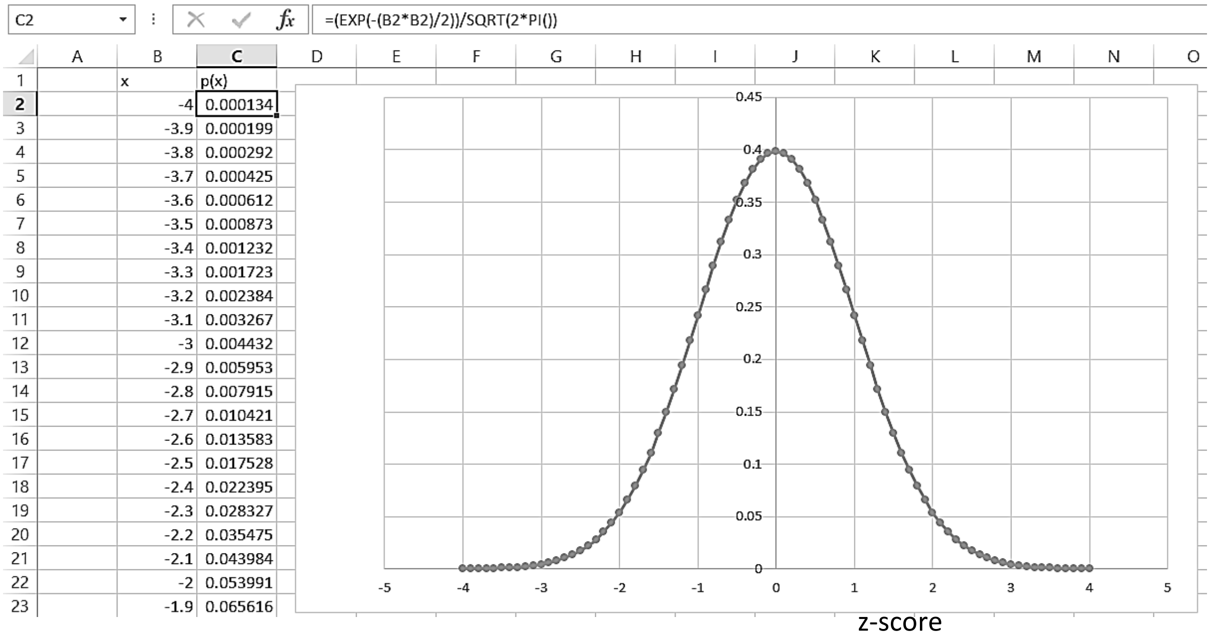
$$f(x) = \frac{1}{\sqrt{2\pi}} \, . \, e^{-x^2/2}$$

We can plot this with a spreadsheet, as shown in figure 307 below. The horizontal variable, which represents units of standard deviation, is called the **z-score**. We are interested in the percentage of babies with a birth weight or 4.0 kg or greater. The first step is to convert 4.0 kg to a z-score. This is done by means of the formula:

$$z = \frac{(x - \mu)}{\sigma}$$

which maps our actual normal distribution to a standard normal distribution, by moving the mean to zero and adjusting the width of the curve to a standard deviation of 1.

$$z = \frac{(4.0 - 3.39)}{0.55} = 1.1$$

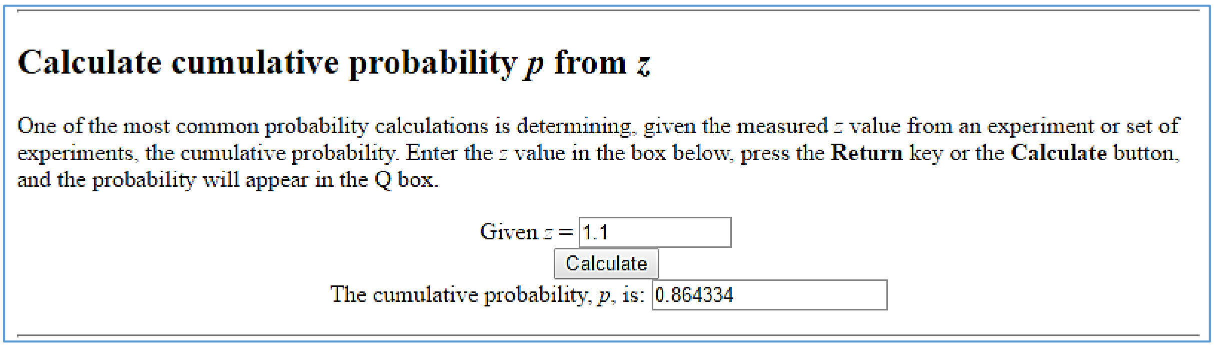


C2 | =(EXP(-(B2*B2)/2))/SQRT(2*PI())

	A	B	C
1		x	p(x)
2		-4	0.000134
3		-3.9	0.000199
4		-3.8	0.000292
5		-3.7	0.000425
6		-3.6	0.000612
7		-3.5	0.000873
8		-3.4	0.001232
9		-3.3	0.001723
10		-3.2	0.002384
11		-3.1	0.003267
12		-3	0.004432
13		-2.9	0.005953
14		-2.8	0.007915
15		-2.7	0.010421
16		-2.6	0.013583
17		-2.5	0.017528
18		-2.4	0.022395
19		-2.3	0.028327
20		-2.2	0.035475
21		-2.1	0.043984
22		-2	0.053991
23		-1.9	0.065616

Figure 307: Standard normal distribution

Once the z-score is known, we can find the percentage of readings above this value by consulting tables or an on-line statistics application:



Calculate cumulative probability *p* from *z*

One of the most common probability calculations is determining, given the measured *z* value from an experiment or set of experiments, the cumulative probability. Enter the *z* value in the box below, press the **Return** key or the **Calculate** button, and the probability will appear in the Q box.

Given *z* = 1.1

The cumulative probability, *p*, is: 0.864334

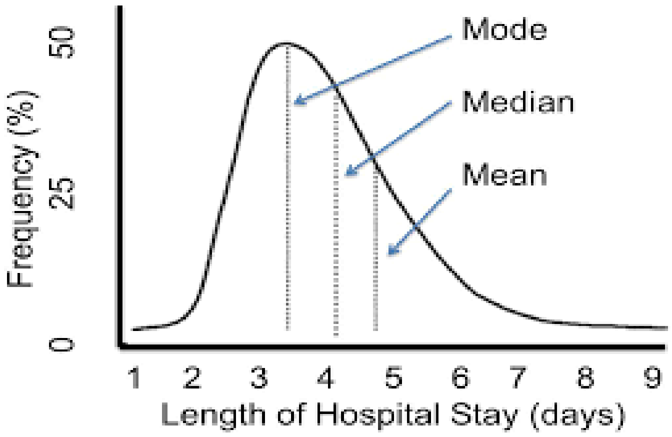
sampson.byu.edu/courses/z2p2z-calculator.html

Figure 308: Cumulative probability calculator for normal distributions

From this result, we find that 0.86 of readings (86%) are below the weight of 4.00 kg, so the percentage of babies with a birth weight at or above 4.0 kg will be 14%.

Skewed distributions

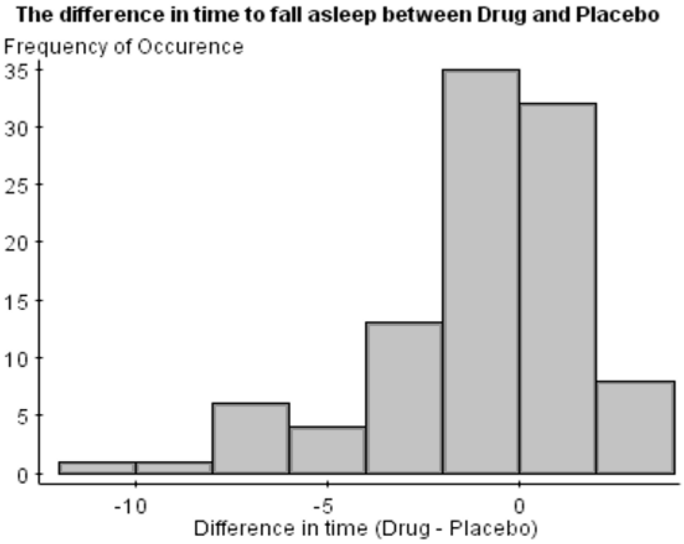
The normal distributions examined so far have been symmetrical in shape, but this is not always the case. In some situations, readings may decline more sharply on one side of the mean than on the other. An example might be the length of hospital stay by patients. Many procedures are completed quickly and patients are discharged within a few days, but some patients may need periods of treatment of several weeks or months. The shape of the distribution curve is stretched out in the positive direction, so is said to be **positively skewed** (figure 309).



sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability8.html

Figure 309: Positively skewed distribution

Situations may also arise where data points are stretched out in the negative direction, producing a **negatively skewed** distribution. An example might be the results of clinical trials of a new sleep medication. The drug may be very effective for a few patients, causing them to fall asleep much more quickly than normal. For a majority of patients, however, the drug may have little effect.



www.statcrunch.com/5.0/viewreport.php?reportid=15020

Figure 310: Negatively skewed distribution

The Excel spreadsheet provides a function to determine the skewness of a set of data values, in a similar way to the calculation of standard deviation.

Bimodal distribution

Another commonly occurring distribution pattern is the bimodal distribution. In this case, the single peak of a normal distribution is replaced by a double peak. This indicates that two separated values are dominant within the data set.

A bimodal distribution often indicates that two sub-sets of data have been sampled together, either deliberately or by accident. For example, a set of simple measurements of student height might show two peaks, representing the dominant mean heights respectively for females and males.

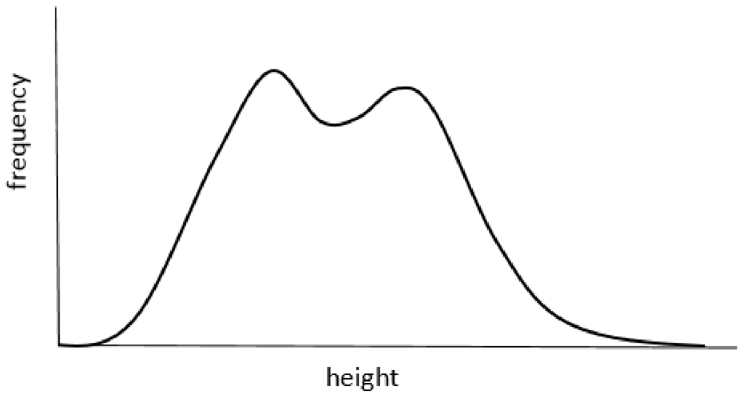


Figure 311: Bimodal distribution of student height

It is usually necessary to separate the sub-sets of data before analysis, since statistical tests are generally only valid for distributions which are close in shape to a normal distribution. However, the identification of a bimodal distribution can sometimes itself be an important diagnostic factor for a data set. We will look at the geographical example illustrated below:

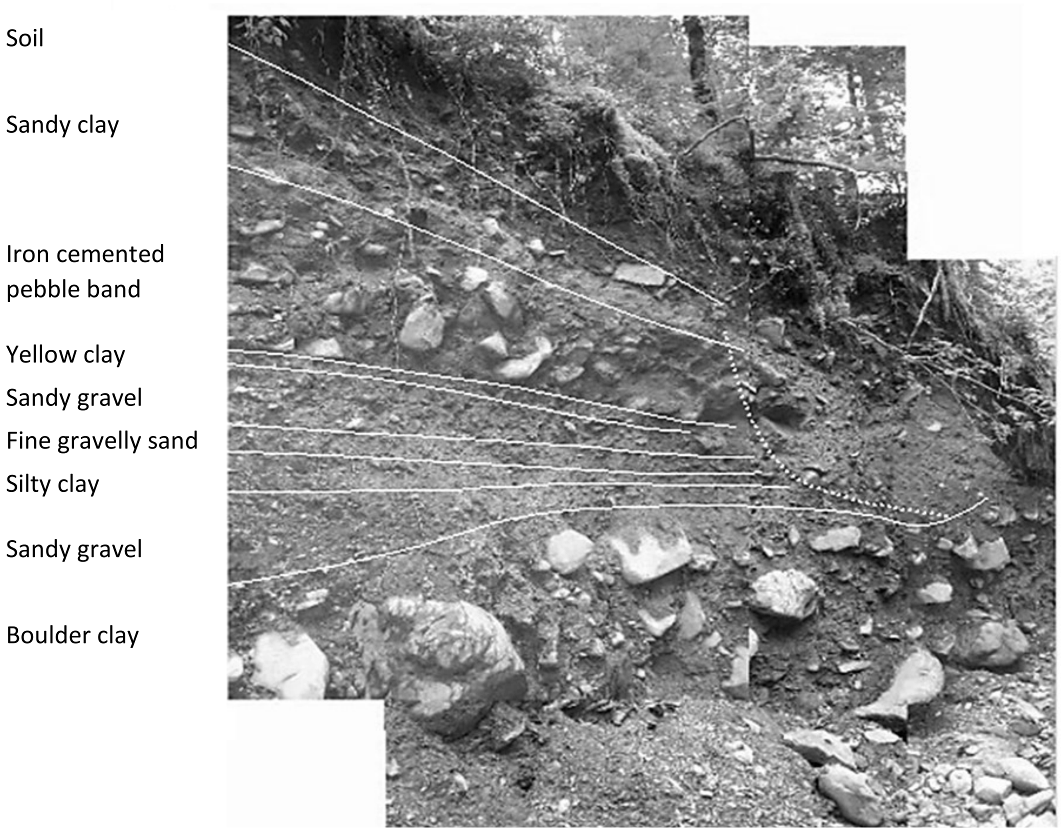


Figure 312: Glacial and periglacial deposits, Afon Wen valley, Snowdonia

At the end of the Ice Age, extensive deposits of clays, sands and gravels were laid down in North Wales. These materials have been well preserved in some deep river valleys such as the Afon Wen valley in Coed y Brenin, Snowdonia.

The sequence contains a number of beds made up from gravel mixed with finer sand and clay. These may have originated in different ways:

- As moraine, laid down beneath moving ice which filled the valley
- As river deposits, laid down by streams flowing from a melting glacier

To investigate the mechanisms of deposition, students collected samples from each layer. The samples were separated into fractions by sieving. The fractions were then weighed to determine the percentages of mud, silt, sand and pebbles present. Example results are shown below.

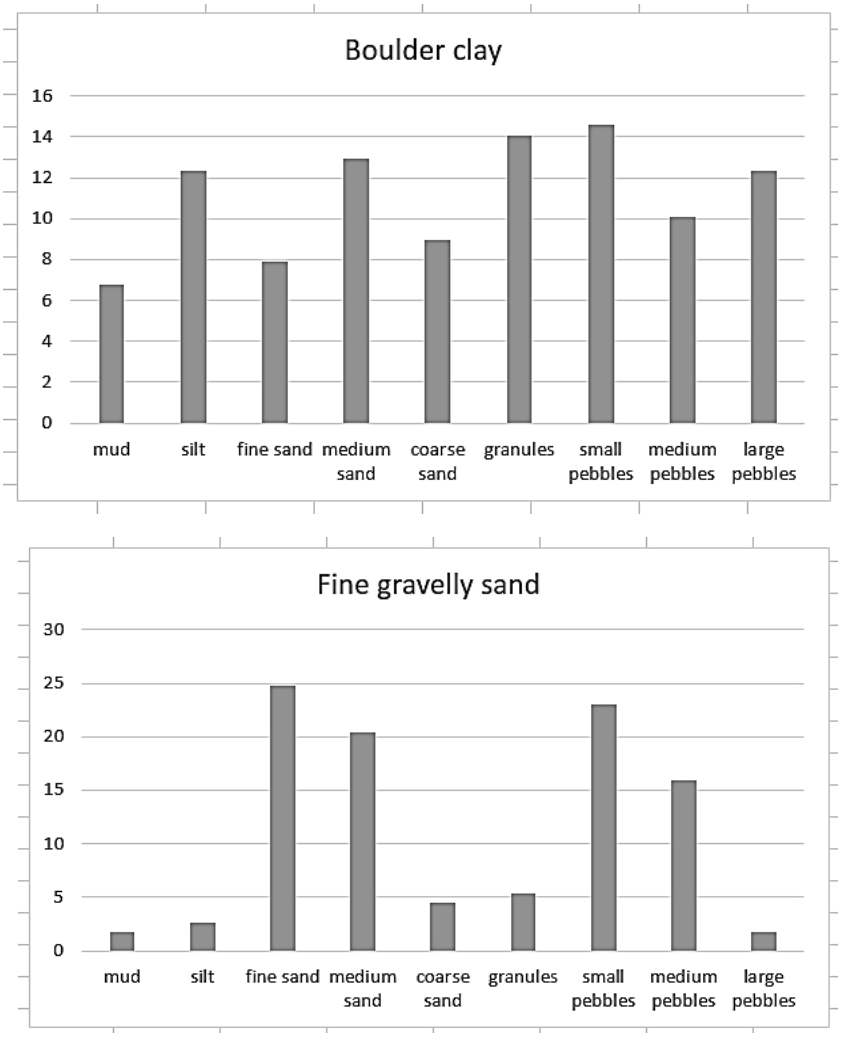


Figure 313:

Examples of sediment grain size distribution from glacial and periglacial deposits, Afon Wen valley, Snowdonia

We see that the size fractions are fairly evenly balanced in the first sample. This is characteristic of the poor sorting of material laid down by ice, and we interpret this as a glacial boulder clay or till deposit.

Sorting is better in the second sample. Two size grades are dominant, and the distribution shows a bimodal pattern. This is characteristic of river deposition, with pebbles deposited at this location at times of high flow, and finer sand deposited in-between the pebbles during times of low flow. We might interpret this as due to differing rates of ice melt on different days, depending on temperature conditions.

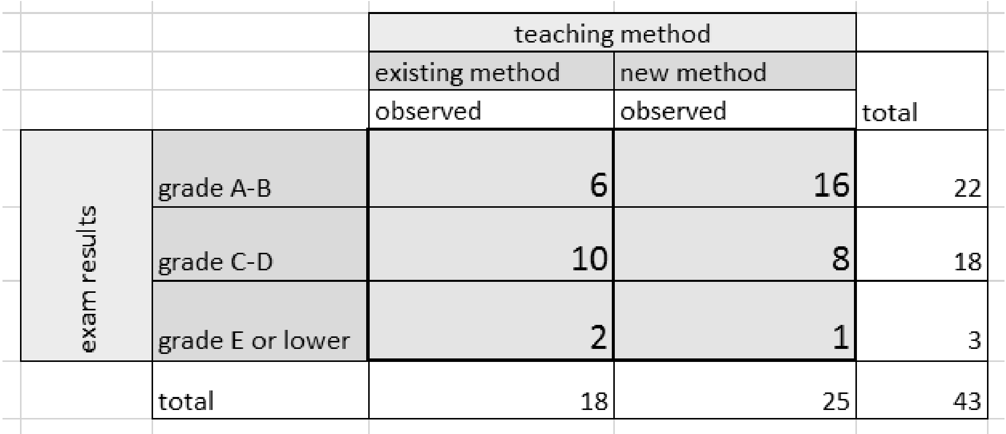
Chi-squared statistic

One of the fundamental tasks in statistics is to determine whether two groups of sampled data belong to the same overall population, or whether they belong to two distinct populations which differ in some important way. For example, education students might be interested in investigating whether different teaching methods have affected the outcomes for two groups of students who were taught in different ways. Assessment grades might be collected for the two groups, then averages calculated.

- If group A had an average of 80% and group B an average of 40%, we would probably conclude that a definite difference exists. However, care should be taken in interpreting these results. We have not yet ***proved*** that teaching is the cause of the difference, and some other factor or factors might be involved.
- If group A had an average of 60% and group B an average of 62%, we would probably dismiss this small difference as being the result of random chance. We would conclude that we have found no difference in outcome between the two groups.

Before advocating a change in teaching methodology for all students, we might wish to establish that the new teaching method used in the research project definitely had an effect in improving grades. The question then arises as to how great a difference between the two experimental groups would be considered significant? Would we accept a grade average of 65% in comparison to 60% as showing that the outcomes for the two student groups are fundamentally different? Fortunately there are objective mathematical methods available for determining whether differences are significant. We will examine one of these, the **chi squared test**, in this section.

We begin by making a table of observed data. The test requires us to provide counts for different subgroups: in this case, we might record the numbers of students gaining different ranges of grades when taught by the two methods.



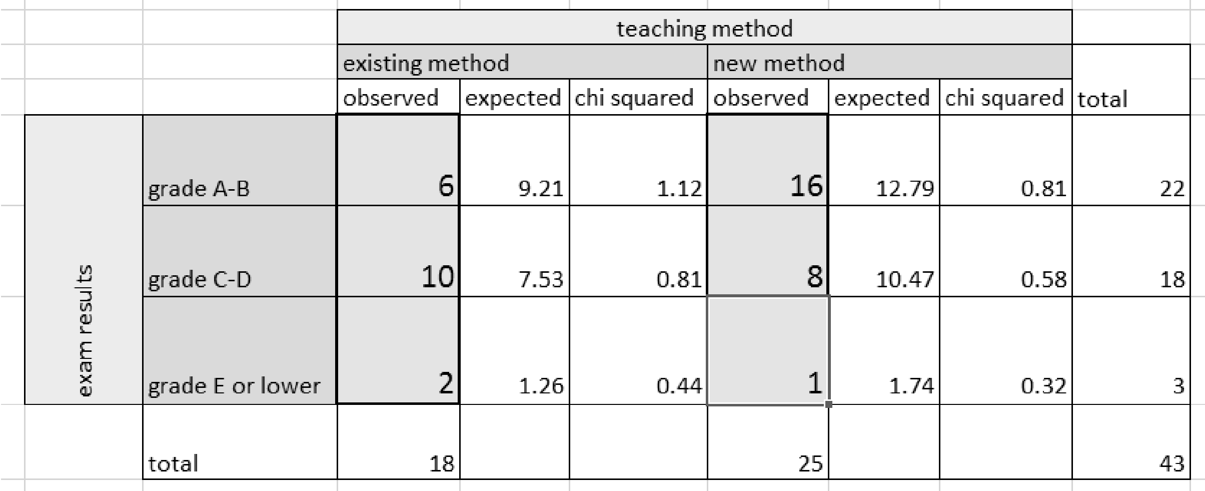
		teaching method		
		existing method	new method	
		observed	observed	total
exam results	grade A-B	6	16	22
	grade C-D	10	8	18
	grade E or lower	2	1	3
	total	18	25	43

Figure 314: Observed values for the chi squared test

We move on to calculate the chi-squared statistic for our set of data values. The first step is to calculate the expected outcomes if there is no difference in grades between the two experimental groups. For example:

> We have recorded a total of 18 students with grades C-D
>
> Overall, there are 18 students being taught by the existing method and 25 students being taught by the new method. We would therefore expect slightly more of the C-D grades to be in the larger group of 25 students taught by the new method.
>
> The 18 C-D grades are shared out in the ratio 18:25, giving a prediction of 7.53 in the smaller class and 10.47 in the larger class.

Predicted values are calculated in a similar way for the other grade ranges, then added to the table (figure 315).



		teaching method						
		existing method			new method			
		observed	expected	chi squared	observed	expected	chi squared	total
exam results	grade A-B	6	9.21	1.12	16	12.79	0.81	22
	grade C-D	10	7.53	0.81	8	10.47	0.58	18
	grade E or lower	2	1.26	0.44	1	1.74	0.32	3
	total	18			25			43

Figure 315: Observed and expected values for the chi squared test

The final step is to calculate a **chi-squared** value for each data item. This is done by means of the formula:

$$\chi^2 = \frac{(observed - expected)^2}{expected}$$

This will be a measure of how close the experimental results were to our theoretical values, which assume no difference in grades between the two experimental groups. The six chi-squared vales are then added to give an overall **chi-squared statistic** of **4.07** for the set of data.

We now determine a quantity known as the **degree of freedom** for the data. The totals of each row and column are now fixed, as these have been used to calculate theoretical expected values. How many of the experimental observation results would need to be specified before all the remaining values in the table are known unambiguously?

Our table consists of three rows of data values in two columns. Suppose that the data value in the top row of column 2 is specified:

	column 1	column 2	
row 1	automatically set	SPECIFIED	row 1 total
row 2	?	?	row 2 total
row 3	?	?	row 3 total
	column 1 total	column 2 total	

For the total of the top row to be correct, the data in column 1 must be a particular value. However, there is still uncertainty in the remaining row values.

	column 1	column 2	
row 1	automatically set	SPECIFIED	row 1 total
row 2	SPECIFIED	automatically set	row 2 total
row 3	automatically set	automatically set	row 3 total
	column 1 total	column 2 total	

We might specify the value in column 1 of the second row. This will then cause all remaining cells to take particular values in order to make the row and column totals correct.

We might choose to specify the values of different cells of the table, but we will only ever be able to set two values independently if the row and column totals are to remain correct. Once the two selected cells have been specified, all others will be set automatically. The table is therefore said to have **two degrees of freedom** because only the totals plus two data values need to be specified before all remaining values are known. In general, the number of degrees of freedom for a table of observations is given by:

$$degrees\ of\ freedom = (rows - 1) \times (columns - 1)$$

We now need to consider how to interpret this result. A chi-squared distribution is obtained by taking a standard normal distribution, then finding the square of the difference between each data point and the mean.

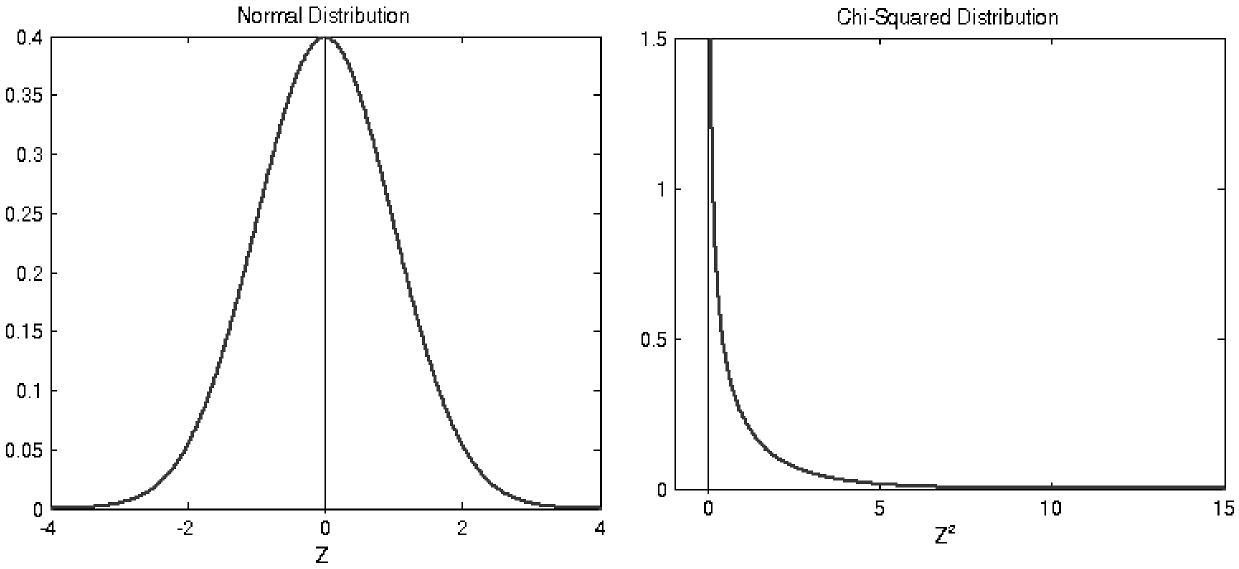


Figure 316: Relationship between the normal distribution and chi-squared distribution

We notice that the chi-squared distribution only has positive values due to the effect of squaring each positive or negative difference. The chi-squared curve appears as a distorted version of the positive normal distribution curve. A fractional z-score of less than 1.0 will be reduced by squaring, so the probability initially falls more steeply. A z-score above 1.0 will be increased by squaring, so the positive tail of the distribution is extended.

The chi-squared distribution can be used to model the variations from the mean when random samples are selected from a normally distributed population. This represents the case of **one degree of freedom**.

To model a data set with two degrees of freedom, we again take a standard normal distribution and find the square of the difference between each data point and the mean. A second set of data points is then randomly selected, the differences from the mean are squared and the results added to each of the first set of differences. The combined probabilities of the two events are then plotted. Mathematically, this can be done by means of calculus. A new chi-squared distribution curve is produced.

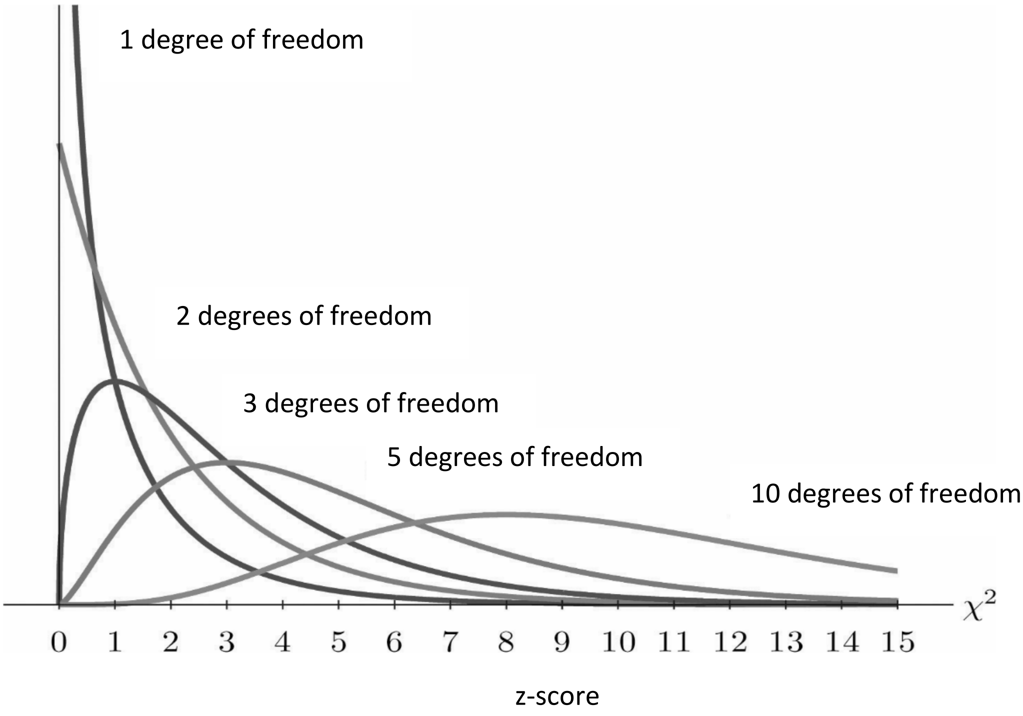


Figure 317: Chi-squared distributions for different degrees of freedom

Another set of random squared differences can be added for each extra degree of freedom. This produces a family of chi-squared probability distribution curves. We see that the curves become less skewed and approach the shape of a normal distribution as the degrees of freedom increase. This is reasonable behaviour. As the number of random events is increased then there is an increasing chance that each point will be generated by adding a mixture of higher and lower values. A bell-shaped curve starts to develop. The **mean** of the distribution moves to the right as more differences are added for each degree of freedom, and the overall sum for each data point increases.

We can now return to the question of how to interpret the chi-squared result from the survey of student grades. Our data has two degrees of freedom, so the chi-squared distribution for two degrees of freedom should be used.

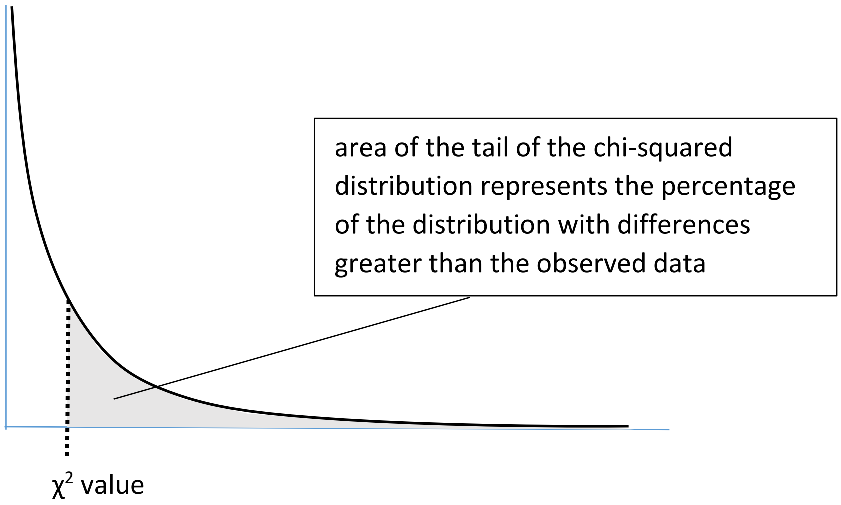
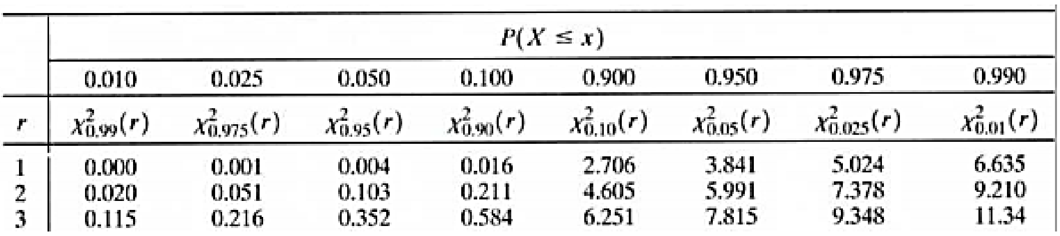

area of the tail of the chi-squared distribution represents the percentage of the distribution with differences greater than the observed data

Figure 318: Interpreting the chi-squared result

We need to determine the area beneath the distribution curve which lies to the right of our chi-squared result. This will indicate the percentage of the reference population with a greater variation from the mean than the observations which we recorded.

- If most of the reference population has a larger variation than our own data, then our data is close to the mean and is likely to belong to a single population.
- If very little of the reference population has a larger variation than our own data, then our data is likely to belong to two populations with different mean values. Students taught by the new method would have grades belonging to a different distribution with higher grade outcomes.

Areas beneath the chi-squared curve can be found from statistical tables, using the required degree of freedom:



	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34

onlinecourses.science.psu.edu/stat414/node/147

Figure 319: Chi-squared probability table

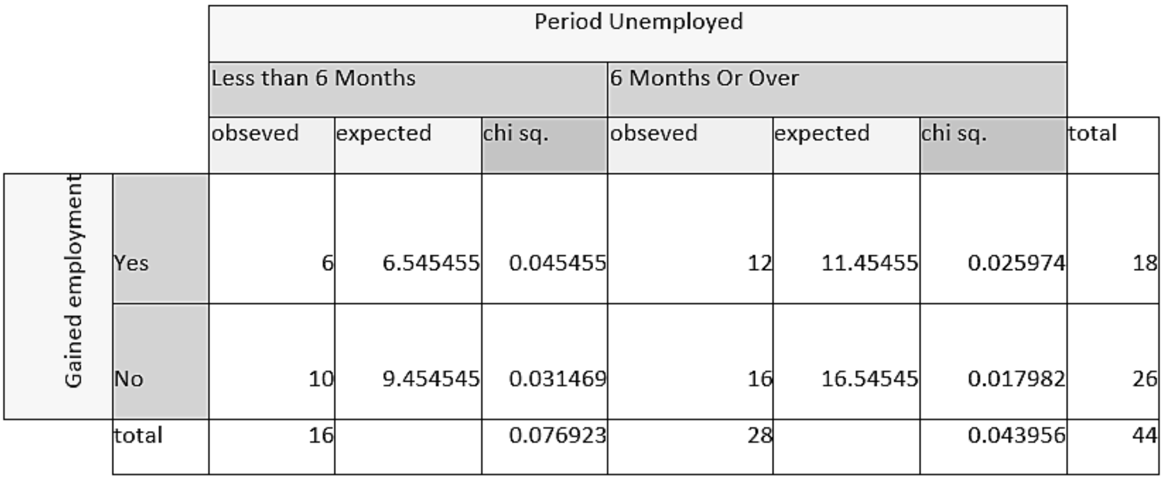
For two degrees of freedom, we find that our result of 4.07 is just below the 0.9 probability value, meaning that only about 10% of the reference population would have a greater variation from the mean than our student groups. The high level of variation in our experiment makes it highly likely that the students taught by the new method belong to a different population distribution which has higher grade outcomes. We might recommend introducing the new teaching method more widely.

Employment in Blaenau Ffestiniog

As an example of a larger scale use of the chi-squared statistical method, we will examine a project carried out by an education student to evaluate the effectiveness of an adult training centre in Bleanau Ffestiniog. The centre provides vocational training courses for unemployed adults, with the objective of improving their chances of gaining employment. Trainees would typically spend six months at the centre, during which they would complete an NVQ course in Information Technology, along with training in a range of work skills. After leaving the centre, the trainees were monitored and supported in their search for employment, and records were kept of outcomes.

The objective of the research project was to examine a group of adults who had passed through the training programme, and to try to identify factors which affected their subsequent success in finding employment. Possible factors identified were:

- The period of **time unemployed** before attending the training course. It was thought that adults who had been out of work for a long period would find it more difficult to return to employment.
- **Qualification** obtained during the training course. It was hoped that the trainees who successfully completed the NVQ qualification would find it easier to gain employment.
- **Age** might be a factor, although the effect was uncertain. It is possible that younger applicants would be preferred for some posts, whilst older and more experienced applicants might be preferred for other roles.
- **Gender** might be significant. This is an area where the number of traditionally male manual jobs has declined, but jobs in tourism-related businesses such as guest houses and restaurants are increasing.
- **Bilingualism**. It was thought that organisations within this Welsh speaking region might prefer to employ staff able to communicate in both English and Welsh.



		Period Unemployed						
		Less than 6 Months			6 Months Or Over			
		obseved	expected	chi sq.	obseved	expected	chi sq.	total
Gained employment	Yes	6	6.545455	0.045455	12	11.45455	0.025974	18
	No	10	9.454545	0.031469	16	16.54545	0.017982	26
	total	16		0.076923	28		0.043956	44

Figure 320: Chi-squared test for *period of employment* against *employment outcome*

Data was collected, and each of the factors listed above was compared against employment outcomes using two-by-two data grids with a single degree of freedom. Expected data values were then calculated, and chi-squared statistics produced. An example, comparing **period of unemployment** with **employment outcome** is shown in figure 320.

The chi-squared statistics from each test were interpreted using the figures for one degree of freedom in a probability table (see figure 319). Results are given below:

Comparator tested against employment outcome	Chi-squared result	Significance
Period unemployed: • Less than 6 months • 6 months or over	0.121	P = 0.25 Slight correlation of employment outcome with period unemployed
Qualification: • NVQ IT achieved • NVQ not completed	1.841	P = 0.75 Moderate correlation of employment outcome with qualification achieved
Age: • Less than 40 • 40 or over	5.200	P = 0.98 Very strong correlation of employment outcome with age
Gender • Male • Female	4.739	P = 0.97 Very strong correlation of employment outcome with gender
Bilingualism • English and Welsh • Monolingual English	0.021	P = 0.12 Very unlikely to be a correlation of employment outcome with bilingualism

The research indicates that the two dominant factors affecting employability were **age** and **gender**. Men were experiencing more difficulty in obtaining jobs than women, and trainees over the age of 40 were finding it more difficult to re-enter employment.

The achievement of the NVQ IT qualification offered at the training centre seems to have been a moderate advantage when seeking work. The length of time unemployed before training had very little effect on employment outcome.

A surprising outcome was that bilingualism appears not to affect employability. It is possible that local employers, who are mainly in the tourism industry, are willing to employ staff who can only communicate with guests in English.

Mann-Whitney U Test

In common with the chi-squared test, the Mann-Whitney U test is used to investigate whether two samples of data belong to the same population or to two different populations. The underlying strategy, however, involves a different approach.

Samples are collected from the two data sets, then arranged into sorted order.

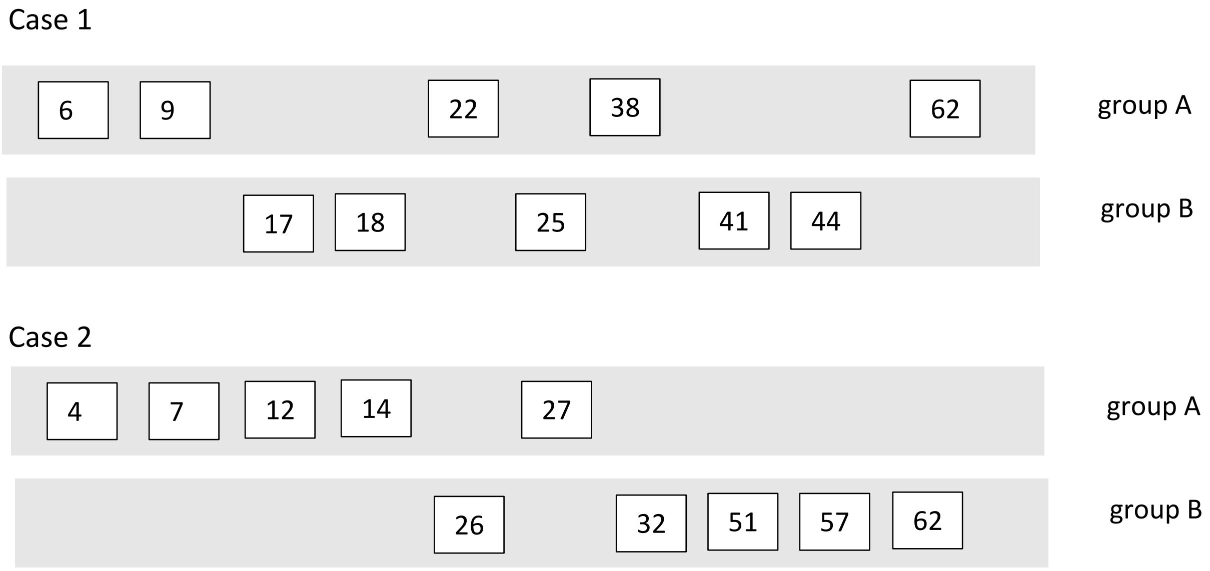


Figure 321: Rationale for the Mann-Whitney U test

We might expect that there will be some overlap between the groups. To carry out a Mann-Whitney U test, calculations are made to determine how many data items in one group are larger or smaller than each of the data items in the other group.

In case 1 above, the overlap is considerable, so it is very likely that groups A and B are samples from the same underlying population. In case 2, however, the overlap is slight. We might conclude that groups A and B are samples from two different populations.

In the Mann-Whitney U test, we are simply examining the sorted order of the samples, rather than considering their actual data values. The test therefore checks for different **medians**, rather than a different mean. The median of a data set is the data item which is in the middle position when the data values are arranged in order.

To demonstrate the use of the Mann-Whitney U test in a substantial project, we present below the report of an experiment carried out by a psychology student to investigate the topic of reconstructive memory:

'Dr Who' experiment

Abstract

To investigate reconstructive memory and the effect of pre-existing schemas on recall, this study, using an independent measures design, compared "Doctor Who" fans' and non fans' recollection of the text of a short "Doctor Who" story. An opportunity sample of thirty participants from internet message boards and from a sixth form college took part using a website which presented the text to them and then recorded their recollection of it, which was then analysed in terms of errors and elaborations. The level of difference between the distortions made by each group was found to be significant at a level of $p < 0.01$ using the Mann- Whitney test, with the "Doctor Who" fans distorting their recollection of the text to become more conforming to the conventions and clichés of "Doctor Who". This provides support for memory being reconstructed from schemas.

Introduction

Memory is vital for understanding and responding to the world around us, so the extent to which it can be relied on as an accurate representation of actual events is an important question, especially with regard to eyewitness testimony. According to Bartlett (1932), people do not recall information precisely as they experienced it; rather, they reconstruct their memories of past events and information through the use of schemas. Schemas are templates that provide a framework of generalities into which the specific details of events and information can be fitted. This can on the one hand aid memory in preventing the need to remember common elements every time they occur, and on the other, distort memory in causing inaccurate stereotypical aspects to be recalled.

To study this, Bartlett asked students to learn and repeat back the short text of the story of "The War of the Ghosts", a North America Indian folk tale, the intention being that it would conflict with the students' cultural schemas. He found that students did indeed distort the style and content of the text and rationalised it in order to make it logical and understandable to them. This could take the form of "flattening" (omitting of unusual elements), "sharpening" (emphasising/elaborating existing elements), "elaboration" (the introduction of new information) as well as omitting original information.

Further research on this includes Wynn and Logie (1998), who found that students' recollections of their first week of university remained consistent when repeated reproduction was requested over time during their first year, which suggests that experiences, as opposed to learned information, may not be subject to distortion in this way. Shank and Abelmarit (1977) said that people have "scripts" for situations, such as visiting a restaurant, and in Bower, Black and Turner (1979) participants recalled details that conformed to a restaurant script but were not actually present. This suggests that distortions resulting from schemas occur in everyday situations.

However, Bartlett has been criticised for not keeping vigorous enough experimental control over the experiment. There was no control group with which to compare the recollected versions of stories, and hence no way of examining to what extent these were due to differences in cultural schemas. This study attempts to do this by using the subculture of "Doctor Who" fandom. "Doctor Who" fans come from the same wider culture as non-fans but have pre-existing familiarity with "Doctor Who" storytelling conventions, clichés, patterns and general information. This allows the interference of pre-existing knowledge and schemas to be isolated from any general patterns of memory distortion that may occur. It is expected that fans will distort the text to become more like their "Doctor Who" schemas.

Aims and Hypothesis

The aim of the study is to investigate whether pre-existing schemas of information about something distort recall of new information on that subject, and the research hypothesis is that pre-existing familiarity with the subject matter of a new text leads to distortions in recall.

The experimental hypothesis is that "Doctor Who" fans will recall the text of a "Doctor Who" story with a significantly different level of distortion as measured in the number of alterations away from or towards "Doctor Who" than non-fans.

The null hypothesis is that there will be no significantly different level of distortion as measured in the number of alterations away from or towards "Doctor Who" between fans and non-fans.

Non-directional hypotheses are used because schema theory predicts that schema can both aid and distort memory, so it is not clear in what way previous familiarity will affect recall.

Method

The method was quasi-experimental, since it is not possible to randomly allocate participants to each condition, and has an independent measures design with both groups doing the same test of memory once. The independent variable was the presence or otherwise of pre-existing familiarity with "Doctor Who", and the dependent variable was the level of distortion. A pilot study was used to test the suitability of the method, in particular methods of measuring the level of distortion.

Participants

The two conditions used were "Doctor Who" fans and people who were not fans of "Doctor Who". Participants were an opportunity sample of volunteers, with an Internet message board being used to obtain sufficient numbers of "Doctor Who" fans. Non-fans participating in the experiment were college staff, fellow students, friends and volunteers from non-"Doctor Who" Internet message boards. The research was carried out by an A-level student who is himself a fan of "Doctor Who".

Participants were allocated to the group on the basis of their self-reported familiarity with the television series "Doctor Who". The pilot study had 5 participants, 2 fans and 3 non-fans, and in the actual study there were 30 participants in total, 16 "Doctor Who" fans and 14 non-fans.

Apparatus

A text was produced for participants to read and recall; specifically, an original brief "Doctor Who" story that would be comprehensible to both fans and non-fans but would in some respects fit, and in other respects depart, from fans' schemas of "Doctor Who", should such things exist.

The text, entitled "The Caves of Death", contained various common elements and clichés from "Doctor Who", although did not reference any particular stories. See Appendix A below for the story "The Caves of Death" and Appendix B for explanatory commentary.

A website was constructed using Microsoft FrontPage on which participants were issued with instructions, given the story to read, and then given a form to fill in on which they would put their recollection of the text along with other information, such as an indication of their consent. It was hosted on web space provided by the ISP. An email address, psy-coursework@...com, was created for the results to be emailed to and for any correspondence with participants.

Procedure

Firstly, participants were given an introduction to the experiment to read, which included that they have the right to withdraw at any stage. Full details of what is being investigated were not disclosed in order to hopefully reduce the effect of demand characteristics. If they wished to continue, they then clicked the "Continue" link at the bottom of the page. Participants were then instructed to read the text three times (increased from two in the pilot study) before clicking on the "Continue" link in order to fill in the form. The website logged when each user accesses each stage of the experiment, so it would be possible to check for any indication of "cheating".

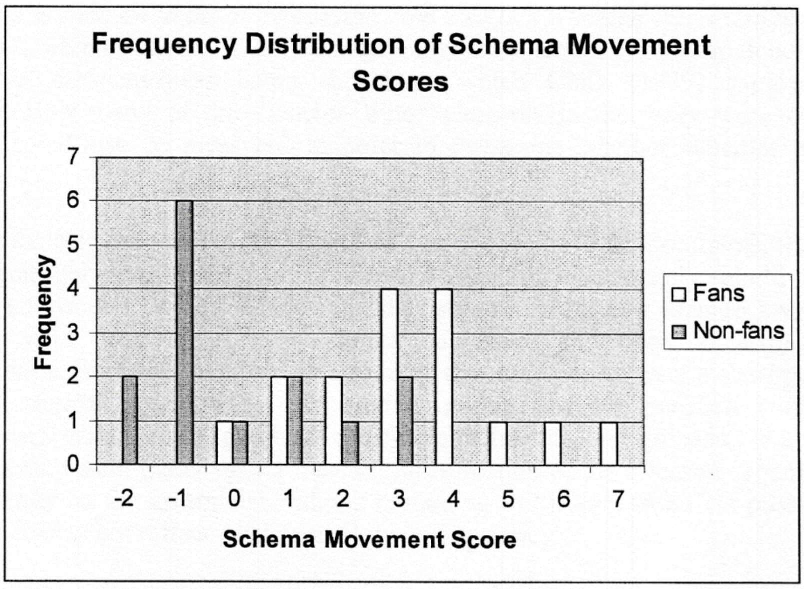
They were then presented with space to type their recollection of the text, and were also asked to fill in their name, email address, level of familiarity with "Doctor Who". The pilot study merely had Fan and No Knowledge (but participant feedback suggested the need for further categories), and to tick a box to indicate their consent. There was also an opportunity to add any comments.

The results were then emailed to the address mentioned above, and the participant was thanked for their contribution and reminded of the contact details should they have any queries or concerns, as well as being provided with links to further information about both reconstructive memory and "Doctor Who".

For the Pilot Study, the results were analysed to produce standardised numerical data of the way in which the text had been recalled, and consolidated into an overall Accuracy Score. However, the process of analysis was impractically time-consuming, and although the data collected was very rich, the necessity to reduce it down to one measurement meant that this method was unsuitable. It was decided that it would be more appropriate in terms of the aims of the study and easier to analyse to focus on the changes with regard to "Doctor Who" (the subject which participants were or were not familiar with). For the actual experiment, the number of errors or elaborations were counted that tended towards the "Doctor Who" schema by inclusion of clichés, conventions, continuity elements and so on not in the original text. Also counted were the number of errors or elaborations that tended away from the "Doctor Who" schema. The latter score was then subtracted from the former to give a measurement of the movement of each person's recollection towards or away from the "Doctor Who" schema - 0 being neutral, a positive score being more "Doctor Who"-like and a negative score being less "Doctor Who"-like.

Results

The schema movement scores proved statistically significant at a level of $p < 0.01$ when the Mann-Whitney test was applied, which means there is a probability of less than 1% of these results being due to chance and so the null hypothesis can be rejected. This is more significant than the level of $p < 0.05$ that was required. The critical value of U for the number of participants, which the calculated value of U in the results must be less than or equal to in order to be considered significant, was 50 for $p < 0.01$ and 64 for $p < 0.05$, and since $U = 22.5$ this shows a clear difference between the two groups. The Mann-Whitney test was applied because it compares the rankings of scores between the two groups and gives an indication of the statistical significance of any difference between the results of the two groups. If there is no difference, then there should be similar rank orders in the groups, but if there is a difference, then one group should display more high rankings and the other group more low rankings.



The graph above illustrates this, with the frequency distribution for non-fan showing lower scores to be much more common compared to fans. The frequency distribution for non-fans shows some irregularities, such as that more participants displayed a shift of 3 towards "Doctor Who" than displayed an overall shift of 0. This is probably because some in the non-fan group had enough familiarity with "Doctor Who" to distort their memories of the text, despite not considering themselves as "fans".

Interestingly, on average the level of distortion for non-fans was exactly 0, since there were equal numbers of distortions towards and away from "Doctor Who", though the mode was -1. The average number of distortions per participant among non-fans was 1.71, while among fans it was almost doubled at 3.00, which suggests that pre-existing schemas not only produce qualitatively different distortions, but also quantitative - they increase the number of mistakes.

Appendix A: Text of "The Caves of Death" story

The Caves of Death

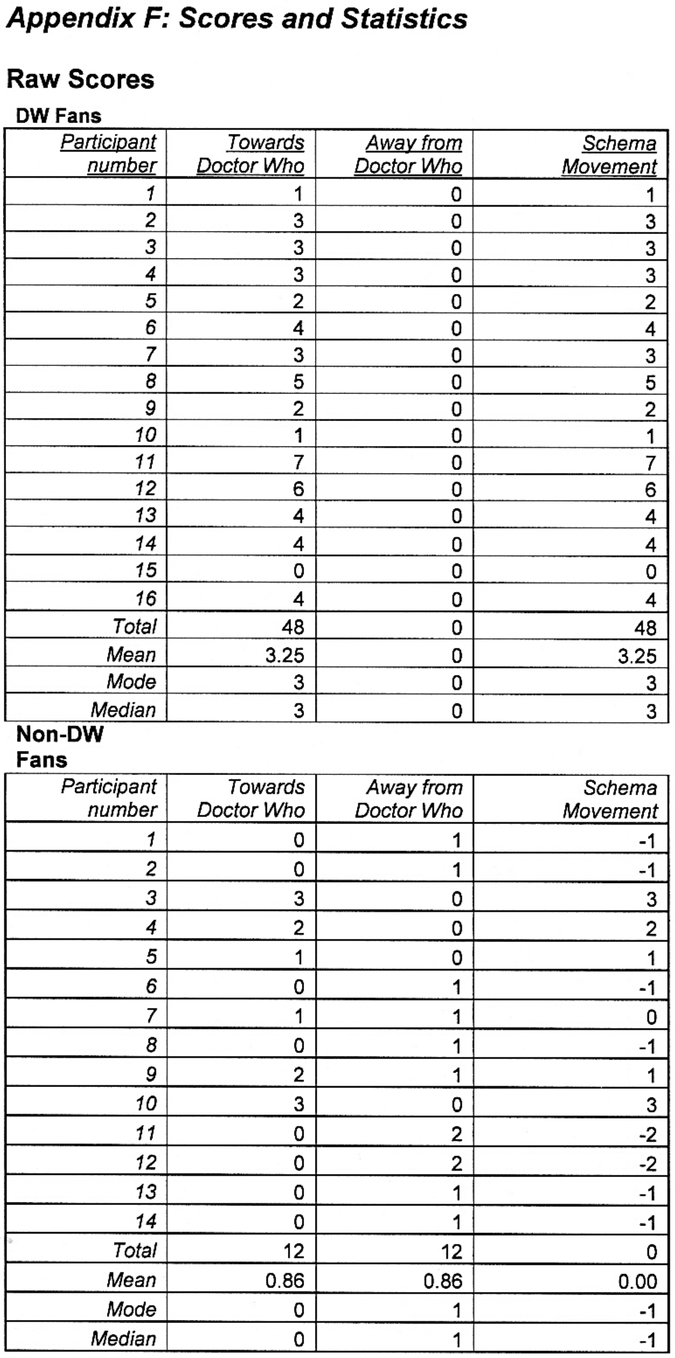
With a strange groaning noise, the TARDIS wheezed into existence. Out stepped the Doctor, currently in his third appearance and therefore a tall, white haired man with a young-old face. Having noticed some mysterious carvings, he started exploring, only for there to be a sudden rock fall! The noise seemed to disturb some being, so instead he headed deeper underground as it gave chase...

Archdeacon Izlambyr led the party of priests, each with the golden skull emblazoned on their black robes, to catch the reported intruder. Puzzled by the discovery of a blue box, he ordered it taken back to the Temple - but beware the Slyrka. Meanwhile, the Doctor desperately fought off the hideous creature, but was saved in the nick of time by the arrival of two of the priests. But his relief turned to horror as he realized he had fallen into the grasps of the Cult of the Holy Death. Taken back to their underground Temple with a heavy heart, the Doctor found himself locked up with a young guard, Froerg, who apologised for keeping him alive so long. The Doctor insisted that he wants to be alive, and that he would be quite happy to just leave quietly. His guard said that this was what they all said, but the Doctor would thank him for it later, metaphorically, for freeing him from the painful and meaningless existence of life. The Doctor used his Venetian Aikido to render the guard unconscious and unlock the door with his screwdriver.

Sneaking around, the Doctor was horrified to discover his ship had been put before the Idol of Death, in the thick of the worshippers. Searching his capacious pockets, he found a dimensional fractation distorter, with which he is able to redirect an unattended transmat beam to teleport himself next to the TARDIS. As he fumbled for the key, pandemonium broke out, but the High Abbot held the acolytes back, and the Doctor dashed into the TARDIS and took off, his hearts racing. Outside, the Cultists watched in grim satisfaction, knowing the impudent escape had sentenced himself to life.

Appendix B: Commentary on Text

- The title, "The Caves of Death", follows a similar construction to many Doctor Who stories, such as "The Web of Fear", "The Brain of Morbius", and "The Invasion of Time".
- The TARDIS is the Doctor's time and space machine. The stock description of the sound it makes is "a wheezing, groaning sound". The terms usually used for its appearance and disappearance is "materialization" and "dematerialization".
- The Doctor is able to regenerate his body into a new form. Each "incarnation" of the Doctor is usually referred to as "The First Doctor", "The Second Doctor" where necessary to distinguish between them.
- The description of the Doctor as a "tall, white haired man with a young-old face" is another stock description.
- A common cliché in Doctor Who was for the Doctor to be separated from his TARDIS by some misfortune that necessitates his exploration of wherever he has arrived. However, it is not stated in the text that the Doctor is cut off from the TARDIS.
- The "blue box" is the TARDIS, which is permanently stuck in the form of a Police Telephone Box due to the Chameleon Circuit having jammed.
- Monsters are one of the most memorable features of Doctor Who, and the "Slyrka" is a combination of "slither" and "lurker", as well as being reminiscent of the names of Doctor Who monsters such as the Slyther, Myrka and Shalka.
- Another staple cliché of Doctor Who was the Doctor being captured, discovering something of the villain's plans, and then escaping, which would be repeated as necessary.
- A deliberate mistake in the text was the Doctor having a "heavy heart" – being a Time Lord of Gallifrey, he has two hearts, as is mentioned at the end of the text when he escapes "hearts racing".
- Aikido is a martial art, and the Third Doctor, played by Jon Pertwee, made frequent use of "Venusian Aikido", although here there is another deliberate mistake as he uses the acoustically similar "Venetian Aikido" instead.
- Throughout much of the series, the Doctor made use of his trusty "sonic screwdriver", this is suggested through his use of a screwdriver to escape.
- "Capacious pockets" is another stock description, although more commonly applied to those of the Fourth Doctor, Tom Baker.
- "Dimensional fractation distorter" is a hopefully Who-ish sounding piece of technobabble. Some stories would resort to such devices to resolve a problem.
- "Transmat" is Doctor Who-speak for the more commonly used "teleport".
- "Taking off means dematerializing into the space-time vortex rather than flying off into the air, which would be rather difficult underground.



Appendix F: Scores and Statistics

Raw Scores

DW Fans

Participant number	Towards Doctor Who	Away from Doctor Who	Schema Movement
1	1	0	1
2	3	0	3
3	3	0	3
4	3	0	3
5	2	0	2
6	4	0	4
7	3	0	3
8	5	0	5
9	2	0	2
10	1	0	1
11	7	0	7
12	6	0	6
13	4	0	4
14	4	0	4
15	0	0	0
16	4	0	4
Total	48	0	48
Mean	3.25	0	3.25
Mode	3	0	3
Median	3	0	3

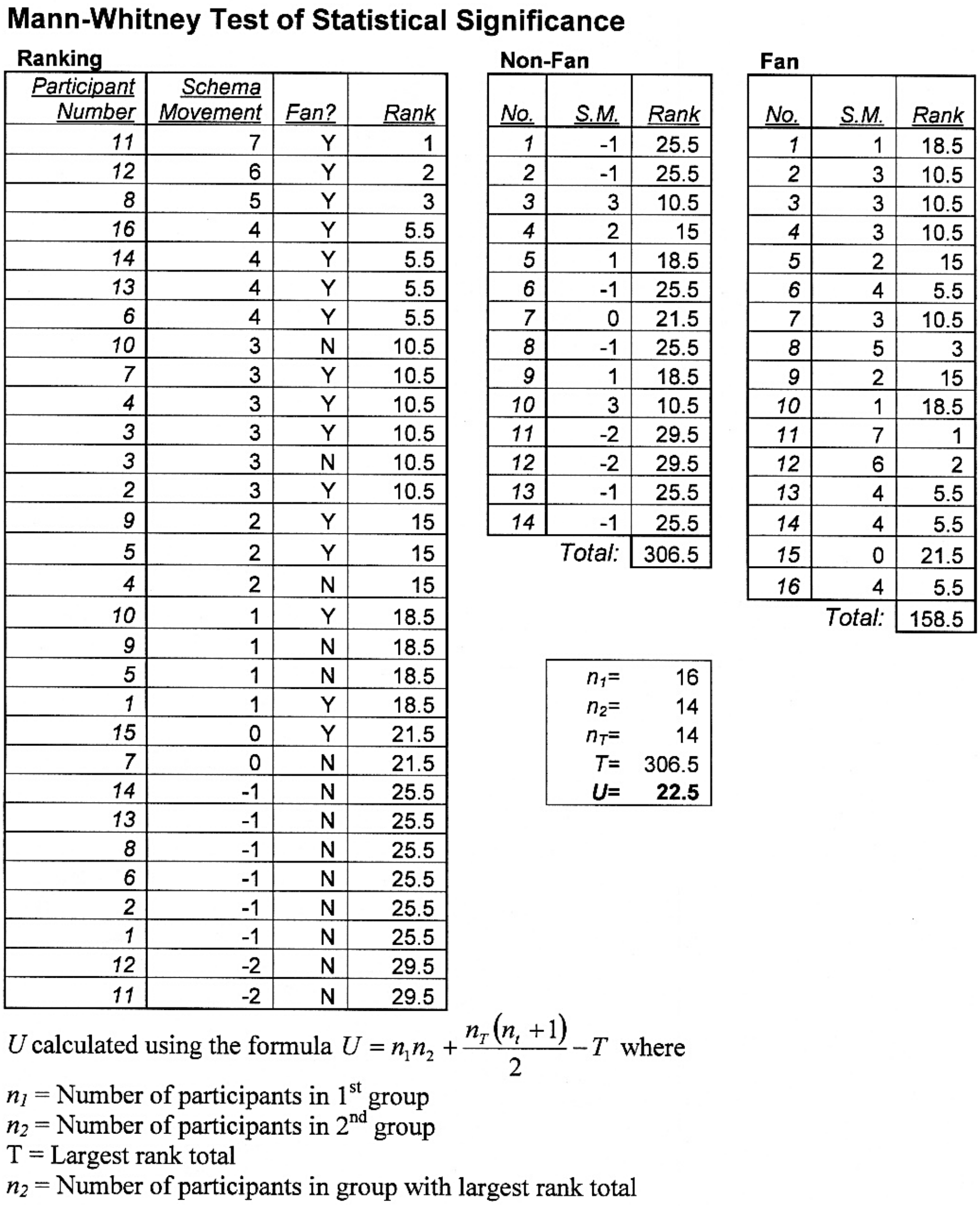
Non-DW Fans

Participant number	Towards Doctor Who	Away from Doctor Who	Schema Movement
1	0	1	-1
2	0	1	-1
3	3	0	3
4	2	0	2
5	1	0	1
6	0	1	-1
7	1	1	0
8	0	1	-1
9	2	1	1
10	3	0	3
11	0	2	-2
12	0	2	-2
13	0	1	-1
14	0	1	-1
Total	12	12	0
Mean	0.86	0.86	0.00
Mode	0	1	-1
Median	0	1	-1

Analysis

The method used to apply the Mann-Whitney U test involved the following steps:

- The participant results were sorted into order of schema movement values, from the highest value of 7 in rank position 1, down to the two lowest values of -2 in joint rank positions of 29.5. Where more than one participant obtained the same schema movement value, the rank positions were averaged.
- Totals were obtained for the ranks within the Non-Fan and Fan groups.
- The Mann-Whitney U statistic was calculated using the formula shown below.
- The result of 22.5 was then compared to a statistical probability table for U values.



Mann-Whitney Test of Statistical Significance

Ranking

Participant Number	*Schema Movement*	*Fan?*	*Rank*
11	7	Y	1
12	6	Y	2
8	5	Y	3
16	4	Y	5.5
14	4	Y	5.5
13	4	Y	5.5
6	4	Y	5.5
10	3	N	10.5
7	3	Y	10.5
4	3	Y	10.5
3	3	Y	10.5
3	3	N	10.5
2	3	Y	10.5
9	2	Y	15
5	2	Y	15
4	2	N	15
10	1	Y	18.5
9	1	N	18.5
5	1	N	18.5
1	1	Y	18.5
15	0	Y	21.5
7	0	N	21.5
14	-1	N	25.5
13	-1	N	25.5
8	-1	N	25.5
6	-1	N	25.5
2	-1	N	25.5
1	-1	N	25.5
12	-2	N	29.5
11	-2	N	29.5

Non-Fan

No.	*S.M.*	*Rank*
1	-1	25.5
2	-1	25.5
3	3	10.5
4	2	15
5	1	18.5
6	-1	25.5
7	0	21.5
8	-1	25.5
9	1	18.5
10	3	10.5
11	-2	29.5
12	-2	29.5
13	-1	25.5
14	-1	25.5
	Total:	306.5

Fan

No.	*S.M.*	*Rank*
1	1	18.5
2	3	10.5
3	3	10.5
4	3	10.5
5	2	15
6	4	5.5
7	3	10.5
8	5	3
9	2	15
10	1	18.5
11	7	1
12	6	2
13	4	5.5
14	4	5.5
15	0	21.5
16	4	5.5
	Total:	158.5

$n_1=$	16
$n_2=$	14
$n_T=$	14
$T=$	306.5
U=	**22.5**

U calculated using the formula $U = n_1 n_2 + \frac{n_T(n_t+1)}{2} - T$ where

n_1 = Number of participants in 1st group
n_2 = Number of participants in 2nd group
T = Largest rank total
n_2 = Number of participants in group with largest rank total

In the next student project, we will examine another statistical method that can be used to investigate data sets:

Spearman's Rank Correlation

It often happens that we collect two sets of measurements from a sample (i.e. a pair of measurements from each member of the sample) and that we wish to discover how strongly these two sets of measurements are associated with one another. A data value might, for example, increase or decrease fairly regularly in response to variations in another parameter, as in figure 322.

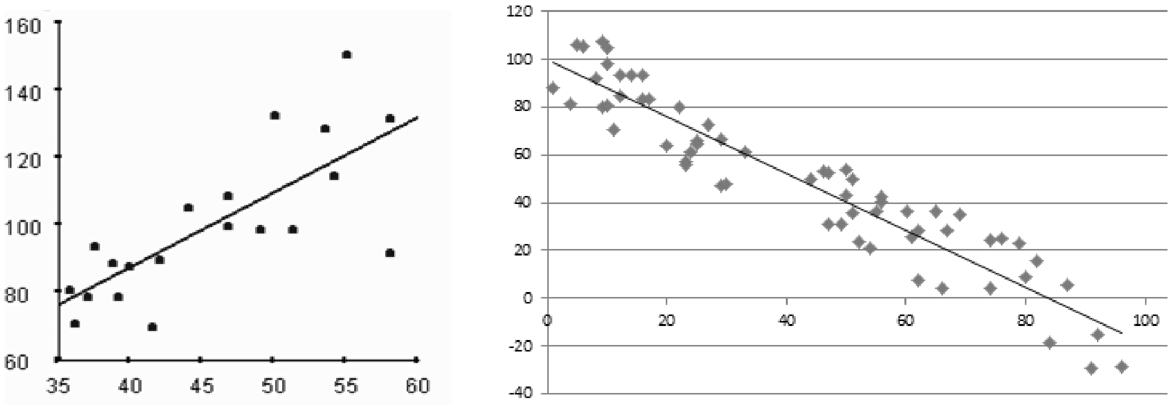


Figure 322: Reasonably strong positive correlation (left) and negative correlation (right)

Spearman's Rank Correlation is a statistical technique for measuring the extent to which data points follow a monotonic relationship. This means that the value of one variable increases as the other variable increases; or the value of one variable increases as the other variable value decreases. A perfect positive correlation is given a coefficient of +1.0, and a perfect negative correlation has a value of -1.0. In practical situations, it is very rare to find perfect correlations. Data values which are unrelated and appear as a completely random distribution of points on scatter graph would have a coefficient close to 0.0.

The Spearman Rank Correlation tests for a correlation, but this need not be linear. The set of data points in figure 323 would have a correlation coefficient of +1.0, since each data point shows a progressive linked increase in both of the variables.

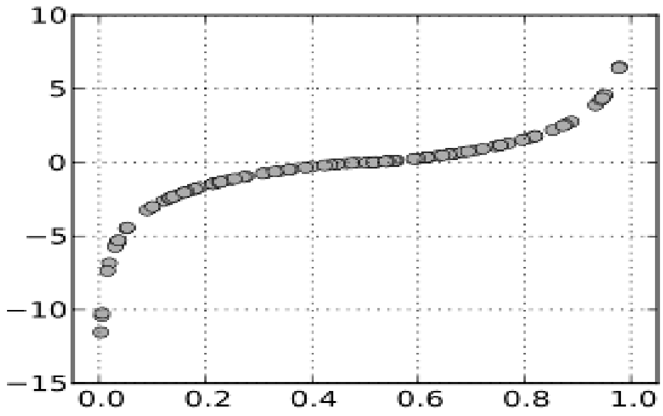


Figure 323: Perfect positive Spearman Rank correlation, although non-linear

To demonstrate the use of Spearman's Rank Correlation in a substantial project, we again present below the report of an experiment carried out by a psychology student:

Connection between stress and illness

Abstract

The hypothesis stated that stress would have a negative effect on students' health and therefore there would be a positive correlation between the two co-variables: high levels of stress and ill-health. This study aimed to test this hypothesis by carrying out a questionnaire on 15 sixth form college students. These students were either 16 or 17, and 11 were female and 4 male. A Spearman's Rank Correlation test was used in order to see if the correlation was significant, which it was (p=0.05, critical value =0.749, observer value = 0.443) So, therefore, the results obtained suggested that the more stressed students get ill more often, suffering from various illnesses. However, there were limitations to this study that may have effected the results, such as all the statements were positively written and some students may not have been comfortable answering some personal questions.

Introduction

Stress is commonly linked with illness. Recent research such as Kielcot-Glaser and Delongis, has increasingly confirmed the importance of stress in mainly cardiovascular disorders and illnesses that occur due to disturbances of the immune system, as a result of stress which may be caused by daily hassles.

When placed in a stressful situation, adrenaline is released by the adrenal medulla which slows down the digestive system in order to conserve energy for flight or fight. This is known as the Sympathetic Adrenal Medullary system. If stress becomes chronic then the Hypothalamic Pituitary Adrenal Axis is activated. This stimulates the pituitary to secrete adrenocorticotrophic hormone (ACTH) which stimulates the adrenal glands to produce a hormone called cortical. This enables the body to cope with the stressors.

According to Seyle's General Adaptation syndrome, prolonged exposure to stressors can have a detrimental effect on the body. The activation of high levels of hormones and shutdown of systems are responsible for most of the damage.

There is increasingly more evidence suggesting that there is a link between stress and cardiovascular disorders. A cardiovascular disorder is defined as any disorder of the heart and circulatory system. Stress can have a direct effect on cardiovascular disorders through the activation of the stress response. This could increase the heart rate which causes blood to pump faster around the circulatory system. Stress can also increase blood pressure, which contributes to the weakening of blood vessels. Stress leads to an increase in the production of glucose and fatty acids which can cause clumps. It can also have an indirect effect by behaviours which individuals may adopt in order to try to combat their stress. These can include sleep deprivation, anxiety and inability to relax.

Krantz et al (1991) found that cardiovascular patients who displayed the greatest myocardial ischemia had the highest increase in blood pressure. This supports the idea that there is a direct link between performing a mildly stressful cognitive task and physiological activity that could damage the cardiovascular system.

Many researchers have researched into stress having an effect on the immune system. The main function of the immune system is to protect the body and does this in three main ways. Firstly, by creating barriers that prevent antigens from entering the body.

Secondly, by detecting and eliminating any antigen that may enter and finally, by eliminating any virus or bacteria which might have started to reproduce.

Kielcot-Glaser et al (1995) aimed to show the direct effects of stress on the immune system by looking at how quickly wounds heal. They found that complete wound healing took longer in the experimental than in the control group. Cytokine levels were also found to be lower in the experimental group and they also indicated feeling more stressed than the control group. This supports the view that stress depresses the function of the immune system.

Delongis et al (1982) believed that chronic everyday strains of living were a more frequent measure of stress than acute events such as bereavement or divorce. They created the 'Hassles and Uplifts Scale' which included 53 possible everyday stressors, such as current affairs and money, as well as events that make you feel good. They compared a life event scale and their own hassle scale to see which was linked more frequently. Participants were asked to complete four questionnaires once a month for a year. These were a hassle scale which included concerns about weight, rising prices, and crime, an uplift scale which included good weather, and relations with friends. There was also a life events questionnaire which included 24 major events and a health status questionnaire which covered overall health status, bodily symptoms and energy levels. There were 100 well educated participants aged between 45 and 64. They found that the frequency and intensity of hassles were significantly correlated with overall health status and bodily symptoms.

Supporting Delongis et al are Kanner et al (1981) who devised a questionnaire entitled the 'Hassle scale' which has been used by several studies. These studies also indicated that there is a positive correlation between hassles and both psychological and physiological symptoms of illness.

Aims

The aim of the study was to see if students who are more stressed suffer from more illnesses than those who aren't. The aim was decided as previous researchers such as Delongis and Kanner believed that daily hassles which causes stress has a negative effect on people's health, and this experiment aimed to see if there is a link between various stressors students may encounter every day in college and the illnesses they may suffer from.

Null hypothesis

There is no correlation between how stressed students are and health problems.

Alternative hypothesis

It is predicted that stress will have a negative effect on students' health and therefore there will be a positive correlation between stress and common health problems. This experiment should obtain a 5% level of significance.

Design

The research method used was a questionnaire. There were two co-variables in this questionnaire which were factors of stress and different illnesses. There was one group of participants who were all given the same questionnaire. A pilot study was carried out on two students. As a result of this pilot study one statement was changed from 'I've had a cold' to 'I suffer from colds'.

Participants

The study was conducted by one further education student. Fifteen participants answered the questionnaire. These participants were chosen by opportunity sampling. These were all students who attend the sixth form college and they consisted of 11 females and 4 males, who were either 16 or 17.

Measures / Apparatus

Fifteen questionnaires that included 11 statements on stress and 11 statements on health were used. A scale from 1 ('Often') to 4 ('Not at all') was used in order to analyse how stressed and how often they suffer from illnesses. All statements were positively written. The questionnaire was designed to measure how often, if ever, the students suffered from various health issues/illnesses since starting college in September. Standardized instructions were used, along with an ethical statement and authorisation.

Procedure

To begin with, a pilot study was carried out on two students to see if the questionnaire was successful and to see if students would understand all statements easily. Both students who carried out the pilot study had difficulty in understanding one statement which was then re-phrased for the actual questionnaire that was used.

The questionnaire was handed out by opportunity sampling to 15 first year students. All were informed that their answers would remain confidential and asked not to put their names on the questionnaire. Once completed, the students were informed that the purpose was to see if there is a link between stress and illness and they were once again told that all answers would be confidential and that they could withdraw them if they wished to. Standardized instructions and an ethical disclaimer were used in order to ensure all students heard the same information.

Controls

Researcher bias was minimised by using standardized instructions when asking participants to fill in the questionnaire.

Examples of the statements on health:

Below is a scale from one to four. Please circle which number relates most to you in connection with the statements below since starting college in September.

1 - Often 2 - Quite often 3 - Not very often 4 - Not at all

- I suffer from headaches.
- I suffer from colds.
- I've had aches and pains in my muscles.
- I suffer from stomach pains.
- I constantly feel tired.
- I feel depressed.

Examples of the statements on stress:

Below is a scale from one to four. Please circle which number relates most to you in connection with the statements below since starting college in September.

1 - Often 2 - Quite often 3 - Not very often 4 - Not at all

- Teachers make me angry.
- I get worried when I think about exams.
- I get annoyed when I have to wake up early for college.
- Other students make me angry.
- Thinking about further education or a career worries me.
- I get upset when I'm late for college.

Results

The tables below show the means and raw scores of both stress and health statements. It is suggested that stress may have had an effect on the participants' health. A scatter graph was used to display the levels of stress and health. This shows that there is a positive correlation between levels of stress in students and how often they suffer from illnesses, with the correlation quite strong.

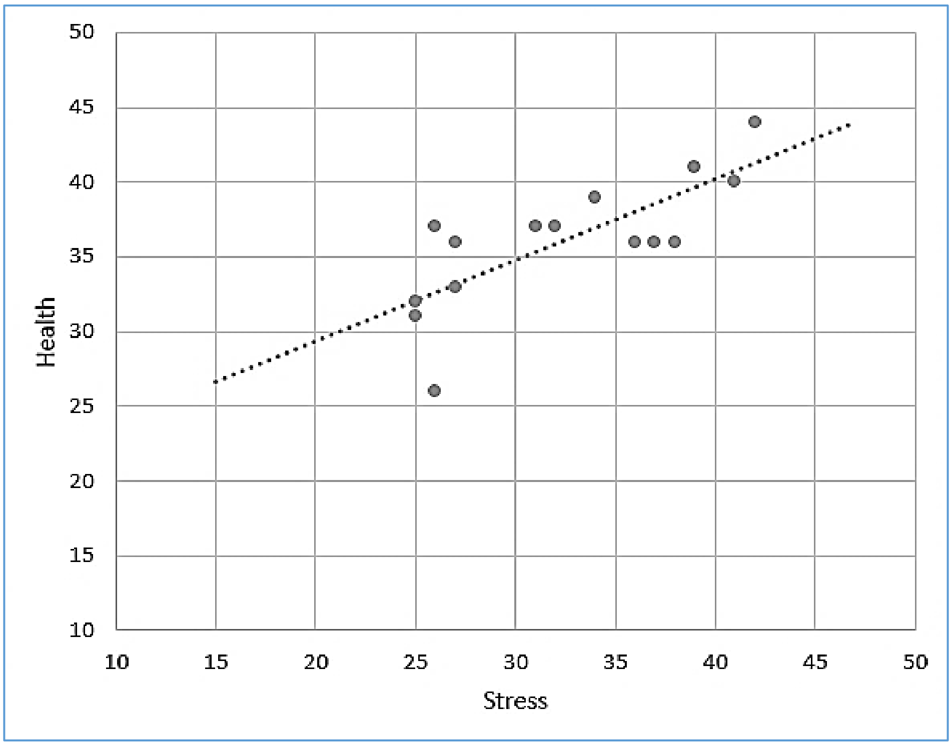
Table 1 - The mean scores of students perceived levels of stress and their health

Statements	Mean
Stress	32.4
Health	36.07

Table 2 - The raw scores of students perceived levels of stress and their health

Participant	Stress	Health
1	32	37
2	42	44
3	39	41
4	27	33
5	37	36
6	38	36
7	27	36
8	31	37
9	34	39
10	41	40
11	25	32
12	26	37
13	25	31
14	26	26
15	36	36

Graph 1 – Students' perceived levels of stress and illness



In order to see if there was a link between stress and illnesses, a Spearman's Rank Correlation Test was used. This was carried out in order to investigate whether the two variables (stress and health) were correlated. The statistical calculations are shown in the Appendix below.

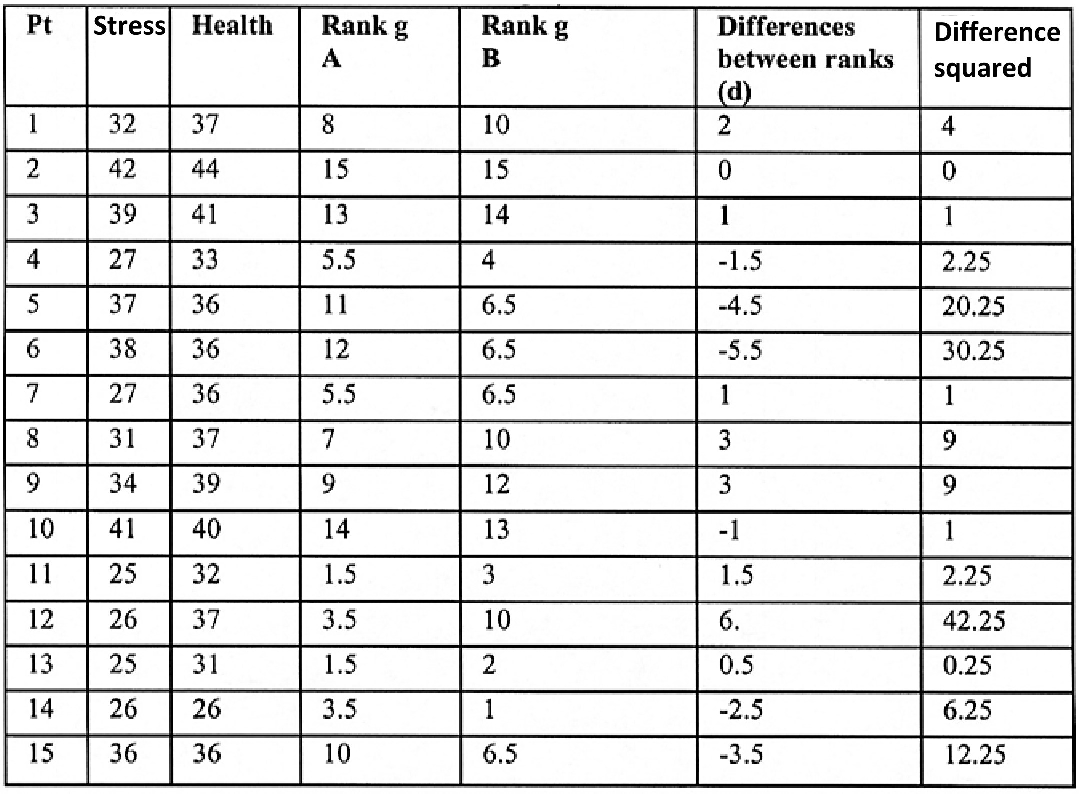
The results found a correlation coefficient of 0.749, which indicates a strong relationship between the two co-variables, stress and illness. The level of significance selected was 5%. The hypothesis was directional, so a one-tailed test was required. The critical value was 0.443. The observed value of r_s of 0.749 is greater than the level of significance of 0.443 with 15 participants (N=15).

As the observer value of rs is greater than the level of significance, the directional hypothesis is retained and the null hypothesis is rejected.

Analysis

The method used to apply the Spearman Rank Correlation involved the following steps:

- The participant results were sorted into order of *stress* and *health* scores, and rank positions allocated for each of these variables. Where more than one participant obtained the same score, the rank positions were averaged.
- Differences in ranks were calculated, by subtracting the *stress* rank from the *health* rank. The differences were then squared



Pt	Stress	Health	Rank g A	Rank g B	Differences between ranks (d)	Difference squared
1	32	37	8	10	2	4
2	42	44	15	15	0	0
3	39	41	13	14	1	1
4	27	33	5.5	4	-1.5	2.25
5	37	36	11	6.5	-4.5	20.25
6	38	36	12	6.5	-5.5	30.25
7	27	36	5.5	6.5	1	1
8	31	37	7	10	3	9
9	34	39	9	12	3	9
10	41	40	14	13	-1	1
11	25	32	1.5	3	1.5	2.25
12	26	37	3.5	10	6.	42.25
13	25	31	1.5	2	0.5	0.25
14	26	26	3.5	1	-2.5	6.25
15	36	36	10	6.5	-3.5	12.25

- The sum of the squares is found:

$$\sum d^2 = 141.0$$

- Spearman's rank correlation coefficient is given by the formula:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

where n is the sample size.

$$r_s = 1 - \frac{6 \quad \times 141.0}{15(15^2 - 1)}$$

$$so \qquad\qquad r_s = 1 - \ 0.251 \quad = \ 0.749$$

Analysis of Variance

A number of statistical tests are available to help determine whether two random samples of data belong to the same population, or to two populations which are different in some significant way. We have looked at examples of projects using the Chi-squared test and the Mann-Whitney U-test. Other methods include the T-test and Z-test.

Sometimes it is necessary to determine whether more than two samples belong to the same or different populations. For example: we may have made measurements of the percentage cover of bracken from sample plots on a number of different hillsides. We might wish to know whether the distribution of the bracken is constant, allowing for expected randomness, or whether there is some definite underlying difference in distribution – perhaps due to differences in animal grazing.

One approach to the multiple comparison problem might be to compare the samples from different hillsides in pairs. However, this can lead to a large number of individual comparisons. For five samples, we would need to carry out ten comparisons. This causes a problem:

> The statistical test cannot tell us ***definitely*** that samples belong to different populations. The test just gives the ***probability*** that the samples belong to different populations.

We generally accept that samples belong to different populations if the probability for this happening is found to be 90% or greater. However, we will occasionally be wrong. When a large number of tests are carried out, there is a high chance of at least one incorrect inference, leading to an incorrect conclusion overall. For comparison of multiple samples, a better approach is to use a technique called **analysis of variance**, which is given the acronym **ANOVA**.

The underlying technique for Analysis of variance is illustrated in figure 324:

- We assume that the different samples have data values which are normally distributed about each mean, so that graphs of the frequencies in each sample would have bell-shapes. However, the ANOVA test is quite robust and will still give valid results if this assumption is not exactly true.
- The test begins by determining the mean and standard deviation for each of the individual sample groups.
- The group mean values are then used to calculate the overall mean and standard deviation for the complete set of data.

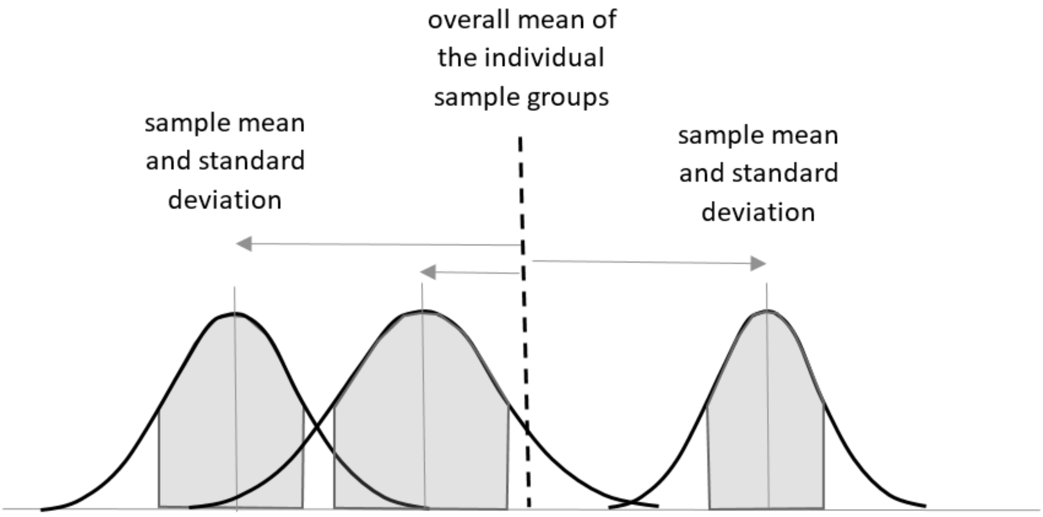

overall mean of the individual sample groups
sample mean and standard deviation
sample mean and standard deviation

Figure 324: Analysis of variance

The Analysis of variance test then makes a comparison between the variance ***within*** the groups and the variation ***between*** groups. Two possibilities for different data sets are illustrated in figure 325, where the box plots represent the mean and standard deviation of each individual sample.

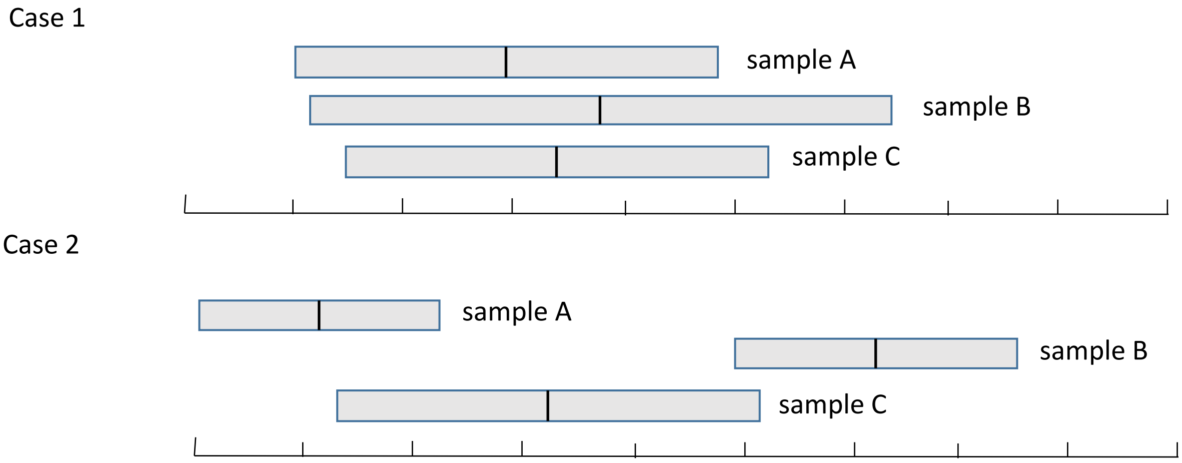


Figure 325: Variance within and between groups

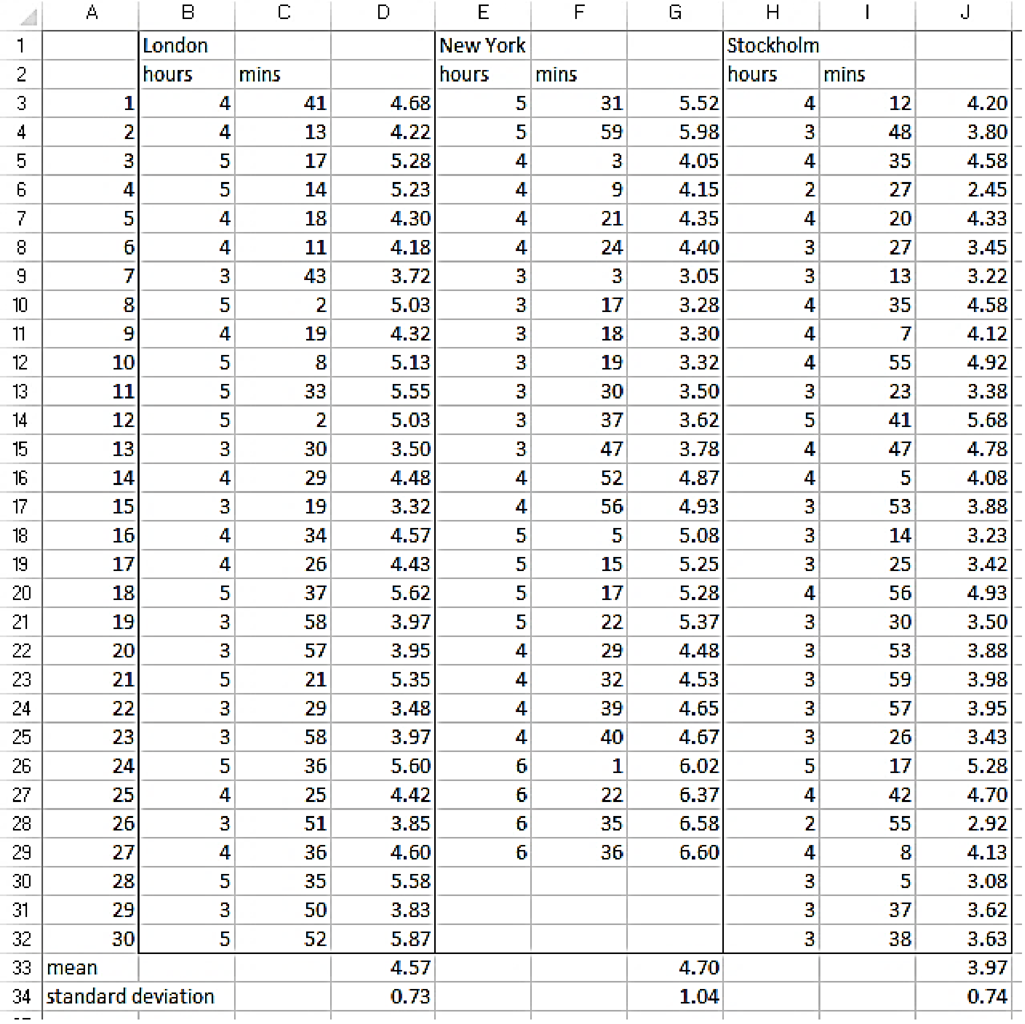
In case 1, the variance within the groups is large, but there is little variance between the mean values of the groups. We might conclude that the samples all belong to the same underlying population.

In case 2, the variance between the mean values of the groups is large in comparison to the variance within groups, so we would conclude that the samples belong to at least two different populations.

The analysis of variance test produces a statistic called the **F ratio**, named after the mathematician Fisher. This is the ratio of how much variability there is ***between*** the groups relative to how much there is ***within*** the groups. The F ratio is then used, along with information about the sample sizes, to find the probability that the samples belong to the same population or different populations.

We will look at a couple of examples of the use of the ANOVA technique. In the first, we examine results for the 2016 marathons held in three different cities: London, New York and Stockholm. It is of interest to see whether the groups of runners taking part in these events are in some way different.

Many thousands of runners competed in each of the marathons. Results have been published for each event, sorted into alphabetical order of the runners' surnames. We might reasonably assume that this list will show competitors in random order of their finishing times. The first page of results for competitors with surnames beginning with the letter 'D' was randomly chosen as data from each of the marathons.



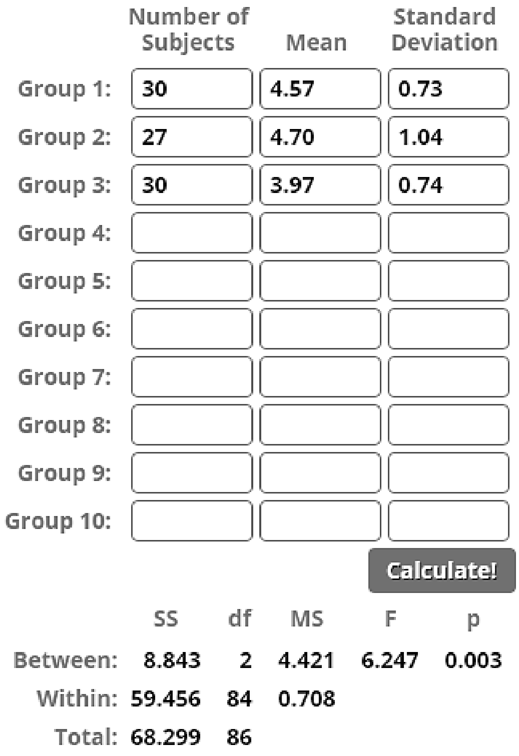
	A	B	C	D	E	F	G	H	I	J
1		London			New York			Stockholm		
2		hours	mins		hours	mins		hours	mins	
3	1	4	41	4.68	5	31	5.52	4	12	4.20
4	2	4	13	4.22	5	59	5.98	3	48	3.80
5	3	5	17	5.28	4	3	4.05	4	35	4.58
6	4	5	14	5.23	4	9	4.15	2	27	2.45
7	5	4	18	4.30	4	21	4.35	4	20	4.33
8	6	4	11	4.18	4	24	4.40	3	27	3.45
9	7	3	43	3.72	3	3	3.05	3	13	3.22
10	8	5	2	5.03	3	17	3.28	4	35	4.58
11	9	4	19	4.32	3	18	3.30	4	7	4.12
12	10	5	8	5.13	3	19	3.32	4	55	4.92
13	11	5	33	5.55	3	30	3.50	3	23	3.38
14	12	5	2	5.03	3	37	3.62	5	41	5.68
15	13	3	30	3.50	3	47	3.78	4	47	4.78
16	14	4	29	4.48	4	52	4.87	4	5	4.08
17	15	3	19	3.32	4	56	4.93	3	53	3.88
18	16	4	34	4.57	5	5	5.08	3	14	3.23
19	17	4	26	4.43	5	15	5.25	3	25	3.42
20	18	5	37	5.62	5	17	5.28	4	56	4.93
21	19	3	58	3.97	5	22	5.37	3	30	3.50
22	20	3	57	3.95	4	29	4.48	3	53	3.88
23	21	5	21	5.35	4	32	4.53	3	59	3.98
24	22	3	29	3.48	4	39	4.65	3	57	3.95
25	23	3	58	3.97	4	40	4.67	3	26	3.43
26	24	5	36	5.60	6	1	6.02	5	17	5.28
27	25	4	25	4.42	6	22	6.37	4	42	4.70
28	26	3	51	3.85	6	35	6.58	2	55	2.92
29	27	4	36	4.60	6	36	6.60	4	8	4.13
30	28	5	35	5.58				3	5	3.08
31	29	3	50	3.83				3	37	3.62
32	30	5	52	5.87				3	38	3.63
33	mean			4.57			4.70			3.97
34	standard deviation			0.73			1.04			0.74

Figure 326: Samples of results for city marathons

Functions within the Excel spreadsheet were used to find the mean and standard deviation for each of the three samples.

A variety of statistical applications are available through the Internet for calculation of ANOVA statistics, both for download and for direct use on-line. A convenient calculator can be accessed on a web page at: www.danielsoper.com/statcalc/calculator.aspx?id=43

To use the ANOVA calculator, the sample sizes, means and standard deviations for our three city marathons are entered.



	Number of Subjects	Mean	Standard Deviation
Group 1:	30	4.57	0.73
Group 2:	27	4.70	1.04
Group 3:	30	3.97	0.74
Group 4:			
Group 5:			
Group 6:			
Group 7:			
Group 8:			
Group 9:			
Group 10:			

	SS	df	MS	F	p
Between:	8.843	2	4.421	6.247	0.003
Within:	59.456	84	0.708		
Total:	68.299	86			

Figure 327: ANOVA result for the city marathon data

Results of the ANOVA test are tabulated:

The column headed SS refers to ***sum of squares***. These values are obtained from the standard deviations which we provided.

The column headed df refers to ***degrees of freedom***. When considering the overall sample, this value is one less than the number of groups, giving a result of 2. When considering the data points within the groups, this is the total number of data values less one for each group. This gives a result of (87 – 3) = 84.

The column headed MS is the ***mean sum of squares***, obtained by dividing the values in the previous two columns, SS and df.

The column headed F refers to the ***F ratio***, which is calculated as:

$$F = \frac{MS\ between\ groups}{MS\ within\ groups}$$

In this case, (4.421 / 0.708) = 6.247

The final column, headed p, indicates the ***probability*** that the sample groups all belong to the same underlying population. This is calculated from the F ratio and the sample sizes.

The result we obtained from analysis of the city marathon data suggests that there is less than a 1% chance that the three data sets belong to the same underlying population. We can investigate this further with a plot of the mean and standard deviation for each of the samples:

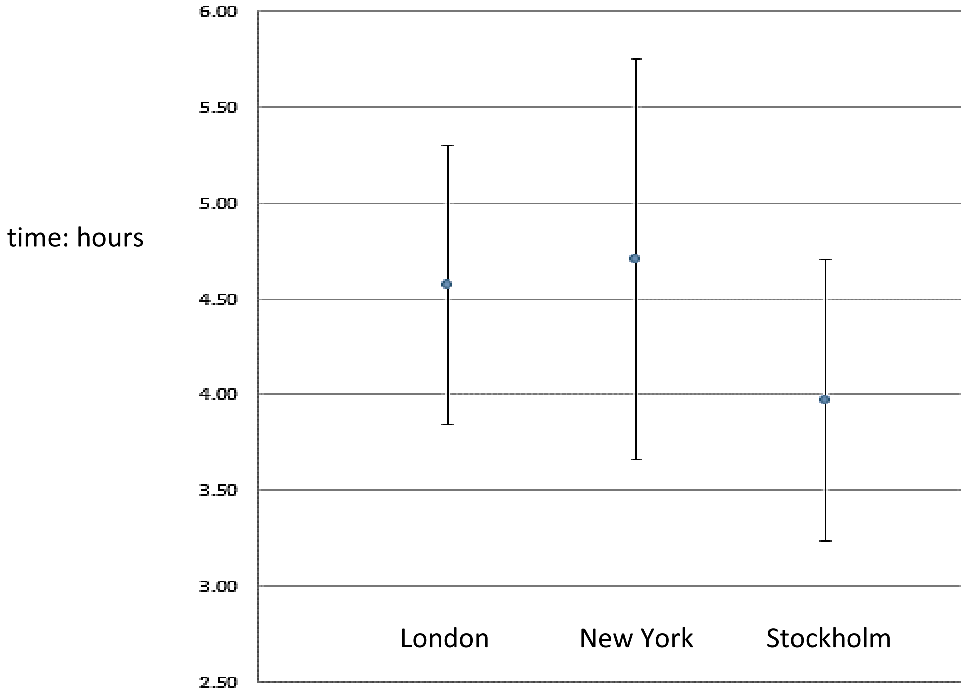


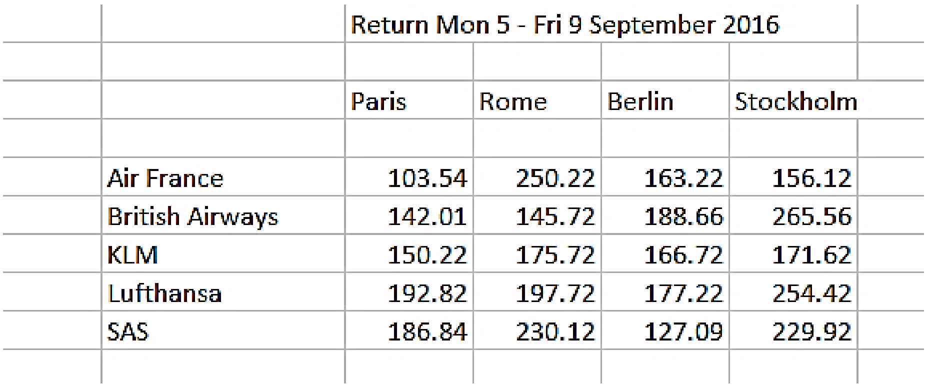
Figure 328: Plot of means and standard deviations for the city marathon data

The plot confirms that the Stockholm marathon has a significantly lower mean than the other two races. The Stockholm marathon might be considered an important sporting event and most competitors are serious athletes. Many competitors in the London and New York marathons, however, may be less serious about sport and are taking part just once as a personal challenge or to raise sponsorship for a charity.

In this example, the three sample groups in London, New York and Stockholm were completely independent. It is very unlikely that any individual runner appeared in more than one of our samples. However, sometimes we deliberately look for corresponding data in each of the comparison groups. We will examine this approach in the next example:

Business students may be interested to investigate the extent to which competition between companies leads to variation or similarity in prices. The table below shows a series of prices quoted by major airlines for return flights from Manchester to four different European cities. The prices were the cheapest fares offered about two weeks before the departure date, with a return flight four days later. The same four destinations have been selected for each airline, so that additional analysis of the result is possible. In this case, an analysis of variance statistical package was downloaded from the website:

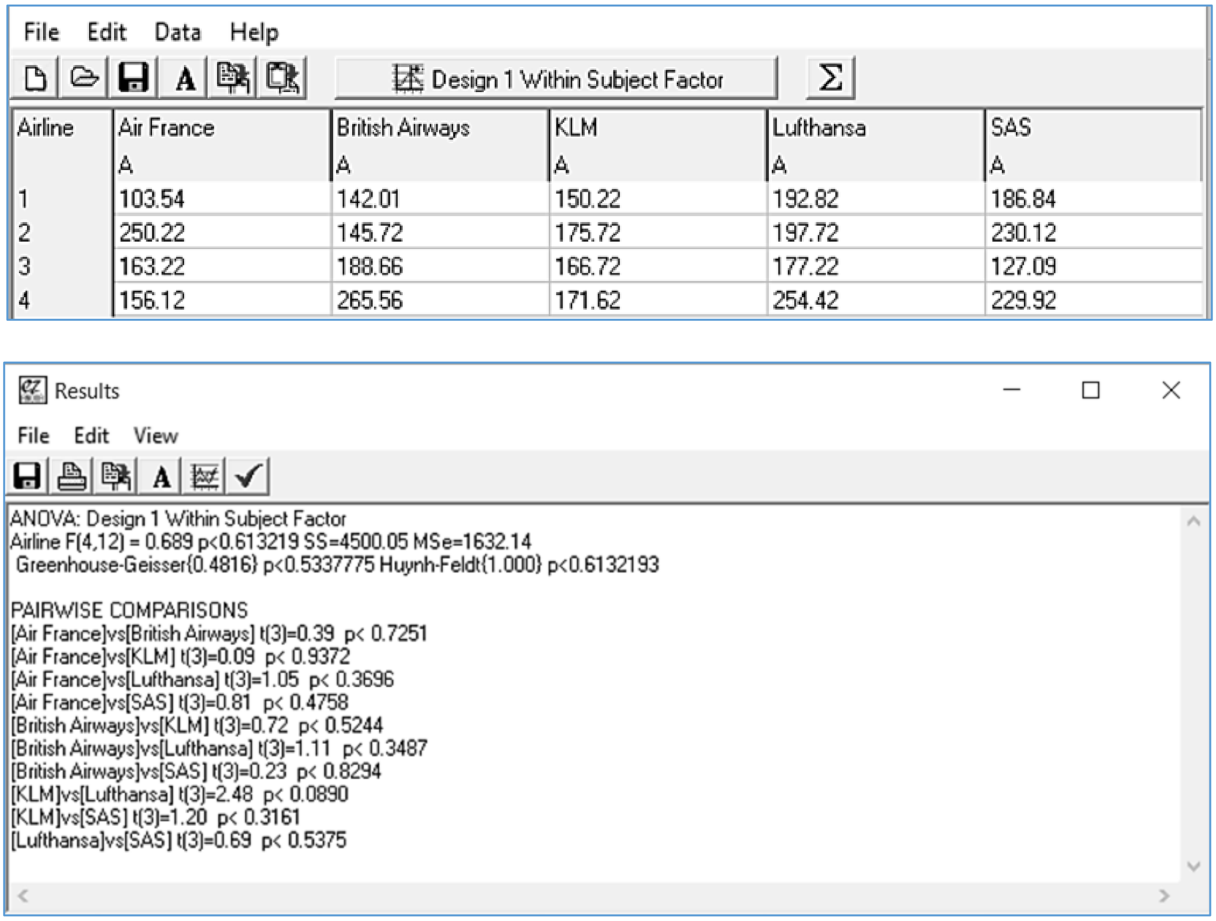
www.cabiatl.com/mricro/ezanova/



	Return Mon 5 - Fri 9 September 2016			
	Paris	Rome	Berlin	Stockholm
Air France	103.54	250.22	163.22	156.12
British Airways	142.01	145.72	188.66	265.56
KLM	150.22	175.72	166.72	171.62
Lufthansa	192.82	197.72	177.22	254.42
SAS	186.84	230.12	127.09	229.92

Figure 329: Samples of air fares from Manchester to four European destinations

After installing the software package, the data values are entered. Processing is then carried out using the 'Within Subject Factor' option which allows us to make further comparisons between the samples.



Airline	Air France	British Airways	KLM	Lufthansa	SAS
	A	A	A	A	A
1	103.54	142.01	150.22	192.82	186.84
2	250.22	145.72	175.72	197.72	230.12
3	163.22	188.66	166.72	177.22	127.09
4	156.12	265.56	171.62	254.42	229.92

```
ANOVA: Design 1 Within Subject Factor
Airline F(4,12) = 0.689 p<0.613219 SS=4500.05 MSe=1632.14
 Greenhouse-Geisser(0.4816) p<0.5337775 Huynh-Feldt(1.000) p<0.6132193

PAIRWISE COMPARISONS
[Air France]vs[British Airways] t(3)=0.39  p< 0.7251
[Air France]vs[KLM] t(3)=0.09  p< 0.9372
[Air France]vs[Lufthansa] t(3)=1.05  p< 0.3696
[Air France]vs[SAS] t(3)=0.81  p< 0.4758
[British Airways]vs[KLM] t(3)=0.72  p< 0.5244
[British Airways]vs[Lufthansa] t(3)=1.11  p< 0.3487
[British Airways]vs[SAS] t(3)=0.23  p< 0.8294
[KLM]vs[Lufthansa] t(3)=2.48  p< 0.0890
[KLM]vs[SAS] t(3)=1.20  p< 0.3161
[Lufthansa]vs[SAS] t(3)=0.69  p< 0.5375
```

Figure 330: Air fares data entry and analysis

As previously, an F-ratio and probability result is calculated for the overall set of data. The probability value of 0.61 indicates that there is a 61% chance that the sets of air fares belong to the same underlying population. We might take this as an indication that competition is working quite well in ensuring that fares offered by competing airlines are broadly similar. However, we might want to investigate more deeply to see whether individual differences exist. A plot of mean and standard deviation for the samples reveals some differences. Air France and British Airways seem to have a much larger range in fares than KLM, and the mean price of Lufthansa flights is higher.

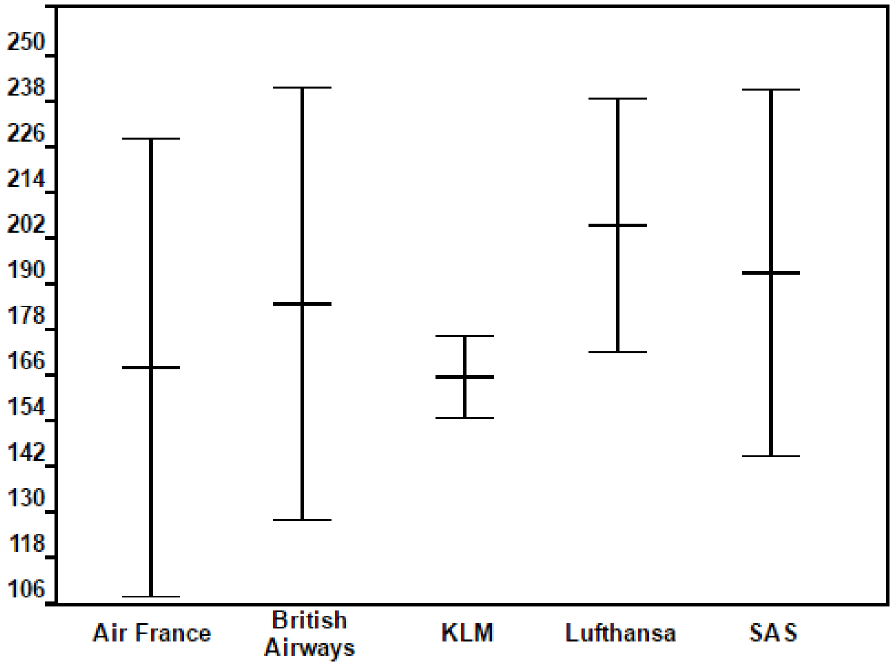


Figure 331: Plot of mean and standard deviation of samples of air fares

If we refer again to the results of the ANOVA test in figure 330, we see a section headed PAIRWISE COMPARISONS. In this section, a test is carried out to determine the probability that each pair of airlines has fares belonging to the same underlying population. In some cases, the probability values are quite high:

British Airways and SAS	0.83
Air France and British Airways	0.73
Air France and KLM	0.94

We might conclude that these airlines are closely following each other's fare structures. In one case the probability value is very low:

KLM and Lufthansa	0.089

These two airlines appear to have very different prices, and perhaps operate different business models.

Overall, Analysis of Variance can provide a valuable statistical method for extending the range of sampling strategies which are possible during numeracy projects.