Unit Support Notes

Unit Support Notes — Statistics (SCQF level 6)

Contents

Introduction

These support notes are not mandatory. They provide advice and guidance on approaches to delivering and assessing this Statistics Unit. They are intended for teachers and lecturers who are delivering this Unit. They should be read in conjunction with:

- the *Unit Specification*
- appropriate assessment support materials

General guidance on the Unit

Aims

The general aim of this Unit is to develop skills that focus on the use of statistical ideas and valid strategies that can be applied to managing statistics in real-life contexts which may be new to the learner. This includes skills in interpreting and analysing graphs and statistical diagrams, applying skills to the normal distribution, applying statistical skills to data analysis, interpretation and communication and determining the equation of linear regression and using it for prediction.

The Outcomes cover aspects of statistics in real-life situations and also skills in statistical reasoning and modelling.

Learners who complete this Unit will be able to:

- Use reasoning skills and statistical skills linked to real-life contexts

In addition, learners will have the opportunity to develop generic and transferable skills for learning, skills for life and skills for work. These include numeracy and thinking skills.

Progression into this Unit

Entry into this Unit is at the discretion of the centre. However, learners would normally be expected to have attained the skills, knowledge and understanding required by one or more of the following or equivalent qualifications and/or experience:

- National 5 Courses or equivalent in a wide range of curricular areas; including Social Subjects, Sciences, Business and Mathematics.

Prior learning, life and work experiences may also provide an appropriate basis for entry into this Unit. This could include relevant skills, knowledge and understanding and appropriate experiences and outcomes from the statistics curriculum area.

Centres wishing to establish the suitability of learners without prior qualifications and/or experiences and outcomes may benefit from carrying out a diagnostic review of prior life and work experiences. This approach may be particularly useful for adults returning to education.

Skills, knowledge and understanding covered in the Unit

Teachers and lecturers are free to select the skills, knowledge, understanding and contexts which are most appropriate for delivery of this Unit in their centres.

Progression from this Unit

This Unit may provide progression to:

- other qualifications in statistics or related areas
- further study, employment and/or training

The Statistics Unit has applications in a variety of subject areas as well as life and work. The skills, knowledge and understanding developed in this Unit could support both breadth and depth of learning in other curriculum areas such as business, science, social studies and health and wellbeing, in addition to life and work contexts.

Approaches to learning and teaching

The purpose of this section is to provide general advice and guidance on approaches to learning and teaching for this Unit.

Effective learning and teaching will draw on a variety of approaches to enrich the experience of learners. In particular, a mix of approaches which provide opportunities for personalisation and choice will help to motivate and challenge learners. Some of these approaches include: interdisciplinary learning, cross-curricular approaches, investigative and problem solving approaches and resource based learning and e-learning.

Examples of how these approaches could be used to combine and integrate the learning and teaching of this Unit are outlined below.

Learners will engage in a variety of learning approaches and activities as appropriate, for example:

- using active and open-ended learning activities such as research, case studies, project-based tasks and presentation tasks
- using real-life contexts and experiences that are familiar and relevant to young people, to meaningfully hone and exemplify skills, knowledge and understanding
- making use of the internet to draw conclusions about specific issues
- recording, in a systematic way, the results of investigation from different sources
- communicating findings/conclusions of research and investigation activities in a presentation
- participating in group work with peers and using collaborative learning opportunities to develop teamworking skills
- a mix of collaborative, co-operative or independent tasks which engage learners
- develop problem solving and critical thinking
- presenting strategies and solutions to others
- use of questioning and discussion to engage learners in explaining their thinking and checking their understanding of fundamental concepts
- making links in themes which cut across the curriculum to encourage transferability of skills, knowledge and understanding — including with technology, geography, sciences, social subjects, mathematics, lifeskills mathematics and health and wellbeing
- using written and/or oral communication and presentation skills to present information
- using appropriate technological resources (eg web-based resources)
- using appropriate media resources (eg video clips)

Outcome 1

Statistics (SCQF level 6)	
1.1 Applying statistical literacy skills to data	
The sub-skills in the Assessment Standard are: ♦ understand: — types of data — the importance of random sampling — outliers ♦ interpretation of, for example: stem and leaf, frequency tables, pie charts, bar charts, diagrams, box plots, contingency tables and histograms	Learning and Teaching Contexts See Exemplar on page 10
1.2 Applying statistical skills to normally distributed data	
The sub-skills in the Assessment Standard are: ♦ interpretation of histogram to indicate the distribution of data ♦ sample measures of location and dispersion including mean, median, standard deviation and interquartile range ♦ emphasis on understanding and interpretation of the above	Learning and Teaching Contexts See Exemplar on page 10
1.3 Applying statistical skills to correlation and linear regression	
The sub-skills in the Assessment Standard are: ♦ scatter plots ♦ perform simple linear regression ♦ interpret the slope and intercept parameters in relation to data ♦ use the linear model for prediction ♦ assess the accuracy of predictions ♦ understand and interpret correlations ♦ understand the applicability of Pearson's correlation coefficient ♦ explore trends in data, eg seasonality	Learning and Teaching Contexts See Exemplar on page 10

1.4 Applying statistical skills to data analysis, interpretation and communication	
The sub-skills in the Assessment Standard are: ♦ interpret and report the results of a hypothesis test ♦ understand and interpret confidence intervals ♦ perform simple analysis using t-tests and paired t-tests ♦ use z-tests to compare two proportions ♦ understand how errors can arise in statistical testing	Learning and Teaching Contexts See Exemplar on page 10

Outcome 2

Statistics (SCQF level 6)	
2.1 Undertaking a correlation and regression analysis	
The sub-skill in the Assessment Standard is: ♦ accessing a given data set and using a software package to assess and model linear relationships using simple correlation and regression modelling	Learning and Teaching Contexts Possible contexts could be to investigate if there is a relationship between: ♦ crime levels and number of police officers on the streets ♦ the distance a golf ball will travel when hit with a golf club at a certain speed ♦ age and mean level of worry about being a victim of crime ♦ temperature and plant growth See further exemplar material on page 10
2.2 Undertaking a data analysis	
The sub-skill in the Assessment Standard is: ♦ accessing a given data set to test a hypothesis, to determine the significance of the test results and communicate the findings from the hypothesis test	Learning and Teaching Contexts Hypotheses could include: ♦ is soil moisture content the same in two sets of random soil samples? ♦ is there a better financial return on Investment A compared to Investment B? ♦ allowing hospital patients more sense of control influences their recovery rates See further exemplar material on page 10

Teachers and lecturers should encourage learners to use an enquiring, critical and problem-solving approach to their learning. Learners should also be given the opportunity to practise and develop research and investigation skills and higher order evaluation and analytical skills. The use of information and communications technology (ICT) can make a significant contribution to the development of these higher order skills as research and investigation activities become more sophisticated.

A calculator or equivalent technologies may be used. A suitable software package, eg Excel, Minitab, SMS, 'R,' SPSS must be used to complete Outcome 2.

Some learning and teaching activities may be carried out on a group basis and, where this applies, learners could also receive feedback from their peers.

Teachers and lecturers should, where possible, provide opportunities to personalise learning and enable learners to have choices in approaches to learning and teaching.

Teachers and lecturers should also create opportunities for, and use, inclusive approaches to learning and teaching. This can be achieved by encouraging the use of a variety of learning and teaching strategies which suit the needs of all learners. Innovative and creative ways of using technology can also be valuable in creating inclusive learning and teaching approaches.

There may be opportunities to contextualise approaches to learning and teaching to Scottish contexts in this Unit. This could be done through mini-projects or case studies.

Developing skills for learning, skills for life and skills for work

For this Unit there are significant opportunities to develop the following skills for learning, skills for life and skills for work. Some of these opportunities are described in the table below:

Skills for learning, skills for life and skills for work	Suggested approaches for learning and teaching
Numeracy is the ability to use numbers to solve problems by counting, doing calculations, measuring, and understanding graphs and charts.	Throughout this Unit, learners will have ample opportunities to use number to solve real-life, social sciences, business and STEM-related problems, and work with information through analysis and interpretation, drawing conclusions and making deductions and informed decisions.
Applying Applying is the ability to use existing information to solve a problem in a different context, and to plan, organise and complete a task.	Wherever possible, learners should be given the opportunity to apply the skills, knowledge and understanding they have developed to solve statistical problems in a range of real-life, cross-curricular, social sciences, business and STEM-related contexts. Learners could be encouraged to think about how they are going to tackle problems, decide which skills and processes to use and then carry out the processes to complete the task. To determine a learner's level of understanding, learners should be encouraged to show and explain their thinking at all times. Learners could be encouraged to think creatively to adapt strategies to suit the given problem or situation.
Analysing and evaluating This covers the ability to identify and weigh-up the features of a situation or issue and to use your judgement of them in coming to a conclusion. It includes reviewing and considering any potential solutions.	Wherever possible, learners could be given the opportunity to identify real-life, social sciences, business and STEM-related tasks or situations which require the use of statistics. Learners should be encouraged to analyse the task or situation to decide how it can be addressed and what statistical skills will need to be applied. Learners should also be provided with opportunities to interpret the results of their calculations and to draw conclusions. Conclusions drawn by the learner could be used to form the basis of a model for making future choices or decisions.

There may also be further opportunities for the development of additional skills for learning, skills for life and skills for work in the delivery of this Unit. These opportunities may vary and are at the discretion of the centre.

Exemplar material

The following exemplar material provides examples of ways in which to display and interpret statistical information in different contexts.

Health and fitness research

As part of an ongoing study, a sociology researcher is investigating why people join health and fitness clubs and what club facilities they prefer to use. Data is obtained by asking a random sample of members to complete a questionnaire. (Edited version of dataset from *Social Research Methods* by Alan Bryman — Oxford University Press, see Appendix 2)

A basic guide to accessing 'R' and Excel software is given in Appendix 3.

Outcome 1

- **Random sampling**
  Random sampling means that each member of a 'population' has an equal chance of being selected. Random sampling is a sampling technique where a group of subjects (sample) is selected for study from a larger group (population). The single and most important advantage of random sampling is obtaining an unbiased sample. Statistics can be used to analyse any set of data but it can be difficult or impossible to interpret the results if there is any possible bias in the data. However, obtaining a truly random sample can be difficult and expensive to achieve.

- **Stem and leaf diagrams**
  The stem and leaf diagram presents all of the data and provides an easy introduction into important concepts, such as, location spread, outliers and the shape of distributions.
  A stem and leaf display is shown below for question 10 from the health and fitness club questionnaire ('During your last visit to the Club, how long did you spend on the cardiovascular equipment?')

  Stem-and-leaf of Q 10 N = 50
  Leaf Unit = 1·0

  Time Spent on Cardiovascular Equipment

0	00
1	0
1	56777788
2	0111222223334
2	55667778
3	001222334444
3	78
4	589
6	4

  Key: 3 | 7 represents 37 minutes

The stem and leaf diagram orders the data as shown above.

The outliers can also be seen at each end of the diagram (0 and 64 minutes).

N = 50 is the number of members who responded to the survey.

The stem and leaf diagram format enables the analyst to order data.

- **Frequency tables**
  Frequency tables can be applied to any type of variable (quantitative or qualitative). The example in Figure 1 shows the count and percentage values of the main reasons why members visit the Health Club. For example 40% of members surveyed visit the Health Club primarily to lose weight.

Figure 1

Aim	Count	Percent
Relaxation	3	6.00
Improve Fitness	16	32.00
Lose Weight	20	40.00
Meet people	1	2.00
Build Strength	10	20.00
N=	50	

Pie charts (Figure 2) and bar graphs (Figure 3) are easy to interpret when working with nominal and ordinal variables. They are particularly useful for emphasising how each category compares in size with the other categories.

The pie chart and bar graph below illustrate the main reasons for visiting the Health Club (counts in the bar graph and percentages in the pie chart).

Area represents frequency in 2D charts.

Figure 2

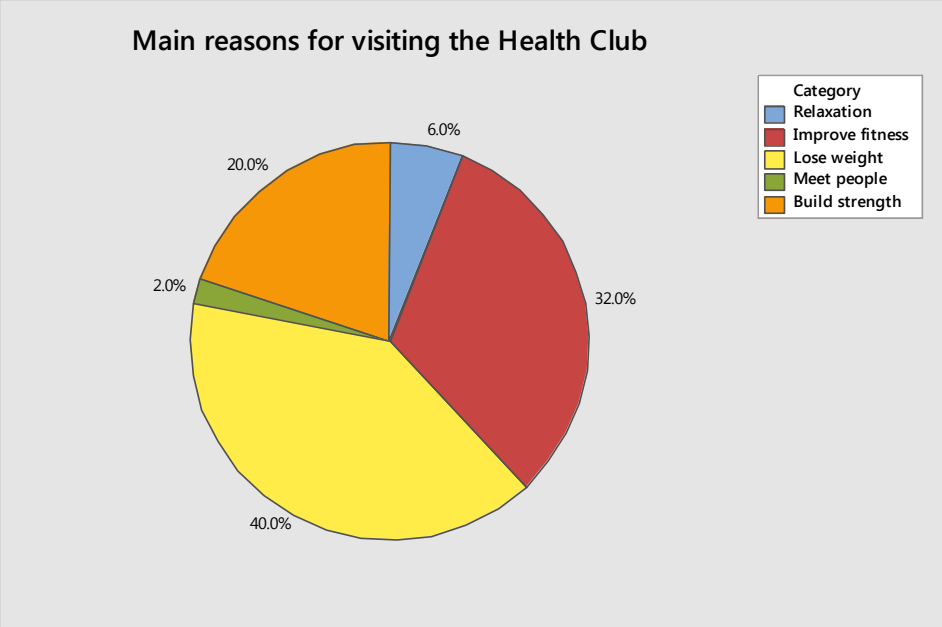
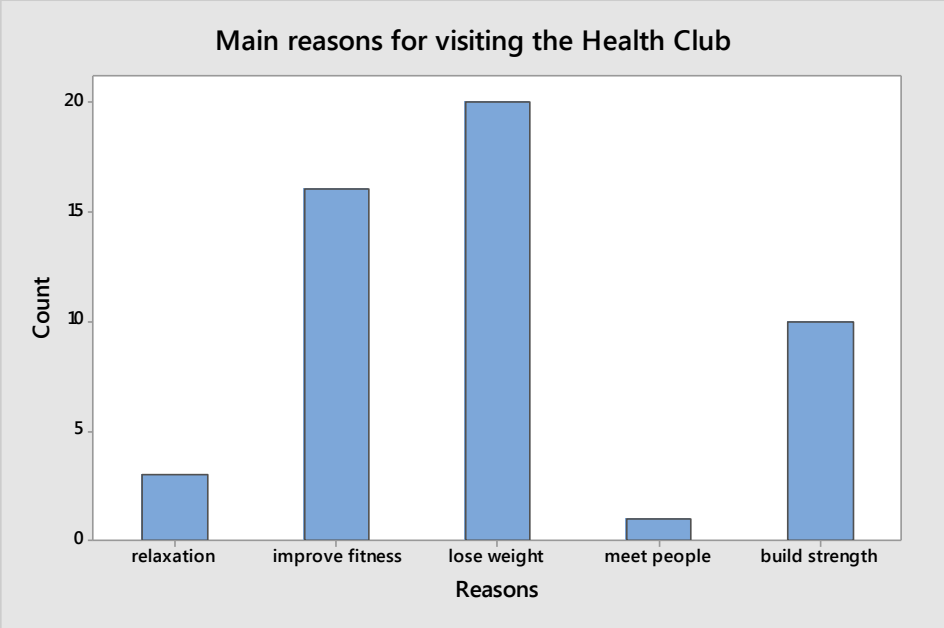


Figure 3



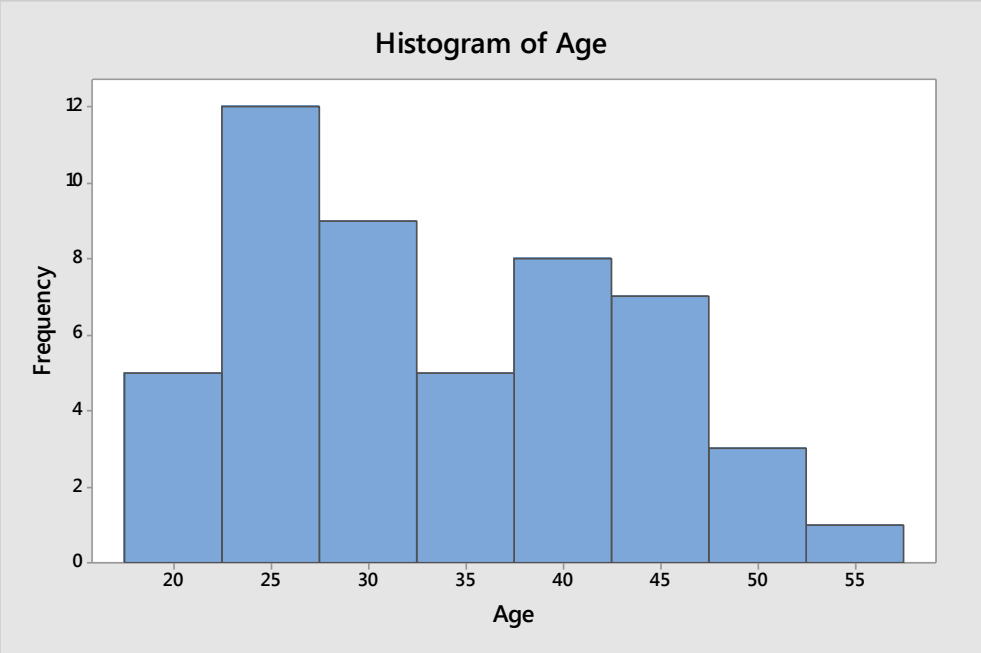
- **Contingency tables**
  A contingency table is essentially a display format used to analyse and record the relationship between two or more categorical variables. A contingency table is like a frequency table but it allows the relationship between two variables to be summarised.
  The contingency table in Figure 4 examines the variables, gender and aim. For example, 3 females (6%) from the sample attend the Health Club for relaxation.

Figure 4

Aim		Male	Female	All
Relaxation	Number	0	3	6%
	Percentage		6%	
Improve fitness	Number	8	8	
	Percentage	16%	16%	32%
Lose weight	Number	4	16	
	Percentage	8%	32%	40%
Meet people	Number	1	0	1
	Percentage	2%	0%	2%
Increase Strength	Number	10	0	10
	Percentage	20%	0%	20%
N=50				

The histogram in Figure 5 shows the distribution of ages visiting the Health Club. The histogram is useful for illustrating the shape of the distribution. The histogram is 'bimodal' — double peaked.

Figure 5



Analysing a histogram

We are trying to infer something about the distribution in the underlying population from which this is a sample, and so sample size will play a part. Some effects are simple, eg symmetry and skewness, and can often be reliably inferred from a sample even of fifty (as in this example). Modality, ie how many peaks, is a more detailed and complex aspect of the shape and correspondingly requires more data to reliably infer anything useful. For example, the histogram above is clearly bimodal, but the sample size is small, and so the shape we see here could be an artefact of how the intervals for the histogram were chosen, ie if we choose a different number of intervals, or slightly different intervals, eg centred on 22, 27, 32 will we see the same bimodality? We may, but it is unlikely in this example. The point is we have a small sample (N=50 is small) for this level of interpretation. So, in practice, we'd register the appearance of bimodality and make some comment on possible interpretation, but we could not deduce that it held for the entire population of all members of this club. So, we'd note that bimodality is suggested, but sample size is small for a reliable interpretation of this; and we'd note that there is clear indication of skewness, with a longer tail into higher values. Notice that the normal distribution has not been mentioned, this is because it is difficult to assert the type of distribution from a histogram. Usually, the best we can say is that the histogram does not obviously suggest that the data are not normal, ie it looks fairly symmetric about a single peak. As regards possible interpretations of such symmetric, unimodal data, it may be that the data are the result of a manufacturing process where a target value is set, and the actual output are mainly close to it, but with a few falling further away, equally on either side, ie symmetrically above and below.

A lot can depend on how (and how many) intervals are chosen, especially when the sample is not huge. This is one reason for caution when making interpretations about shape. On the other hand, the sample size may be fairly large or we might feel the nature of the appearance of bimodality (etc) is convincing. In that case, we would suspect that this means that there are really two distinct populations here and not one, and try to find a way to explore that, and find out what they are. In the above histogram, if we think the bimodality is valid, we would seek to find an interpretation for the two groups, eg clearly there is a population of younger members (aged around 25) and another population of older members (aged around early 40s). Sometimes, the groups we 'spot' correspond to variables already in the sample (eg gender — although not in this case).

- **Types of data**

  The types of data, and examples of each from the Health Club survey, are shown in the table below.

Type of data	Example from Health Club survey
Quantitative (Numerical) Discrete and continuous	Age — continuous
	Time spent on cardiovascular or weights equipment — continuous
	Gender — discrete
Qualitative (Categorical) Nominal and Ordinal	How often do you use the weights machines? (Ordinal)
	Which of the following are the main reasons for going to the gym? (Nominal)

♦ Measures of location and dispersion

Measures of location are: the mean, the median and the mode.
When data is perfectly normal the mean, median and mode are identical.

However, when data becomes skewed the mean loses its ability to provide the best central location.

The mean is the arithmetic average of the data, calculated by dividing the sum of the data by the number of data points. Thus, if the data are:

2, 2, 1, 5, 3, 4, 2, 3, 2, 3 then the mean value is given by:
$(2+2+1+5+3+4+2+3+2+3)/10 = 2\cdot 7$.

The mean is a model of all data within the data set. It takes account of every entry in the data set and is, therefore, susceptible to the influence of outliers.

The median is the 'middle' value when the data are written in order of increasing value. For the sample above, the data are written in order as:

1, 2, 2, 2, 2, 3, 3, 3, 4, 5.

If there were an odd number of data points then there would be a unique middle value in the sequence, but there are 10 data points so there is no unique middle value. In that case the median is the average of the two middle values, ie the average of the fifth and sixth data points $(2+3)/2 = 2\cdot 5$.

Which of these two calculations is better?
In general, the mean is better because it is less variable than the median. However, note that because it is an arithmetic function of all the data, it can be seriously affected by unusually large or small values in the sample. To see this, consider the sample above with one additional observation: 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 50.

Now, the mean is 77/11 = 7, and it is difficult to see how a value which is greater than 10 of the 11 observations can be regarded as a useful measure of central tendency.

On the other hand, the median of these data is the middle value, ie the sixth observation, which is 3, and this is clearly still a useful measure of central tendency for this sample. We say that the median is robust against the effects of extreme values, and we prefer it to the mean in situations where we have skew data which may easily yield extreme high or low data points, or outliers.

Our sample above: 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 50, looks strange, and we may ask if such samples really do occur in practice. In fact, skew distributions are very common, and occur in many contexts.

Skewness simply implies one tail much longer than the other, common in data which are essentially positive, eg times and income. Histograms for salaries in the West

(and world, generally) are always skew, having incredibly long tails into high values. The difference between mean and median incomes is huge (and so it's important in practice to know which is being used). Often too, survival times after the diagnosis of a condition, or after an operation, will be very skewed.

Famously, the palaeontologist Stephen Jay Gould was diagnosed with a rare but dangerous cancer in 1982. At the peak of his career and with a young family he checked out the known statistics and found it was incurable with a median time to death after diagnosis of eight months. After the shock, he searched further and found that the distribution of this survival time has a very long tail, and, in fact, he died in 2002, 20 years later, from a different and unrelated cause.

However, the best examples are to do with money. We need only think of UK salaries: the median is, approximately, £20,000, maybe slightly less, but we all know about all the 'big earners' from University principals earning an annual salary £200–300,000, to bankers on several millions (and the mean is about £25,000, because of this very long tail). The point, for comparison of mean and median, at least, is that a small number of very huge numbers can raise an average substantially.

The same sensitivity of the mean can be seen in the presence of extreme low values. For example, the average age at death in one African country is very low, 32 years. In fact, 32 is not a common age for death there. Indeed, most who live that long will live a lot longer. The problem is that when we consider ages at death we find that the country has a shockingly high rate of infant mortality and so there are a lot of values which are very close to zero, and so they bring the average down.

The mode is a third measure of location which is sometimes useful. It is the most frequent entry in the data. Thus, in the sample above: 1, 2, 2, 2, 2, 3, 3, 3, 4, 5, the mode is 2 and seems to be a useful measure in this case. Further, the mode of the extended sample (with the observed value of 50) is still 2, so it appears to be robust against extremes like the median.

However, when we examine real data in practice, we find that there are usually very many numbers, most of which are different, ie there are few repetitions and they usually have nothing to do with location or central tendency.

The mode tends to be useful for data which takes only a small number of distinct values and we are interested in the most commonly occurring value. It is often useful for categorical data where we simply wish to specify the commonest category.

Guide to measures of central tendency

Variable	Measures of central tendency
Nominal	Mode
Ordinal	Median
Interval — not skewed	Mean
Interval — skewed	Median

Measures of dispersion include the range, the standard deviation and the inter-quartile range.

The range is the difference between the maximum and minimum values in a set of data.

For the data generated by Question 10 ('How long did you spend on cardiovascular equipment during your last visit to the Health Club?') the range would be 64 minutes minus 0 minutes resulting in a range of 64 minutes.

The standard deviation is the average amount of variation around the mean. There is a mathematical formula used to calculate the standard deviation but essentially the standard deviation is another **measure of spread** of the data.

The inter-quartile range is obtained by ordering the data in ascending order and then dividing the data into quarters. A simple example of calculating the inter-quartile range is as shown below.

For the list of numbers: 2, 3, 4, 4, 6, 7, 8, 27

	$Q_1 = 3{\cdot}5$		$Q_2 = (4+6) \div 2 = 5$			$Q_3 = 7{\cdot}5$	
2	3	4	4	6	7	8	27

The first quartile $Q_1 = 3{\cdot}5$
The second quartile, the median, $Q_2 = 5$
The third quartile $Q_3 = 7{\cdot}5$

The range for the above list of numbers is 27 – 2 = 25. This is not very representative of the list of numbers as 27 appears to be an outlier. To obtain a more accurate measure of the range we can calculate the inter-quartile range.

Inter-quartile range: $Q_3 - Q_1$: $7{\cdot}5 - 3{\cdot}5 = 4$.
This answer gives a more reasonable indication of the dispersion in this list of numbers than calculating the full range, 27 – 2 = 25.

Learners often use the term ‘spread without fully understanding it. They often feel they know what to say about the mean and median, eg whether a value is big or small, but when they think about spread it’s much less clear. They will often ask of a particular value of standard deviation, ‘Is this big?’ In fact they often have difficulty in fully understanding the means, but often just have an insight into the values.

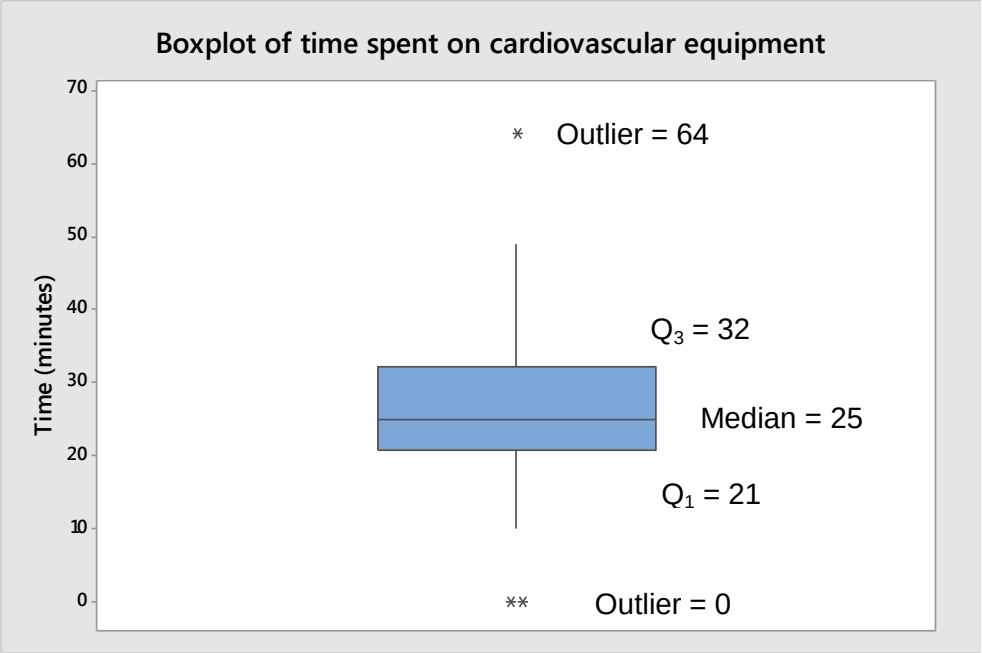
For example, if the data are men’s heights, we know a lot about such data intuitively, and so we can easily interpret mean and median heights. Learners need to know that the different measures, whether mean and median (or standard deviation, interquartile range and range) actually measure different aspects of location (or scale) and so are not comparable. Also, we compute them to compare with similar measures of other data sets. We rarely ever quote and interpret a single measure; we usually compare it with the same measure of another data set, eg we might compare the interquartile range values for men’s and women’s heights and deduce that one is more (or less) variable than the other.

The two most important measures of spread are standard deviation and interquartile range. The standard deviation is based on the mean and the interquartile range is derived from the ordered sample, so these two behave like the mean and median, respectively.

Thus, we would prefer the standard deviation, but if the data are very skewed or we suspect the presence of extreme values we may prefer interquartile range which will be robust against the effect of extreme values. The range is useful and easy to compute, but uses so little of the data, and only the most extreme values, that it is not nearly as reliable as either of the other two.

- **Boxplots**
  A boxplot provides a visual representation of the measures of dispersion. The boxplot in Figure 6 below shows the range, median and quartiles for question 10 (see Appendix 2).

Figure 6



- **Outliers**

  An example of an outlier can be seen in the data set in Appendix 2, in answer to the question (Q 10): 'During your last visit to the Club, how long did you spend on cardiovascular equipment?' Member number 41 answered 64 minutes. This can be seen more clearly in the boxplot shown in Figure 6 above.

  There are several different tests to indicate that a value is a possible outlier. Minitab software identifies outliers on boxplots by labelling observations that are at least 1·5 times the interquartile range ($Q_3 - Q_1$) from the edge of the box. Dixon's test is another means of identifying outliers.

Outcome 2

The following Outcome 2 exemplar material illustrates what information should be extracted from the software output when undertaking correlation and regression and hypothesis testing analyses.

Correlation and regression analysis

(with permission from Dr. David Young, Senior Lecturer, Strathclyde University)

The correlation coefficient should always be interpreted with care since there may be no direct connection between two highly correlated variables.

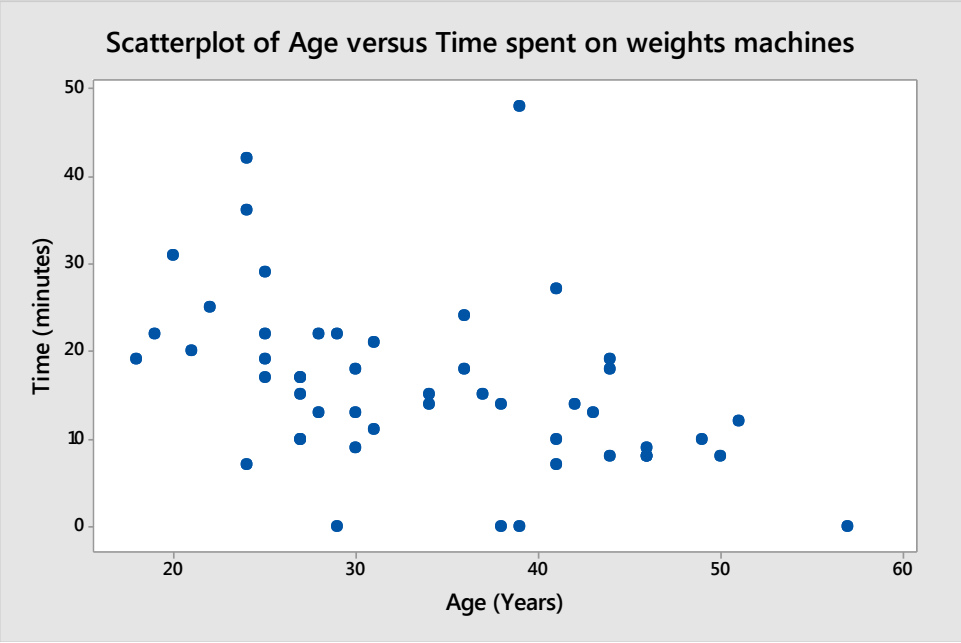
For example, the price of clothing increasing over the last 10 years and growth in mobile phone use would be positive (since both have increased over this period of time). However, there is clearly no causal relationship between these two variables — the increase in clothing prices is not causing the increase in mobile phone usage.

This may appear to be a rather exaggerated example but it is a common mistake in research to make claims about causal relationships between variables based on high correlation values.

Returning to the Health Club data set, a scatter plot can be used to show a relationship between two variables. In Figure 7 the variables are Age and Time spent on weights machines.

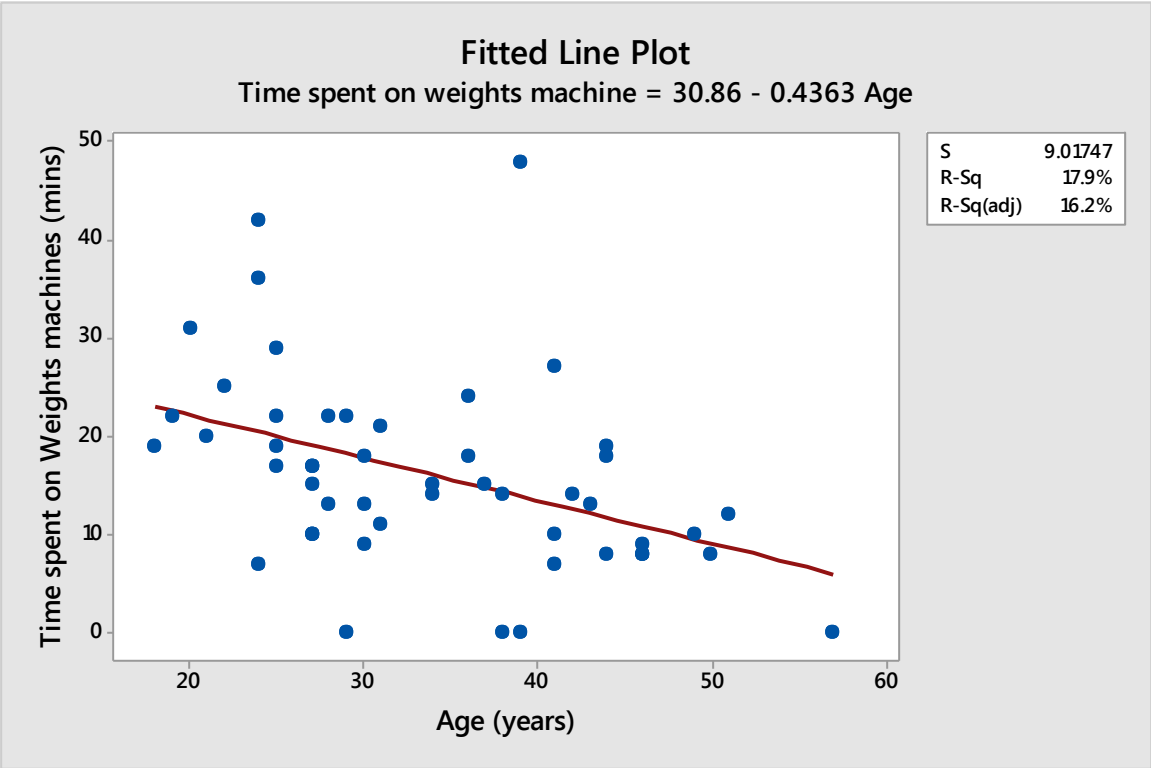
The scatter plot in Figure 7 would suggest a weak negative relationship between Age and Time spent on weights machines, ie as age increases the time spent on using weights machines decreases.

Figure 7



A fitted line can then be added to the scatterplot as shown in Figure 8. This line may then be used to predict future values. However, in this example we will investigate the strength of the linear relationship between the variables before using the line for prediction.

Figure 8



To further explore this relationship, a Pearson's correlation coefficient value can be calculated as shown below.

The Pearson correlation coefficient of age and time spent on weights machines = -0·423.
P-value = 0·002.

The value -0·423 suggests a weak relationship or correlation.

If we square this number a coefficient of determination is obtained: 0·179. If we multiply this number by 100 this gives a percentage of 17.9%.

This means that 17.9% of the variation in the use of weights machines is accounted for by this linear relationship with age. This means that the actual data are scattered widely about the fitted line, ie some are close and some are far away from it.

Therefore, if the fitted line was used to predict further answers then we could not be sure they would be close to the actual values. In other words our confidence intervals for such predictions would be very wide.

Outcome 2

Set up and test a hypothesis

A researcher may have a question regarding their field of study, such as, ‘is there a significant difference in blood pressure between male and female patients?’ or ‘does the use of a partricular fertiliser improve crop yield?’

Other examples could be: a crime reduction initiative claiming that: ‘more police officers on the beat improve crime rates in the area’ or ‘can a person’s level of income can be used to predict who they will vote for in an election?’.

Statistical analysis considers the probability of an event being due to chance. For example, what is the chance that a third child in a family will be female if the first two were male? It is usually not possible to be 100% certain of an event or outcome but mathematically it is possible to say how likely it is to occur. This forms the basis for statistical testing or hypothesis testing.

An hypothesis test begins by posing the question ‘is the observed numerical difference (ie based on sample data) evidence of a real difference (in the underlying populations) or simply a result of the random variation in the population values?’

Example (with permission from Dr. David Young (Senior Lecturer, Strathclyde University)

A study was conducted to compare the effect of two different painkillers on blood glucose levels. Fifteen subjects were given painkiller A, and twelve were given painkiller B, and their blood glucose levels were recorded in mg/kg as shown in the table below. The objective of the study is to determine if the mean blood glucose levels are higher with one painkiller than the other.

A	B
Painkiller A	Painkiller B
40	48
47	60
48	64
51	73
54	76
58	79
64	80
64	84
65	91
67	93
67	*
72	*
78	103
87	*
104	112

Null hypothesis: there is no difference in the mean blood glucose levels achieved by the two painkiller groups

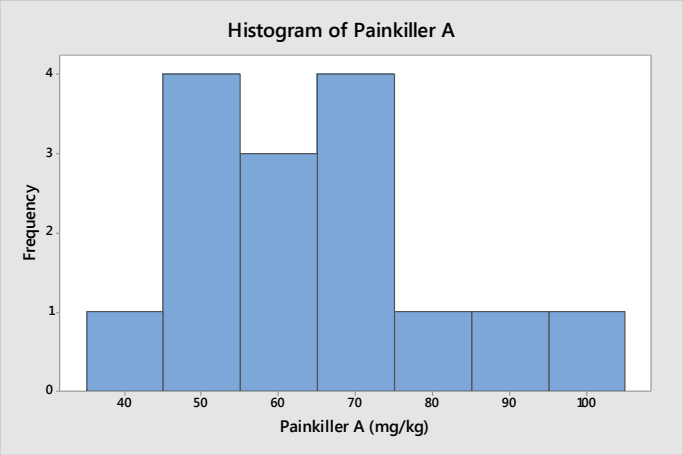
Alternative hypothesis: there is some difference in the mean blood glucose levels achieved by the two painkiller groups

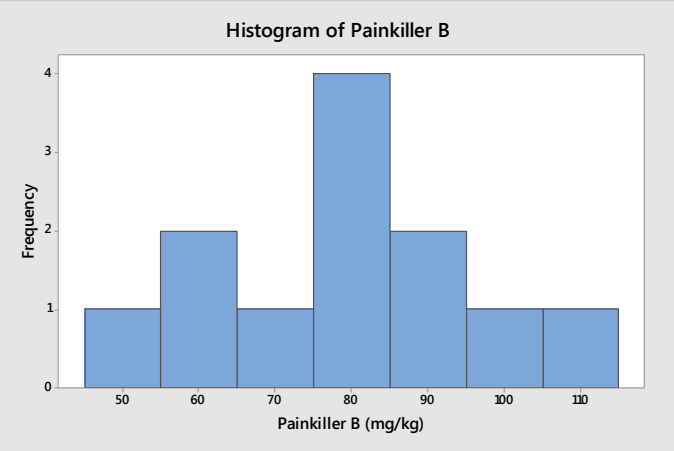
The summary statistics for the blood glucose levels in these two groups is shown below:

Descriptive Statistics: Painkiller A, Painkiller B

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Painkiller A	15	0	64.40	4.27	16.54	40.00	51.00	64.00	72.00	104.00
Painkiller B	12	3	80.25	5.23	18.10	48.00	66.25	79.50	92.50	112.00

Assuming that the data are normally distributed, which is obviously not violated by the mean/median proximity in the table above and the histograms shown below, we can perform a t-test.





From the summary statistics table it would appear that the average blood glucose levels in patients given painkiller A $(64 \cdot 4 \text{mg/kg})$ is lower than in those given painkiller B $(80 \cdot 25 \text{mg/kg})$.

On average therefore, patients given painkiller B have blood glucose levels that are $15 \cdot 85 \text{mg/kg}$ higher than those given in painkiller A.

The hypothesis test will estimate the probability of this happening purely by chance. More specifically, under the null hypothesis (ie that the glucose level in both groups is not different), the hypothesis test estimates the probability that a difference of

$15 \cdot 85 \text{mg/kg}$ would occur in a sample of this size purely by chance (ie simply due to random variability).

Obviously if the probability of such a difference happening by chance is very small, it would be unlikely to happen if the null hypothesis were true. The conclusion would then be that there is a significant difference between the two painkillers.

The appropriate test to perform in order to examine differences in mean values between two groups of normally distributed data is known as **a two-sample t-test.**

The computer output below show the results generated from performing this test:

Two-Sample T-Test and CI: Painkiller A, Painkiller B

```
Two-sample T for Painkiller A vs Painkiller B

               N  Mean  StDev  SE Mean
Painkiller A  15  64.4   16.5      4.3
Painkiller B  12  80.3   18.1      5.2


Difference = μ (Painkiller A) - μ (Painkiller B)
Estimate for difference:  -15.85
95% CI for difference:  (-29.85, -1.85)
T-Test of difference = 0 (vs ≠): T-Value = -2.35  P-Value = 0.028  DF = 22
```

The p-value is the first entry to consider. The p-value is the probability of getting data as extreme as those actually observed in the experiment if the null hypothesis were true.

The lower the p-value, the more evidence there is against the null hypothesis.

The conventional cut-off for significance is 5%. Therfore, if the p-value is 5% or less, there would be evidence to suggest that the null hypothesis is false and that the study hypothesis of interest (ie the alternative hypothesis) is true.

From the two sample t-test output above, it can be seen that the p-value is $0 \cdot 028$.

This means that the probability of getting a difference of $15 \cdot 85 \text{mg/kg}$ in mean blood glucose levels between the two groups by chance is $0 \cdot 028$ or $2 \cdot 8\%$.

Since this probability is less than 5% it is unlikely that it would happen by chance. Since it would be unlikely to happen by chance, the difference must therefore be due to the different effects of the painkillers.

We can conclude that the mean blood glucose levels for painkiller A is significantly lower than it is for painkiller B at the 5% level.

Note that it is possible to make this claim only if there are no other differences between the groups. It is therefore important that at the start of the study the subjects are allocated randomly into two groups. If they are not then the observed differences between the two groups may be due to another factor. For example, if the average age in one group is higher than the other, age may be the factor that influences the blood glucose levels rather than the painkiller.

Approaches to assessment and gathering evidence

The purpose of this section is to give advice and guidance on approaches to integrating assessment within this Unit.

The Statistics Unit can be assessed in a variety of ways and could include, for example:

- elements of practical assignments such as a project or investigation
- specific assessment tasks or activities
- discrete tests or question papers

The following table gives some examples of how these approaches could be used within the Unit to provide a varied and integrated assessment experience. This approach aims to make the assessment more coherent and meaningful for learners. Please note that these approaches are not exhaustive and other possibilities also exist.

Approach to assessment	Outcome	Examples of approaches to assessment
Assessment tasks/activities	Outcome 1	For Outcome 1, learners could be asked to analyse a set of data using a variety of techniques, to answer questions from non-statistical users, and to present the analysis and conclusions in verbal or written form. Learners will be expected to use a variety of graphical displays, such as, boxplots, scattergraphs, stem and leaf diagrams and histograms. Learners should be able to interpret and comment upon the wide variety of ways of displaying statistical data.
	Outcome 2	For Outcome 2, learners will undertake correlation and regression analysis, and a data analysis. Learners will be expected to upload a data set onto software for analysis and comment.

It would normally be expected that considerable learning and teaching would have taken place prior to the collection of evidence for assessment purposes.

Equality and inclusion

It is important that where possible, inclusive approaches to learning and assessment encourage personalisation and choice for learners. The additional support needs of learners should also be taken into account when planning learning experiences and when considering any reasonable adjustments that may be required.

Any additional support provided to learners to help them access assessment tasks should maintain the integrity of the Outcomes and Assessment Standards.

Examples of support which may be appropriate for this Unit are as follows:

- practical helpers under direct learner instruction could assist with practical activities
- adapted equipment
- the use of a calculator or similar aid
- ICT and other assistive technologies

Other types of support are also possible and would be determined by the teacher/lecturer in response to the specific needs of the learner.

It is recognised that centres have their own duties under equality and other legislation and policy initiatives. The guidance given in these *Unit Support Notes* is designed to sit alongside these duties but is specific to the delivery and assessment of the Unit.

Alternative approaches to Unit assessment to take account of the specific needs of learners can be used. However, the centre must be satisfied that the integrity of the assessment is maintained and that the alternative approach to assessment will, in fact, generate the necessary evidence of achievement.

Appendix 1: Reference documents

The following reference documents will provide useful information and background.

- Assessment Arrangements (for disabled learners and/or those with additional support needs) — various publications on SQA's website: http://www.sqa.org.uk/sqa/14976.html
- *Building the Curriculum 4: Skills for learning, skills for life and skills for work*
- *Building the Curriculum 5: A framework for assessment*
- Design Principles for National Courses
- *Guide to Assessment* (June 2008)
- *Principles and practice papers for curriculum areas*
- *Research Report 4 — Less is More: Good Practice in Reducing Assessment Time*
- *Coursework Authenticity — a Guide for Teachers and Lecturers*
- *SCQF Handbook: User Guide (published 2009)* and SCQF level descriptors www.sqa.org.uk/sqa/4595.html
- *SQA Skills Framework: Skills for Learning, Skills for Life and Skills for Work*
- SQA Guidelines on e-assessment for schools
- SQA Guidelines on online assessment for further education

Appendix 2: Extract from Health Club dataset

	Questions							
Membership No.	1	2	3	4	5	10	11	12
1	1	21	2	1	1	33	20	5
2	2	44	1	3	1	10	18	10
3	2	19	3	1	2	27	22	12
4	2	27	3	2	1	30	17	3
5	1	57	2	1	3	22	0	15
6	2	27	3	1	1	34	17	0
7	1	39	5	2	1	17	48	10
8	2	36	3	1	2	25	18	7
9	1	37	2	1	1	34	15	0
10	2	51	2	2	2	16	12	11
11	1	24	5	2	1	0	42	16
12	2	29	2	1	2	34	22	12
13	1	20	5	1	1	22	31	7
14	2	22	2	1	3	37	25	12
15	2	46	3	1	1	26	9	4
16	2	41	3	1	2	22	7	10
17	1	25	5	1	1	21	29	4
18	2	46	3	1	2	18	8	11
19	1	30	3	1	1	23	18	6
20	1	25	5	2	1	23	19	0
21	2	24	2	1	1	20	7	6
22	2	39	1	2	3	17	0	9
23	1	44	3	1	1	22	8	5
24	1	49	4	2	2	15	10	4
25	2	18	3	1	2	18	19	10
26	1	41	3	1	1	34	10	4
27	2	38	2	1	2	24	14	10
28	1	25	2	1	1	48	22	7
29	1	41	5	2	1	17	27	0
30	2	30	3	1	1	32	13	10
31	2	29	3	1	3	31	0	7
32	2	42	1	2	2	17	14	6
33	1	31	2	1	1	49	21	2
34	2	25	3	1	1	30	17	15
35	1	46	3	1	1	32	8	5

36	1	24	5	2	1	0	36	11
37	2	34	3	1	1	27	14	12
38	2	50	2	1	2	28	8	6
39	1	28	5	1	1	26	22	8
40	2	30	3	1	1	21	9	12
41	1	27	2	1	1	64	15	8
42	2	27	2	4	2	22	10	7
43	1	36	5	1	1	21	24	0
44	2	43	3	1	1	25	13	8
45	1	34	2	1	1	45	15	6
46	2	27	3	1	1	33	10	9
47	2	38	2	1	3	23	0	16
48	1	28	2	1	1	38	13	5
49	1	44	5	4	1	27	19	7
50	2	31	3	1	2	32	11	5

Q1 Are you male or female? 1 – male, 2 – female.

Q2 How old are you?

Q3 Which of the following best describes your main reason for going to the Health Club?

1. Relaxation
2. Improve Fitness
3. Lose weight
4. Meet other people
5. Build strength

Q 4 When you attend the Club, how often do you use cardiovascular equipment?

1. Always
2. Usually
3. Rarely
4. Never

Q 5 When you attend the Club, how often do you use the weights machines?

1. Always
2. Usually
3. Rarely
4. Never

Q10 During your last visit to the Club, how many minutes did you spend on the cardiovascular equipment?

Q11 During your last visit to the Club, how many minutes did you spend on the weights machines?

Appendix 3: Using ‘R’ and Excel software

‘R’ software

There are now many good introductions to R, the most useful of which are on free websites, such as http://www.openintro.org/.

The website of the R project has an introductory manual available at http://cran.r-project.org/doc/manuals/R-intro.html.

An example of helpful web support is at:
http://ecoviz.csumb.edu/wiki/index.php/R_Cheat_Sheet

A very good (and not expensive) introduction is Stowell, S. (2014). Using R for Statistics. New York: Apress. This is an extended and updated version of the same author’s Instant R (2012).

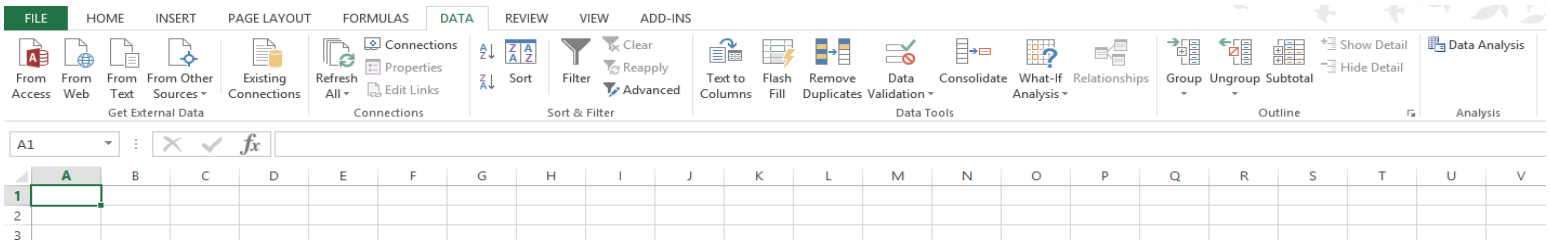
You also get a comprehensive reference manual when you download R to your computer. The manual is placed in the folder where you choose to install R. By default on a Windows computer the folder is called:Program Files>R>R-3.1.0>doc>manual.

Excel software

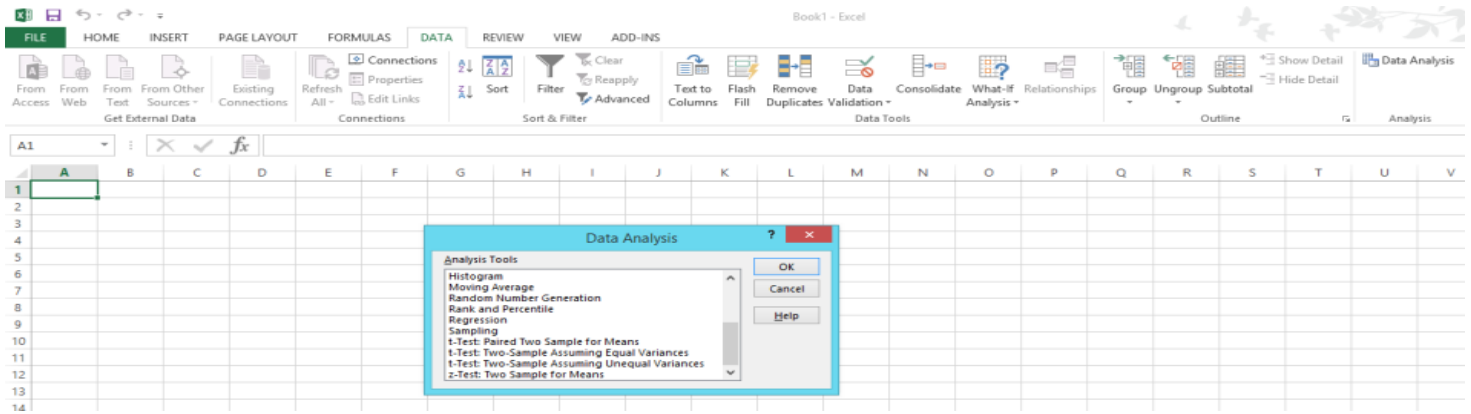
To use Microsoft Excel software for statistical analysis, ensure that the Data Analysis tool is installed as described below:

Under the ‘File’ menu access ‘Options’ then ‘Add-Ins’ and then select ‘Data Analysis’ tool.

Now select the ‘Data’ tab and the screen below should be visible.



Click on the ‘Data Analysis’ tab and the menu box shown below should appear enabling access to statistical analysis tools:



Administrative information

Published: June 2015 (version 1.0)

History of changes to Unit Support Notes

Version	Description of change	Authorised by	Date