Statistics for School Biology Experiments and Advanced Higher Projects

Graeme D Ruxton
Jim Stafford

Two bottom cover images by:
Left: Renjith Krishnan/FreeDigitalPhotos.net
Right: Jannoon028/FreeDigitalPhotos.net

Acknowledgements

This guide is published by SSERC: a registered educational charity and a company limited by guarantee, the corporate members of which are the thirty two Scottish Education Authorities.

SSERC's science, technology and safety services are provided for the elected members, officers, teachers and technicians of those Local Authorities and its other school and college members.

Graeme Ruxton is a Professor of Biology at the University of St Andrews. His research interests are mostly about how plants and animals evolve to avoid being eaten by other animals. Graeme has a particular interest in making the principles of experimental design and the use of statistics in analysing experimental results accessible to the widest range of learners including school students. His book *Experimental Design for the Life Sciences* influenced the development of the Investigative Biology Unit of Advanced Higher Biology.

Jim Stafford is a Senior Associate with SSERC. Previously he has been a Principal Teacher of Biology, a Local Authority Science Adviser and Quality Improvement Officer. During his career Jim has worked with various partner organisations on a number of areas of science education, including leadership training, health and safety guidance, and the development of national qualifications in biology.

Contents

Page

Introduction

The introduction deals with different types of experimental designs and the data they produce - a vital consideration when deciding on which statistical test to use for your data.

We will help you to:

- *distinguish between quantitative and qualitative variables (Measuring variables section)*
- *understand the effect of natural random variation on collected data and how to deal with it (Variation and true values section)*
- *think about different types of experimental design (Types of experiment section)*
- *choose the correct statistical test for the data collected for your particular experimental design (Choosing your statistical test section)*
- *understand how a statistical test helps to deal with chance and uncertainty in your data by using a null hypothesis to determine statistical significance (Dealing with chance and uncertainty section)*
- *set up the statistical package **R** (How to actually carry out statistical tests section).*

We have tried to write this booklet in a way that is accessible to both teachers and senior school students, particularly those students who will undertake a Project investigation in Advanced Higher Biology. If you think incorporating statistics into the report of a biology investigation is difficult, we want to change your mind. Coming up with a project that is both interesting and achievable can require a lot of thinking and ingenuity; collecting the data can require a lot of patience and really test your practical skills. By comparison, simple but effective statistical treatment of the data should be much less challenging. Also, if you skimp on statistics, you are missing a trick in really getting maximum information out of your hard-won data.

Many classroom experiments in biology also lend themselves to a statistical treatment of their results, improving the conclusions we can draw from our data. Such analyses, although perhaps not familiar, are no more demanding than the well established processes of tabulating and graphing results that we carry out at present. Students in the Senior Phase will be familiar with the concepts of variation and chance in biological systems as they apply to the scientific ideas of selection, adaptation and evolution. Statistics applies the concepts of variation and chance to the analysis of experimental results, developing understanding of the important scientific ideas of a null hypothesis and statistical significance.

We don't deny that the mathematics of statistics is complex, and carrying out the calculations can be time consuming and tedious. However, computer software is now available that removes these barriers by allowing collected raw data to be entered directly into the software programme which will then generate the results of a statistical test. Thus the scientist only needs to know the type of statistical treatment that his or her data requires and to be able to understand the computer's output when that treatment is applied. In this booklet we will help you do that for a range of very-commonly encountered types of data.

Collected raw data needs to be processed in some way so that we can make sense of our results. Results may be averaged, counted in different categories, or percentages and ratios calculated. We can also visualise our results by displaying them in tables, graphs, charts or diagrams. All of these help us to make sense of our results and to see patterns and trends from which we can draw conclusions. If you are doing these things already, then you are carrying out what people would describe as descriptive statistics. Statistical tests, what some people would call inferential statistics, provide us with more information that these other methods of analysing results cannot provide. These two types of statistics are often complementary and together can offer a compact but effective description of your data. In order to select the best treatment for your data, you need to be able to categorise the nature of your data, and we tackle that in the next section.

MEASURING VARIABLES

Variables are the traits that we measure in an experiment or field study. Variables may be *quantitative* or *qualitative*.

Quantitative variables are measured on a linear numerical scale. The scale may either be *continuous* (where the variable may take any value on the scale: e.g. mass, time, temperature) or *discrete* (where the variable is measured in whole numbers e.g. number of heart-beats per minute, number of eggs laid). Although discrete variables are measured in whole numbers it is acceptable to express derived data, such as the mean, as a fraction (for example 2.4

children per family). For derived discrete data the mode (most frequently occurring value) might be a more useful statistic than the mean. For example town planners might find the mean family size of 2.4 children a useful statistic when planning the need for primary schools in an area, but the mode of 2 children per family might be more useful when planning the number of rooms in houses to be built. The statistical treatment of quantitative variables is dealt with in Chapters 1, 2 and 3.

Qualitative variables are measured as counts in separate categories (e.g. the numbers of each fish species in a pond, the numbers of males and females in a group, the numbers of pink and white flowers in a sample of plants). Care should be taken to ensure that categories are mutually exclusive so that any one individual can only be allocated to one category. Such categorical data can be processed to produce frequencies, ratios or percentages to compare the counts in each category. Categorical data can also include measurements on an *ordinal scale* where the categories are ranked in order of magnitude, for example a five point scale to describe the abundance of an organism (1 = rare, 2 = occasional, 3 = frequent, 4 = common, 5 = abundant). These are sometimes referred to as ranked variables. The points on an ordinal scale are not at even intervals, therefore it is not a numerical scale (although at first glance it may look like one) and so the data has to be treated as categorical data. In our abundance example, a species scored as 4 is not necessarily twice as common as another species that is scored as 2. To make this clear, the values "1" to "5" in the scale above are arbitrary; we could have coded them as "0" to "4" or even "A" to "E". It is for this reason that such variables are different from a variable like for example "number of eggs in a bird's nest". The statistical treatment of categorical data is dealt with in Chapter 4.

VARIATION AND TRUE VALUES

Random variation is everywhere in biology. It is unlikely that any two individuals will have the same measured value for any variable. So how do we determine the representative value for a large collection of individuals? For example, how do we know what the typical height of 11-year old girls in Scotland is, or the typical size of a Scottish agricultural field? We need to measure a *representative sample* of individuals to determine the average measurement and the extent of the variation around that average. A representative sample is a sample that would be expected to show very similar average measurement and variation around the average as the whole population. Describing a sample of measurements in this way is the substance of Chapter 1. So to understand the heights of Scottish 11-year-old girls we don't have to measure all 45 000 of them; rather we just need to measure a representative sample of them. One of the key skills of experimental design is being sure that your sample is truly representative.

When dealing with variation in measurements we need to make an important distinction between the variation between *repeated measurements* on one individual and measurements on a sample of different individuals (*replication*). Repeated measurements on one individual will show any variation in our measurement method (sometimes called measurement or systematic error). Measurements on a sample of different individuals (replication) will show the variation between individuals (sometimes called natural variation or random error). Be wary of confusing repeated measurements on one individual and replication where you measure several different individuals; failure to distinguish between these is a common error when analysing results. For example, if we wanted to compare the energy intake of boys and girls through use of a food diary we might expect that a given individual will vary quite a lot from day to day so we might ask each individual to record their food intake every third day until we have ten repeated measurements for that individual. Our data point for that individual will be their average daily intake. This gives us a better description of a given individual's characteristic intake; in effect the repeated measurements allow us to improve our measurement of that individual. However, to make a valid comparison between boys and girls we need to replicate this procedure across several boys and several girls. Our number of independent data points will be equal to the number of children in our study, irrespective of how many repeated measurements we have on each.

If a set of repeated measures on one individual shows a small degree of variation then a small number of repeated measurements should lead us to a good estimate for that individual; if the degree of variation between the measurements is large then a greater number of repeated measures will be required. A set of close results that show little variation is referred to as *precise*. However such a set of results may not be *accurate*. That is they may not cluster around the true value, perhaps due to a calibration error in the measuring equipment so that all the results show a similar degree of error. This type of inaccuracy is referred to as *bias* or systematic error. Although such

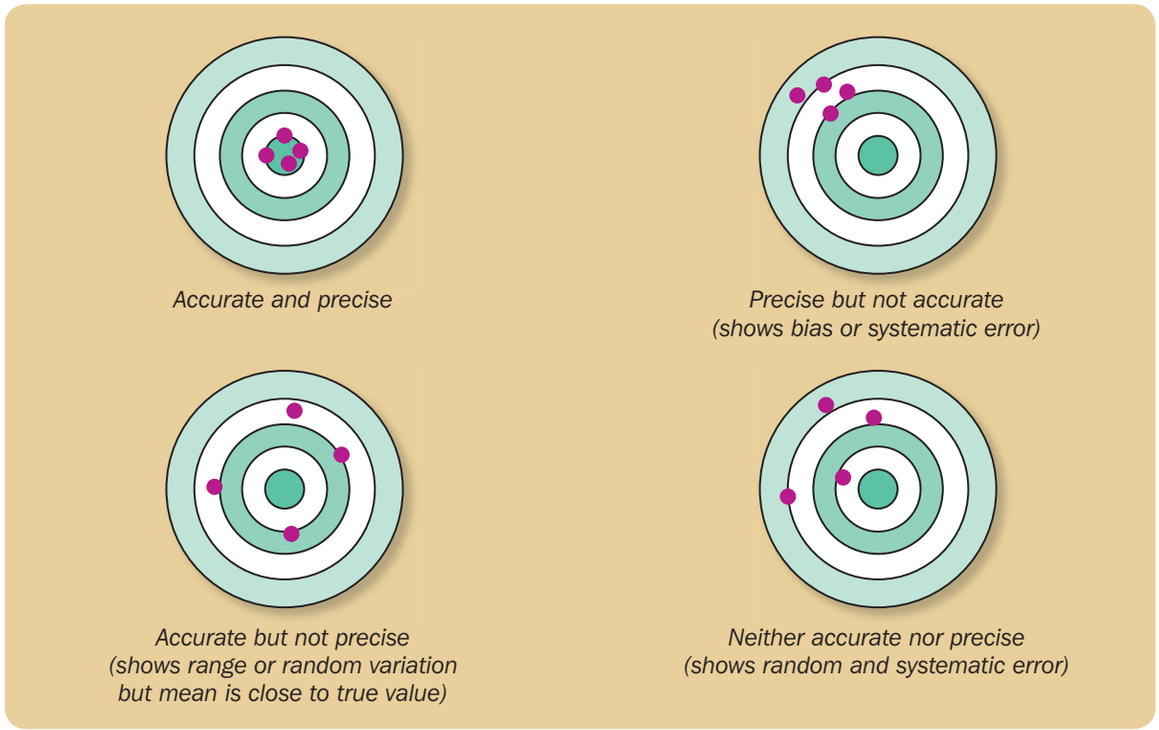


Figure 1 *- Target diagrams illustrating accuracy and precision.*

results will not lead to the true value, any comparison between samples will still show the trends in the data as the measurement error or bias is the same for each sample. The diagram (Figure 1) shows the relationship between accuracy and precision.

We would expect repeated measurements of the height of a single child to show greater precision than their daily energy intake. Any difference in measurement should be very small if we measure them carefully, thus in practice we would only measure a child's height once in most studies. In our energy intake study if we mistakenly used a table that converted different amounts of food into kilocalories rather than joules, but recorded our final results as if they were in joules, we would introduce bias into the results. Although our results would have a systematic error (they would be precise but not accurate) we could still potentially detect any difference between boys and girls although our values for energy intake would be incorrect. In a similar way if a set of replicates shows a large degree of variation in their measured values then we will need a relatively large sample to get close to the true value. If the variation between replicates is small then a low number of replicates should lead to the true value. We can determine the approximate number of replicates required to get close to the true value by starting with a small number of replicates, calculating the mean and then adding further replicates and recalculating the mean (a *cumulative mean*). Once the cumulative mean does not alter then we probably have sufficient replicates to give a true value. The same holds true for ecological sampling - once the cumulative mean of your samples no longer changes then your number of samples is likely to be representative of the whole population.

TYPES OF EXPERIMENT

Because of variation we rarely measure one individual in an experiment or field study; rather we measure a group of individuals. We refer to the measurements we collect from such a group of individuals as our sample. This sample may be the measurements on a group of individuals collected in a field study or the replicates of an experimental treatment. Too often in biology we simply average the measurements in a sample without exploring the nature of the data more fully. This analysis of our sample can be considered as an 'experiment' in its own right and often comprises an important first part of any biological study. Dealing with such data is covered in Chapter 1.

Often in biology we want to make comparisons between two or more different samples. These might be samples collected from different areas in a field study, for example leaf litter from deciduous and coniferous woodland; or they may be different treatments in a laboratory experiment, for example algae grown in the presence of different pollutants. The statistical analysis of the data from these types of experiments is covered in Chapter 2.

Another common type of biology experiment is where we measure two variables for each individual in a sample to see if there is a relationship between one variable and the other. For example we could measure the number of 'eye spots' on a male peacock's tail feathers and the number of females that they mate with. Such a study is a *correlation* or *association study*; that is we are not trying to determine cause and effect, we are simply interested to see if there is a relationship between the two variables. In a correlation study we make no attempt to control any other potentially confounding variables that might affect (or be the cause of) the relationship between the two variables being measured. Such studies are still valuable, as the variables are measured in a 'real life' (in vivo) situation rather than in a manipulated laboratory (in vitro) setting where the experimenter may create unintended effects. To determine cause and effect we use a *controlled* (*manipulative*) study. Here we attempt to control, or failing that measure, any potentially confounding variables so that we can eliminate or take into account their effect. We also set (manipulate) one of the variables (the independent variable) and measure the other variable (the dependent variable) so that we can determine the effect the set variable has on the measured variable. Chapter 3 deals with the statistical analysis of correlation and cause and effect biology experiments. However if the set variable in a cause and effect experiment has a small number of values (say 4 or less) it is better considered as a categorical variable and we should use the methods in Chapter 2.

All of the above types of experiment involve quantitative variables measured on a linear scale. Where we are dealing with qualitative variables, our collected data will be counts in each category. For example, we might want to count the number of birds of different species that visit feeding stations with different seed types to determine their food preferences. The statistical analysis of this type of data is dealt with in Chapter 4.


Image: Goldy/FreeDigitalPhotos.net

CHOOSING YOUR STATISTICAL TEST

Our previous sections should help you to choose a suitable statistical test for your experiment. First, decide whether you are dealing with quantitative or qualitative (categorical) data. Then decide on what type of experiment you are doing. Finally take into account the number of values you will have for your variables. At this point you might want to reconsider the design of your experiment so that it is suitable for statistical treatment. The time to decide upon a suitable statistical test is at the design stage of your experiment, not when the experimental work is complete; by then it may be too late to identify a suitable statistical treatment for your data.

To help you decide on which statistical test to use we have summarised the nature of the variables dealt with in each of Chapters 1 to 4 at the beginning of each Chapter and described what each Chapter does. We have also summarised the information here and the contents of Chapters 1 to 4 in the *Statistical Test Finder* flow chart in Appendix 1.

DEALING WITH CHANCE AND UNCERTAINTY

Statistics help us deal with uncertainty. Imagine we tossed a coin 100 times, we would expect roughly 50 heads and 50 tails, but we know that random chance means that even if the coin was unbiased we would not necessarily expect exactly 50 heads. If we observed 51 heads and 49 tails then we would probably just put the difference down to chance; we wouldn't worry that the coin was biased. Conversely, if it produced 90 heads and 10 tails, we probably would be pretty convinced that it is biased towards heads. However, if it is 60 heads and 40 tails, is that sufficient evidence that the coin is biased? Here two people might disagree based on their gut feelings.

Statistics offers us something more objective than our instincts to help us decide; in the case of this coin-tossing experiment the chi-squared test that we talk about in Chapter 4 is just the test we need.

It is important to realise what statistical tests can and cannot do. In the case above, a chi-squared test will not tell you for certain that the coin is biased or unbiased. The statistical test gives you a *p-value* and it is important to understand how to interpret this. A statistical test looks at how surprising your data is if the *null hypothesis* is true. The null hypothesis is the assumption that nothing interesting is truly happening, so for the coin-tossing experiment the null hypothesis is that the coin is unbiased and equally as likely to produce a head as a tail (see Box 1 for more examples of null hypotheses). The p-value is the likelihood of getting data like yours or even more extreme than yours if the null hypothesis is actually true. For our coin tossing experiment where we observed 60 heads this is the probability of getting sixty or more heads or sixty or more tails from tossing the coin 100 times, if the coin is truly unbiased. If we actually do the chi-squared test for this experiment we get a p-value of 0.0455. So the probability of seeing a result like you have obtained (or an even more extreme result) if the coin is actually unbiased is only 4.55%. This is less than one chance in twenty so feels pretty unlikely, thus we would be justified in thinking this coin was quite probably biased based on our observations and this statistical test.

The question you are now wondering is how small a p-value should be before we get excited. There is a convention in science that p-values less than 0.05 give grounds for rejecting the null hypothesis but values bigger than 0.05 do not. So in the case above we would say that there is statistical evidence to justify rejecting the null hypothesis and suspecting the coin is biased. Whereas if we had obtained 59 heads, then the associated p-value would be 0.072 and we would conclude that our experiment did not provide strong evidence that the coin was likely to be biased.

BOX 1

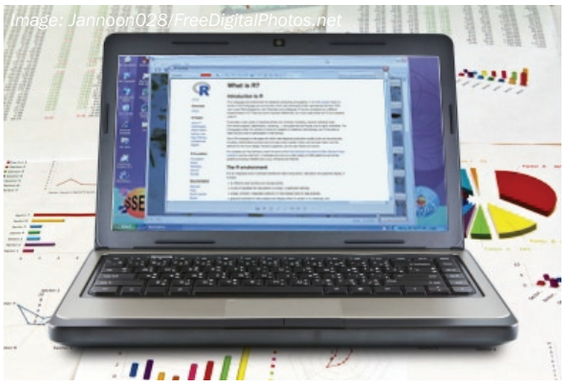
More examples of a null hypothesis

- If we had an experiment comparing the heights of a sample of girls and a sample of boys, then the null hypothesis would be that there is no difference in average height between girls and boys.
- If we had an experiment where we compared plant growth under three different fertiliser treatments, then the null hypothesis would be that plant growth was the same under all three treatments.
- If we had an experiment where we measured the height and running speed in a sample of children then the null hypothesis is that there is no relationship between these two variables, and height is in no way a predictor of running speed.
- If we trapped field mice and scored each as either female or male and as either carrying parasites or not, then we could investigate whether sex influences the likelihood of carrying parasites; the null hypothesis in this case is that sex does not influence likelihood of carrying parasites.

HOW TO ACTUALLY CARRY OUT STATISTICAL TESTS

Statistical tests can often require relatively complex calculations. We would rather that repetitive calculations were left to a computer, freeing up your time to think about biology. Hence we will suggest a number of different methods for getting a computer to carry out statistical tests for you.

We think by far the best way for you to carry out statistical tests is via a fabulous statistical package called ***R***. This is entirely free to download and use without any restriction. You can download it onto any computer and as many computers as you want, you never need to enter any credit card details, you will never be charged, and it will never stop working. There is no catch. It is also a very powerful and well designed package that more and more professional scientists use; and not just for statistics, all of the figures for this booklet were created in ***R***. If you go to university, the most commonly used package for statistics is ***R***. Don't be fooled by the fact that it is free: ***R*** is the Rolls-Royce option for statistical analysis. Once you download it and open it, you will see a large window with a "**>**" cursor, just type in commands here (and press return after each) and it will do all your calculations for you. We will describe which specific commands to use for different statistical tests as we introduce each test. To download ***R*** simply follow the instructions in Box 2.



If for some reason (that we find hard to imagine) you don't like ***R*** or can't download it, we won't leave you high and dry. There are often websites that can help you perform statistical tests, and we will point you to those. Further, most people will have Excel on their computer, and we will also point to its statistical capabilities when relevant. But, we promise you nothing is easier to use, more powerful, or more reliable than ***R***. In particular, ***R*** is more reliable than many other statistical packages with relatively small data sets, a situation that often occurs in school biology. It does not have the most attractive of user-interfaces; but typing in its commands to perform simple statistics is not difficult; and we will walk you through how to do this throughout this booklet.

BOX 2

How do download *R*

1) Go to the ***R*** homepage: http://www.r-project.org/

2) On the left hand panel, just below the title "Download", click on the word "CRAN" to see a page of countries.

3) Scroll down to the UK and click on one of the options, e.g. https://www.stats.bris.ac.uk/R/.

4) Click on the version of ***R*** appropriate to your computer's operating system (probably Mac or Windows).

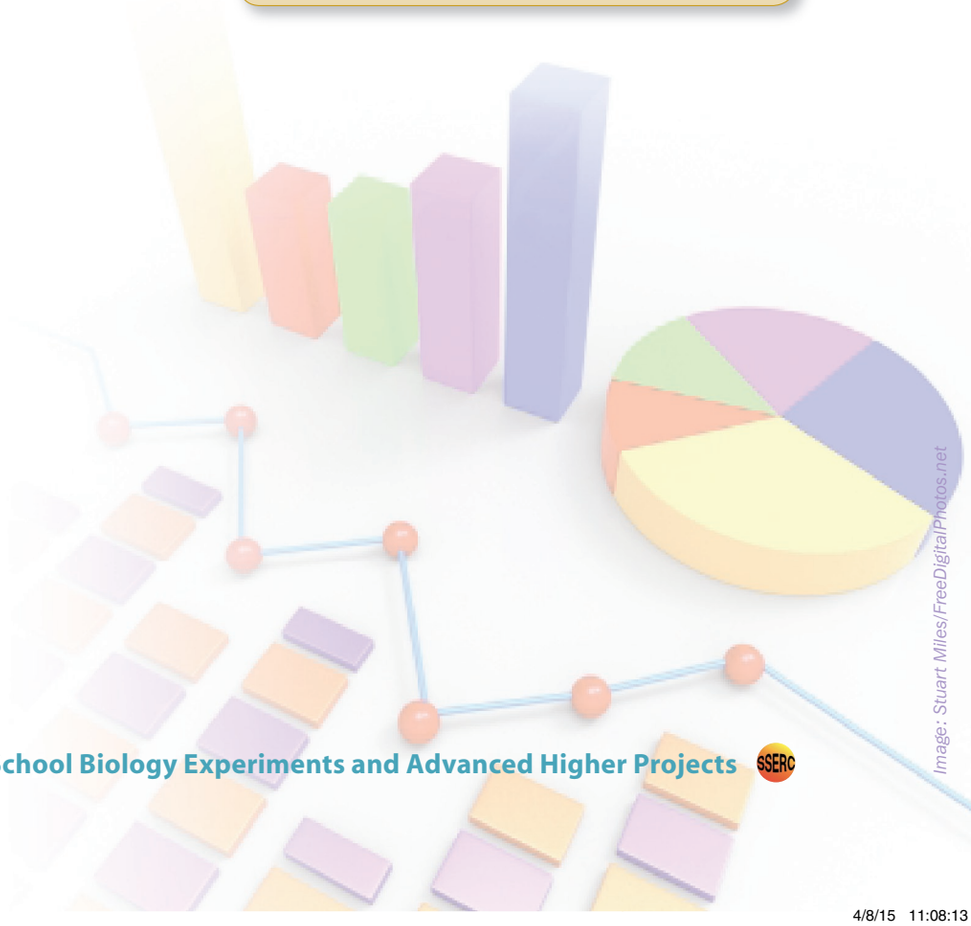
5) Click on the "base" subdirectory.

6) Click on the link to the ***R*** setup program (e.g. "Download R-2.13.2v for windows").

7) When prompted save the programme to your computer's hard drive.

8) Open the folder, click on the "setup" file, agree to everything, select default installation, and say "yes" to a shortcut icon on your desktop.

9) Click on the desktop icon (a big letter "***R***") to start using ***R***.



1 Describing a sample of measurements

This Chapter deals with examining a sample of quantitative data.

In this Chapter we will show you how to:

- *visualise a sample of data in a box plot (section 1.1)*
- *obtain the range, inter-quartile range, median and mean of a data sample using **R** (section 1.2)*
- *draw and interpret a histogram of a sample of data (section 1.3)*
- *obtain a measure of the central tendency (mean or median) of a sample of data using **R** (section 1.4.1)*
- *obtain a measure of the dispersion of a sample of data (the standard deviation) using **R** (section 1.4.2).*

We often measure quantitative variables using a numerical scale (for example length, temperature, time) on one or more samples of individuals. These measurements might be from a sample of individuals of a particular species observed in a field study, or a set of replicate treatments in an experiment. Statistics can help us to describe and make sense of the measurements we make on a sample, and this is what we explore in this Chapter. These statistics will help us to understand the nature of the data we collect, and will give us more information from which we can draw conclusions from the experiment or field study.

1.1 DESCRIBING THE COLLECTED DATA

Let us suppose as part of a study we have measured the heights of a group of 30 11-year-old girls. Our measurements are (in centimetres): 135, 146, 153, 154, 139, 131, 149, 137, 143, 146, 141, 136, 154, 151, 155, 133, 149, 141, 164, 146, 149, 147, 152, 140, 143, 148, 149, 141, 137, 135. How can we make sense of this data? Perhaps the most obvious thing to do is to calculate the *mean* of the data, which will give us the average height of the girls in the sample. The mean is easy to calculate, it is the total of all the values divided by the number of values; in this case the mean height is 144.8 cm. Notice that the convention is to give the mean to one decimal place beyond the actual measurements made. Since our measurements were taken to the nearest centimetre, we quote the mean to the nearest tenth of a centimetre. ***R*** can calculate the mean for us, and we will show you how to do this later, along with some other useful statistics.

The mean still does not give us a great deal of information about our sample; it gives us an idea of the average height of our girls but not how much variation there is around that average. When stating a mean it is also useful to also give the *range* of the data. The range is simply the minimum and maximum measurements; in the case of our data the range is 131 cm to 164 cm.

Now we know a bit more about our data but we still do not have much of an idea about the spread or *dispersion* of the data across the range - are the data points evenly spread or are there more to one end or the other of the range? One way to have a quick look at this is arrange the data in order to find the middle value in the range (the *median*) and compare it to the mean. If we have an even number of values, the median is the mean of the middle two values in the ordered list. The mean or median are sometimes called measures of the *central tendency* of the data; if the mean and the median are similar in value then we are likely to have a relatively symmetrical spread of the data.

The range gives us some idea of the dispersion of the data but we can have a closer look at the dispersion of the data by using the *inter-quartile range*, which is the two values that enclose the "middle 50%" of the sample when we order the values (139-149 cm in our case). In other words, 25%" of the values are less than 139 cm and 25% of the values are greater than 149 cm. The attraction of the inter-quartile range is that it is less sensitive to exceptional values than the range is. The sample can be graphically represented by a box plot (also known as a box and whisker plot) as shown below.

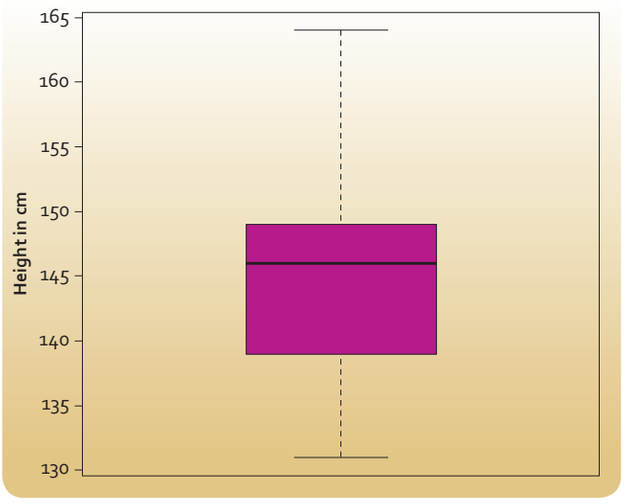


Figure 1.1 *- Box plot of heights of 30 11-year old girls from Rottenrow Primary School.*

1

As with all our figures, we have generated this in ***R***, and the code to do it can be obtained by emailing the authors (email Graeme.Ruxton@st-andrews.ac.uk). The dark bar in the middle of the box tells us the median value of our data, whereas the top and bottom values of the box give us the inter-quartile range. The quartiles are the three values that divide our ordered list of values into quarters. The lower value of the box is the 1st quartile, the median is the 2nd quartile and the upper value of the box is the 3rd quartile. You can see that our distribution isn't symmetric, because the median does not lie in the middle of the box. The exact meaning of the whiskers takes a little bit of thinking about. The whiskers extend to the most extreme data points that lie within 1.5 times the length of the inter-quartile range from the box. Data points that lie beyond the whiskers are known as *outliers*. This definition isn't very easy to use, but the key message is that the box tells us about the spread of values of the middle 50% of the data, and the two whiskers tell us about the spread of values in the top 25% and bottom 25% of ordered values. What we see in our particular case is that there is asymmetry in the values, with the top 25% being much more spread out than the bottom 25% (because the top whisker is longer). We can also conclude in our case that the data are not very spread out because the whiskers extend across the whole range of values from 131 cm to 164 cm. This will not always be the case, imagine that our tallest girl was not 164 cm but 177 cm, then the box plot would look like the case as shown in Figure 1.2.

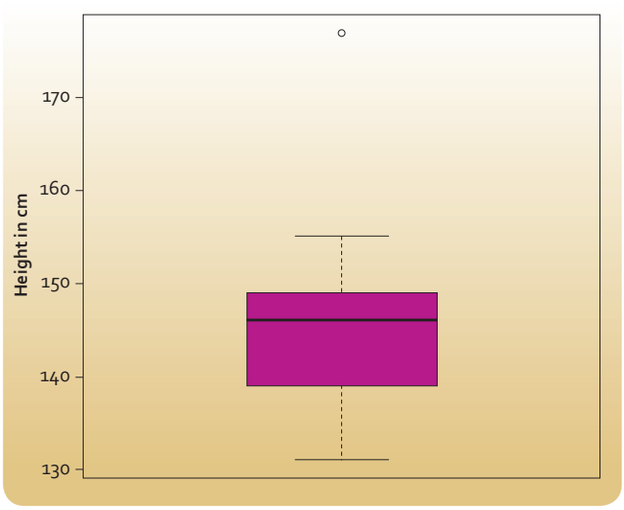


Figure 1.2 *- Box plot of heights of 30 11-year old girls from Rottenrow Primary School, with one very tall girl.*

This shows that box and whisker plots can be used to highlight unusual cases (such as our single very tall girl). There is a single dot on the plot at 177 cm to indicate that there is a single individual that lies outside the span of the whiskers. These unusual cases are *outliers*. You can get more than one such outlier. We will discuss outliers later in this Chapter.

In summary, box and whisker plots can be a useful way of visualising your data in a compact form.

1.2 USING *R* TO PROVIDE SUMMARY STATISTICS ON A SAMPLE OF DATA

We can use ***R*** to provide a summary of the statistics referred to above using the dataset of girls' heights.

We can enter this data into ***R*** using commands like those below:

```
first10 <- c(135,146,153,154,139,131,149,
137,143,146)

second10 <- c(141,136,154,151,155,133,149,
141,164,146)

third10 <- c(149,147,152,140,143,148,149,
141,137,135)

allheights <- c(first10,second10,third10)
```

This is not as mysterious as it looks. We need to tell ***R*** the list of all 30 values in our sample. Here we have divided the data into three sub-lists of ten values each. ***R*** likes you to give it a list as a series of numbers (or existing lists) separated by commas, with the whole list enclosed in round brackets, then the letter "c" in front. You can think of the "c" as telling ***R*** to *connect* all the list of values. You can call a list anything you like by giving the name that you choose (like *first10*) followed by a less than sign immediately followed by a *negative* sign and then the list that you want to assign to that name. You can think of these two symbols (<-) as making an arrow that points to the name that you want to call that list. We do this for all three sublists of ten, then join these three lists together so that *allheights* contains all of our data. If we then type **summary(allheights)** then we get a number of useful statistics,

R responds with

Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
131.0	*139.2*	*146.0*	*144.8*	*149.0*	*164.0*

Specifically we get the minimum and maximum values (and hence the range), the positions of the 1st and 3rd quartiles (and hence the inter-quartile range), the mean and the median.

1.3 HOW TO DRAW AND INTERPRET A HISTOGRAM

It's always useful to visualise our data. Box plots are suitable for relatively small samples of data (say sample sizes less than 20) or where we want to visually compare two or more samples. For a single larger sample of data the most straightforward way to visualise the data is with a histogram like Figure 1.3. The data shown in this figure are the heights in cm of our sample of 30 11-year old girls.

When producing a histogram make sure to give an informative title and label the x and y axes. Histograms essentially group the data into regular intervals also known as *bins*. Here each bin is an interval covering a 5 cm range of heights, and we have 7 bins (one of which has no-one in it). Where a value falls on a boundary between two bins the convention in ***R*** is that it is rounded down (other packages may round up); for example a value of 135 cm is counted into the bin 130 - 135. Selecting a good number of bins may need some trial and error; if you have too few bins then you lose valuable detail about the data, but if you have too many then the data becomes too scattered and it is difficult to see trends in the data. For example, in Figure 1.4 we have used 20 bins and we think the trends in the data are more difficult to immediately absorb than in Figure 1.3.

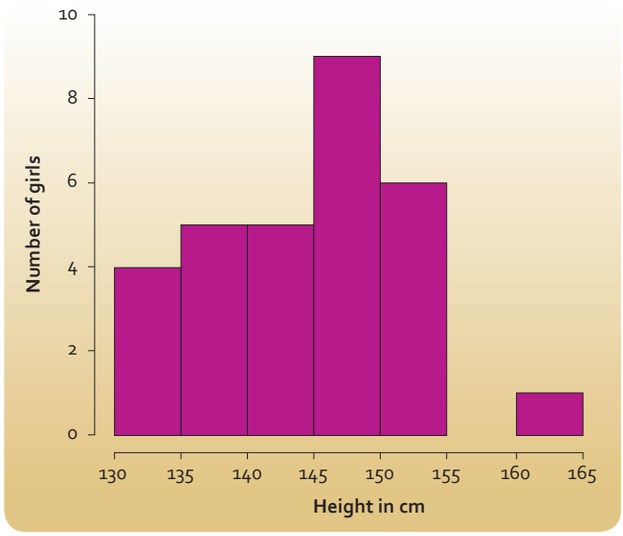


Figure 1.3 *- Heights of 30 11-year old girls from Rottenrow Primary School.*

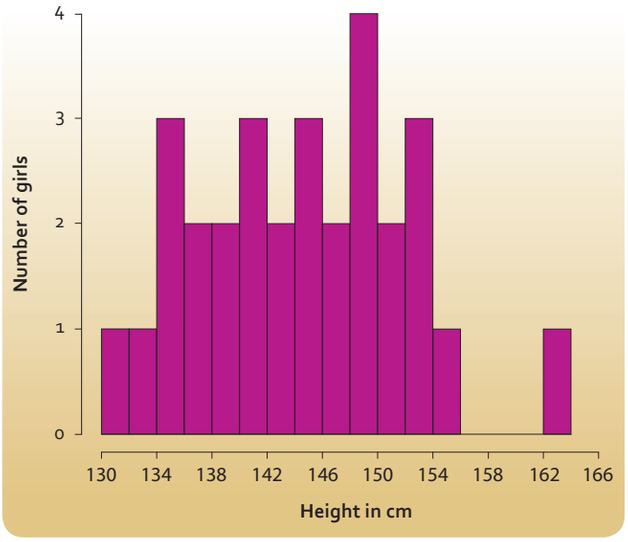


Figure 1.4 *- Heights of 30 11-year old girls from Rottenrow Primary School.*

In general you should probably have at least seven bins in your histogram. A good way to decide on bin size and number of bins is to record your data in a stem and leaf diagram (see Figure 1.5 below). This will give you a rough idea of how a histogram might look and you can then adjust the interval sizes to something that you think will produce a suitable histogram. Another useful rule of thumb to use is the *Rice Rule* where the number of bins should be around two times the cube root of the number of individuals in the sample ($2n^{1/3}$) where n is the number of individuals in the sample (30 in our case). This is more of a guide than a hard and fast rule, and you might want to alter the number of bins so that the boundaries between bins are easy-to-digest round numbers like we have in Figure 1.3.

Stem	Leaves
13	1 3 5 5 6 7 7 9
14	0 1 1 1 3 3 6 6 6 7 8 9 9 9 9
15	12 3 4 4 5
16	4

Figure 1.5 *- Heights of 30 11-year old girls from Rottenrow Primary School as a stem and leaf diagram.*

The stem and leaf diagram is formed by splitting each number into two parts based on the last two significant digits, in this case between the tens and units. The first column is the group of all the numbers beginning with one hundred and thirty, one hundred and forty and so on and the second column records each individual number. So the numbers in the first

1

row of the second column are 131, 133, 135 and so on. We can see at a glance what a 4 column histogram with a bin width of ten would look like; but as we should normally have at least 7 columns, it looks like 5 cm width bins might be suitable for our histogram (just as we used in Figure 1.3).

Let us turn from drawing a histogram to interpreting one. The first useful thing your histogram can do is help you spot errors in your measurements. If our histogram suggested that one of our girls had a height of 13 cm or 650 cm (the height of a giraffe) then this would suggest to us that something strange has occurred, and we would check to see if we have simply mistyped one of the values into the computer (in which case we can correct the mistake), or we have made a mistake in our original measurement or recording of that measurement (in which case we can either go back and re-measure the girl in question or remove the erroneous measurement from our sample). It is perfectly appropriate to remove an individual from your sample if this type of checking indicates that you have clearly made a mistake in measuring them.

However in a case like ours all the data seem plausible and we can use the histogram to get a feel for the data. Aspects of the data we might be interested in are explored below.

The typical or characteristic value
In terms of describing the typical or characteristic height of girls in our sample, we can see that the modal (most common) bin has heights from 145 cm - 150 cm. But we note that there are more girls (14) with heights in lower bins than the modal one than in higher bins. So we would expect that typically girls are around 145 cm or a little less. We would expect the mean or median values of our sample to be around 145 cm or a little less.

The spread of values around that characteristic value
We can say that there is relatively modest spread of heights in the sample. If the median height is expected to be around 145 cm the smallest and largest heights are only about 15 cm (about 10%) different from that median value. So girls in our sample don't differ by very much from each other in height. This is not too surprising, we would be surprised if one 11-year old girl was twice as tall as another; but if we were measuring swimming speed it would not be so surprising if one girl took twice as long as another to swim 25 m.

The degree of symmetry in the distribution
We can see a bit of a suggestion that the distribution of heights is not particularly symmetric. To see how scientists describe such asymmetry we need to introduce you to some jargon. Firstly, we would expect that in any distribution there will be a few values that are either extremely high or extremely low compared to the rest; the groups of these extreme values are called the *tails of the distribution*.

For our height data, we can see that the tail of extremely high values is further away from the bulk of the values than the extremely low tail. In such a situation the bulk of values are bunched on the left side of the histogram. This is a situation called positive skew. However, there are other names for *positive skew*: *right-skewed*, *right-tailed* or *skewed to the right*. Whilst it is reasonable to say that our data are suggestive of a positive skew, we would be wise to be cautious in avoiding putting too much emphasis on this, since the tail of extremely high values is based on only a single individual.

Unsurprisingly, the opposite case where the tail of low values is further from the concentration of most values, and most values are concentrated to the right of the figure is called negative skew (or *left-skewed*, *left-tailed*, or *skewed to the left*). A good example of data that we might expect to show negative skew is the age at which people die in developed countries (see Figure 1.6 below). Currently, most people dying are aged between 70 and 100; there is a short tail of higher values with no-one living beyond 110, but a long tail of lower values with small numbers of people at all younger ages dying.

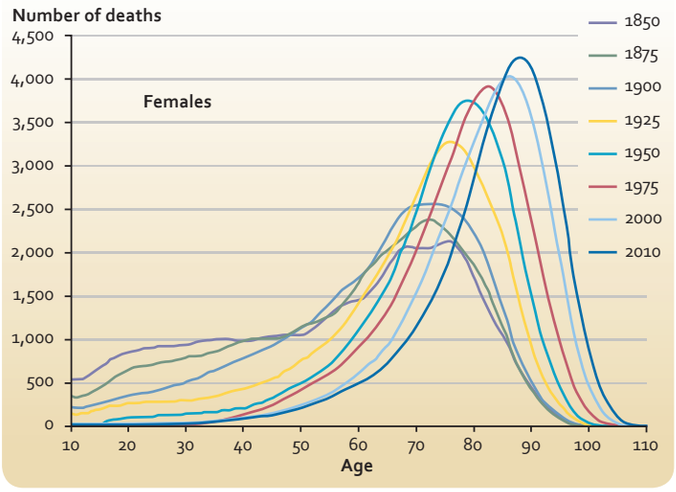


Figure 1.6 *- Distribution of ages at death for 100,000 women in England and Wales in different years.*

Any unusual values (outliers)
There is one girl in our sample who is over 160 cm tall (Figure 1.3), and is at least 5 cm taller than anyone else in the sample. This is noteworthy, but her height isn't sufficiently far from the bulk of the data to consider that our measurement should be in question.

In summary, plotting a histogram of any sample can give us a good feel for the nature of the data in our sample, and can allow us to describe it quite fully at least qualitatively. We can complement this with a number of summary statistics to describe aspects of a sample quantitatively (see below).

1.4 DESCRIBING THE CENTRAL TENDENCY AND STANDARD DEVIATION OF A SAMPLE OF DATA

1.4.1 Describing central tendency (mean or median)

The *summary* function in ***R*** gives us two measures of the typical or "average" value of the data: the mean and median. Which should we use? A good rule of thumb would be to use the mean to describe the typical value unless a histogram of your data shows that the distribution is very skewed, in which case use the median. If the distribution is exactly symmetrical then the mean and median will be exactly the same; if the distribution is not strongly skewed (as in our case), then the two values will be similar. In this case, it makes sense to use the mean because this is the measure that most people will be familiar with. However, if the data are strongly skewed then the median is a better description of the average than the mean. Imagine if we asked the girls in our survey how many pets there were in their household, most would likely answer a number like 0 or 1 or 2, but it's easy to imagine why one girl might respond that there are 500 pets in her household since her father is a tropical fish enthusiast. This one girl would have a substantial effect on the mean, which might be something like 17. But 17 does not feel like a useful description of the typical number of pets per household, whereas the median will likely take the value 1, which does feel like a more useful description of the typical number of pets. This is the attractive characteristic of the median in general; it is much less influenced by extreme values than the mean is.

1.4.2 Standard deviation

The mean or median are measures of the *central tendency* of the data. The spread of the data (sometimes called the dispersion of the data) can be described by the *range* or *inter-quartile range* that we have already discussed in this Chapter. Another commonly-used measure of dispersion is called the *standard deviation*. The standard deviation is best used to describe the dispersion of a data set that is pretty much symmetrical around its mean. As a generality, to interpret this measure, we would expect for most symmetrical data sets that almost all the sample values (in fact, 95% of them) will lie within two standard deviations of the mean. A low standard deviation means the values are close to the mean value and a high standard deviation means the values are spread out over a large range.

You can calculate the standard deviation for our data on girls' heights in ***R*** using the *sd* function. When we type **sd(allheights)** we get the response

```
[1] 7.694154,
```

which means that the standard deviation is approximately 7.7 cm. So for our data we would summarise the mean and standard deviation as 144.8 ±7.7 cm. Notice that we round our values down to avoid giving the impression that we think our measurements are more precise than they really are (see section 5.1 for more on this).

Now in the case of the girls' heights twice the standard deviation is 15.4 cm, and the mean is 144.8 cm; so we would expect most of our data to fall between 129.4 cm and 160.2 cm. This pretty much holds true, with only one (3%) of our 30 measurements being outside this range. Just to introduce you to a final piece of statistical jargon that you might encounter, the standard deviation squared is called the *variance* of the sample.

So which measure of dispersion should you use? As a second rule of thumb we would recommend that if your data shows a symmetric distribution quote a mean and give a standard deviation alongside it, and if you quote a median give the inter-quartile range too.

1

1.5 ADDITIONAL NOTES

If you can't or don't want to use ***R*** then all of the measures described in this section can be obtained if you type your data into Excel. How to do it will vary a little depending on exactly what version of Excel you have, but Excel is so commonly-used that you will be able to find how to do it in a flash on the internet, just type something like "calculating quartiles in Excel 2010" or "calculating standard deviation in Excel 2003" into your favourite search engine and you will find the answer.

One last word of caution, all the examples we have talked about so far have been data that clumps together into a distribution with only one peak. This will be true for most datasets, but just occasionally we will encounter data that has more than one peak. If you imagine being stationed beside an arterial road into a major town and logging the times of all the cars that pass you, you might expect one peak of activity corresponding to the morning rush hour and another peak corresponding to the evening rush hour. For such data (called *multimodal* rather than *unimodal*) the measures described in this section can be quite misleading. If you draw a histogram of your data and it appears to be multimodal (have more than one peak) then do not attempt to describe the data quantitatively, but present the histogram and describe observed trends in that data qualitatively only.

It is also possible to test if the mean or median of a given sample is statistically significantly different from a specific value, and we discuss how to do that in the next Chapter.

1.6 CHAPTER CONCLUSION

Even if your research question involves comparing different samples of data, we think your understanding of your data will be improved if you first explore and describe every sample separately using the methods described in this Chapter. Anyone reading a report on your study will also benefit from these preliminary descriptions of the data before moving on to more complicated comparisons between samples. We will explore such comparisons in the next Chapter.

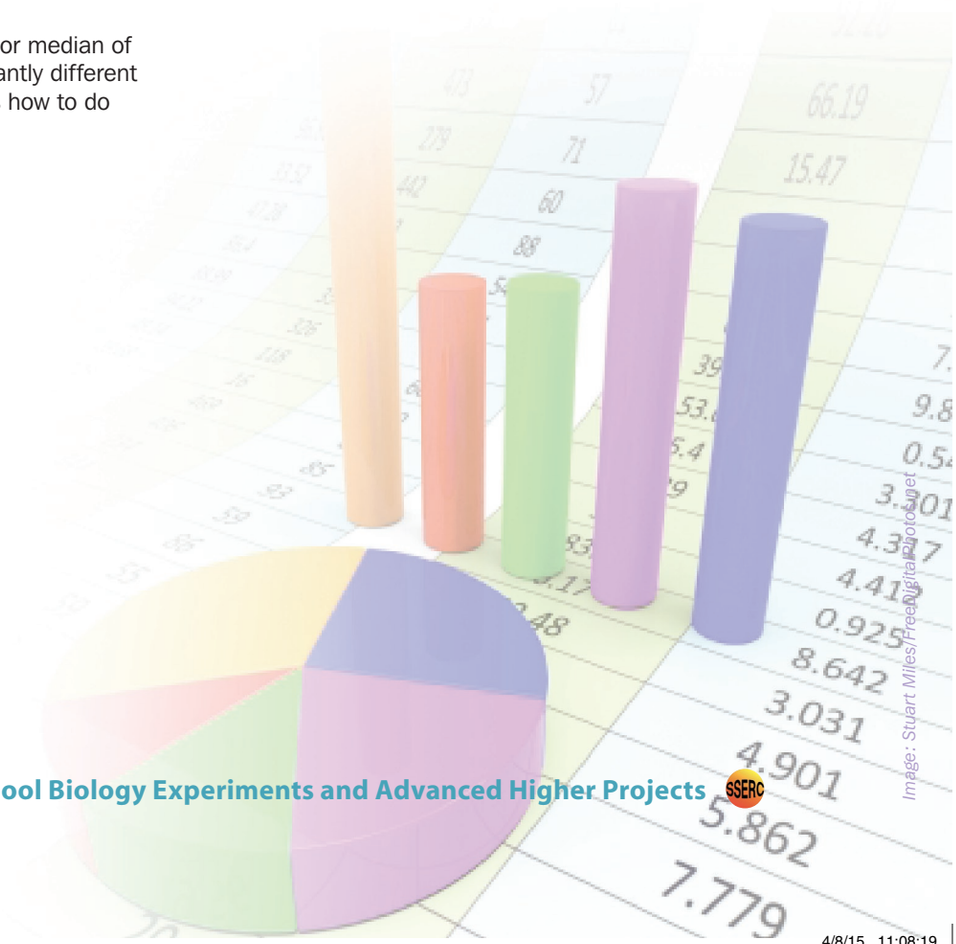


Image: Stuart Miles/FreeDigitalPhotos.net

Comparing two or more samples
(and comparing one sample with a theoretical prediction)

2

This Chapter deals with comparing two or more samples of quantitative data. If you wish to compare two or more samples of qualitative (categorical) data refer to Chapter 4.

In this Chapter we will show you how to:

- *get a preliminary feel for the nature of the data by comparing box plots of the samples (section 2.1)*
- *use the correct statistical test to compare the difference between two samples (section 2.2)*
- *use statistical tests to compare more than two samples (section 2.3)*
- *use a statistical test to compare the mean or median of a sample to a predicted or specified value (section 2.4).*

In the last Chapter we considered how to extract as much information as we can from a single sample. However, many research questions require you to compare between samples. For example, we might sample both girls and boys to ask whether running speeds are different between the two sexes, or we might grow ten seedlings in each of four different types of compost to explore the effect of rooting substrate on growth. This Chapter will look at how we can best compare between samples.

2.1 GETTING A PRELIMINARY FEEL FOR THE SAMPLES

We should begin by taking the approach discussed in the last Chapter to get a feel for our data. Imagine that we select two lines of fruit fly of the genus *Drosophila*, a *resistant* strain that shows resistance to a pesticide and a *control* strain that shows no such resistance. We sample 25 females from each of the two strains and for each measure their fecundity (reproductive rate) as the mean number of eggs laid per day over the first 14 days of life. We could type the data into ***R*** using the code below.

```
resistant <- c(12.8,21.6,14.8,23.1,34.6,19.7,
22.6,29.6,16.4,20.3,29.3,14.9,27.3,22.4,27.5,
20.3,38.7, 26.4,23.7,26.1,29.5,38.6,44.4,23.2,
23.6)

control <- c(35.4,27.4,19.3,41.8,20.3,37.6,36.9,
37.3,28.2,23.4,33.7,29.2,41.7,22.6,40.4,34.4,
30.4,14.9,51.8,33.8,37.9,29.5,42.4,36.6,47.4)

summary(resistant)

summary(control)
```

For the *resistant* strain

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.80	20.30	23.60	25.26	29.30	44.40

And for the *control* strain

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14.90	28.20	34.40	33.37	37.90	51.80

From this we see that in both cases the mean and median are relatively similar, suggesting both samples are at least approximately symmetric. The inter-quartile ranges are broadly similar in length, suggesting that the spreads of values are similar in each case. Since the samples are not strongly asymmetric, we can ask ***R*** to get the standard deviations for the two samples by using the commands **sd(resistant)** and **sd(control)**: yielding SD values of 7.8 and 8.9 respectively. This suggests that the spread of values might be a little higher in the *control* group but not by much. By comparison the mean and median are substantially higher for the *control* group than the *resistant* group. We can see these effects more clearly if we also plot the data using box plots for the two groups alongside each other.

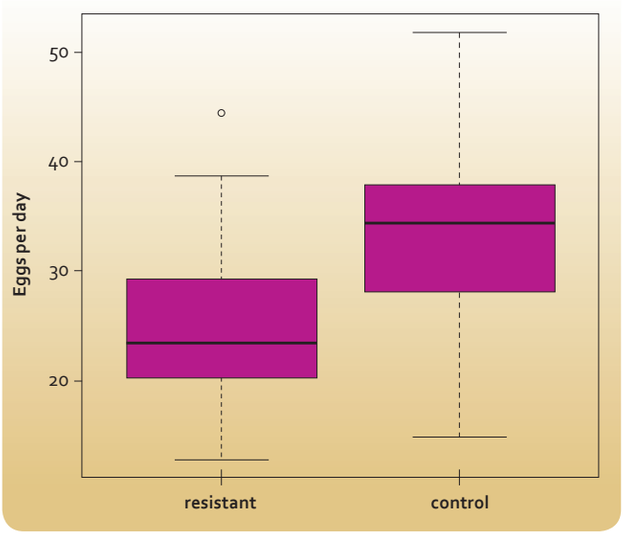


Figure 2.1 *- Daily number of eggs produced for 25 female fruit flies from each of resistant and control strains.*

2

It would seem reasonable on the basis of our preliminary examinations to formally test whether on average resistant females are less fecund than control ones; that is explore if selection for pesticide resistance also selects for reduced fecundity. We tackle such testing in the next section.

2.2 A STATISTICAL TEST FOR A DIFFERENCE BETWEEN TWO SAMPLES

A statistical test that can be used for any two independent samples is the Wilcoxon rank sum test (sometimes called the Mann-Whitney U-test). This is a good general test to use as it ranks all of the data in order and then adds the ranks in each sample and then compares the total sum in each sample. This makes the Wilcoxon rank sum test suitable for small samples where we are not sure what the underlying shape of the data would be if we could get a big sample - a situation that often applies in school biology experiments. In our case this tests the null hypothesis that selection for resistance has no effect on fecundity. To implement this in ***R*** we simply type the additional command

```
Wilcox.test(resistant,control),
```

and this will generate the following output:

```
Wilcoxon rank sum test with continuity correction

data: resistant and control

W = 156.5, p-value = 0.002547

alternative hypothesis: true location shift is not
equal to 0

Warning message:

In wilcox.test.default(resistant, control):

cannot compute exact p-value with ties
```

We can safely ignore the warning message. All that matters to us is that the p value (0.002547) is less than 0.05. Thus we have evidence to reject the null hypothesis. Since on the basis of our graphing and summary statistics the biggest difference seems to be in central tendency rather than dispersions or shapes of the distributions, we can conclude that the median daily fecundity is statistically significantly lower in the resistant strain than in the control strain.

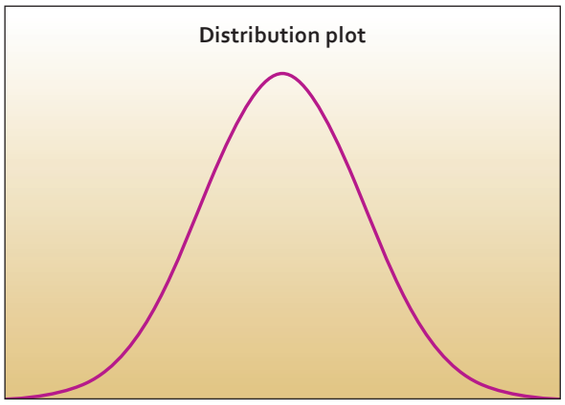


Figure 2.2 *- A normal curve.*

As a result of how the Wilcoxon test actually works, the key thing to remember is that we should discuss any differences that this test shows up in terms of medians and inter-quartile ranges, rather than means and standard deviations.

There is an alternative test (called a t-test) that can be used if the data approximates to a normal distribution. The word *normal* has a special meaning in statistics, and refers to a special type of distribution of data that when plotted as a histogram has the shape shown in Figure 2.2.

The key properties of the normal distribution are that (i) it is symmetric, (ii) it has a single peak, and (iii) values cluster strongly around the peak value so that most of the data are relatively close to the mean value but with smaller numbers of values being significantly lower or higher. These properties are often summarised as suggesting that the histograms of the samples should be shaped like a church bell. The t-test compares the differences between the two means relative to the spread of their values. If the spread of values between the two samples overlap considerably then the difference between the two means is less likely to be significant than if there were only a small degree of overlap between the two sets of values. Let us produce histograms of our two distributions (Figure 2.3).

Surprisingly, there is no hard and fast rule for how similar to a perfect normal distribution histograms should be for a t-test to be appropriate. Perhaps even more surprisingly to you, most practising scientists would probably say that the two histograms in Figure 2.3 were close enough.

If we simply type **t.test(resistant,control)** then ***R*** gives us the following output:

Welch Two Sample t-test

data: resistant and control

t = -3.4251, df = 47.087, p-value = 0.001283

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-12.882696 -3.349304

sample estimates:

mean of x mean of y

25.256 33.372

The t-test compares the means and the p-value relates to the null hypothesis that the means of the two samples are the same. In our case the low p-value gives us grounds to reject this hypothesis and conclude that we have statistically significant evidence that pesticide resistance affects mean daily fecundity. We can see from our earlier investigation that this effect seems to be that pesticide resistance *reduces* fecundity.

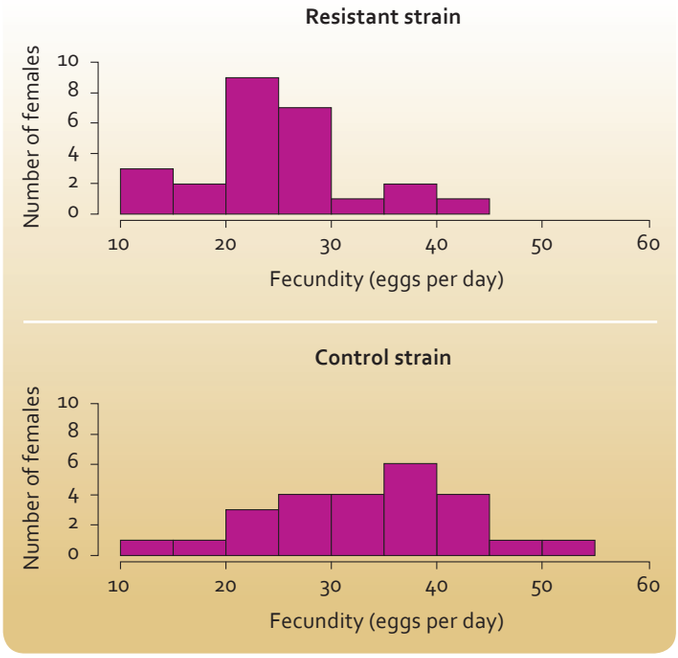


Figure 2.3 *- Histograms of the daily fecundity scores of 25 randomly selected female fruit flies from the resistant and control groups.*

So which of these two tests should you use? The Wilcoxon test is always valid, and so should probably be your default option. The t-test is more powerful (more able to detect differences that exist between the samples) if the samples are reasonably close to a normal distribution, but its p-values are unreliable if even one of the samples is too different from normal. Since there is no simple rule for deciding when a sample is close enough to normal, then you should use the t-test with caution. You should use it with even greater caution if using a package other than *R*: *R* uses a clever version of the t-test which does not mind if the two distributions have quite different dispersions, this is sometimes called the *Welch test* or the *unequal variances t-test*. Most packages do not use this version, instead using a simpler version that assumes the two distributions have similar dispersions. Using this simpler version when one sample is more spread out that the other can lead to unreliable p-values.

The t-test is available in Excel. You can use your search engine to find online tutorials on how to implement it in your particular version. You should select the "two-sample assuming unequal variances" option. For the other test we recommend using an on-line calculator if you don't want to use *R*. Just type "Mann-Whitney U-test calculator" into your favourite search engine to find some. One that we found that is easy to use is http://vassarstats.net/utest.html.

2.3 COMPARING MORE THAN TWO GROUPS

Imagine that we selected a third strain of flies which had an enhanced susceptibility to the pesticide - so that this *susceptible* strain showed the opposite selection of that in the *resistant* strain discussed in the last section. We might reasonably want to test for differences in fecundity across all three groups. Let's start by looking at the data as box plots (Figure 2.4).

It looks as though selection either way (for or against resistance) reduces fecundity compared to the unselected control groups, but is this effect statistically significant, and is the fecundity the same for both the selected groups. We would like a statistical test that explores all these questions. The Kruskal-Wallis test can help us do just that, you can think of it as the generalisation of the Wilcoxon test to more than two groups.

2

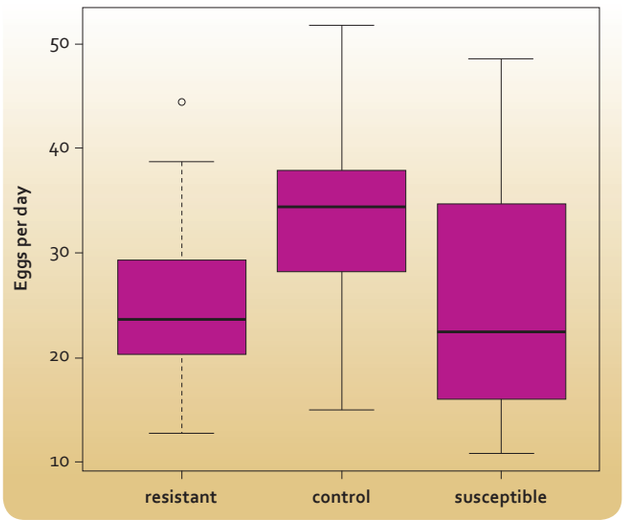


Figure 2.4 *- Daily number of eggs produced for 25 female fruit flies from each of resistant, susceptible and control strains.*

If we type the following into ***R***

```
resistant <- c(12.8,21.6,14.8,23.1,34.6,19.7,
22.6,29.6,16.4,20.3,29.3,14.9,27.3,22.4,27.5,
20.3,38.7, 26.4,23.7,26.1,29.5,38.6,44.4,23.2,
23.6)

control <-c(35.4,27.4,19.3,41.8,20.3,37.6,36.9,
37.3,28.2,23.4,33.7,29.2,41.7,22.6,40.4,34.4,
30.4,14.9,51.8, 33.8,37.9,29.5,42.4,36.6,47.4)

susceptible <- c(38.4,32.9,48.5,20.9,11.6,22.3,
30.2,33.4,26.7,39.0,12.8,14.6,12.2,23.1,29.4,
16.0,20.1,23.3,22.9,22.5,15.1,31.0,16.9,16.1,
10.8)

kruskal.test(list(resistant,control,susceptible))
```

Then we get the response:

```
Kruskal-Wallis rank sum test

data:  list(resistant, control, susceptible)

Kruskal-Wallis chi-squared = 14.0456, df = 2,
p-value = 0.0008913
```

The Kruskal Wallis test examines the null hypothesis that all the medians of the different groups are the same. In this case, the small value for the p-value suggests that the null hypothesis is not true. However, this test does not tell us which groups have different fecundities from which other groups. To find that out we would now have to carry out three Wilcoxon tests to compare each group with each of the other two groups using exactly the methodology described in the last section. It is still worth doing the Kruskal Wallis test first, because if the p-value was greater than 0.05 we could have concluded that all the groups were similar and would not have needed to do the additional Wilcoxon tests.

There is also an extension of the t-test to more than two groups - this test is called *one-way Analysis of Variance or one-way ANOVA*. If we type in the following commands to ***R***

```
resistant <- c(12.8,21.6,14.8,23.1,34.6,19.7,
22.6,29.6,16.4,20.3,29.3,14.9,27.3,22.4,27.5,
20.3,38.7, 26.4,23.7,26.1,29.5,38.6,44.4,23.2,
23.6)

control <- c(35.4,27.4,19.3,41.8,20.3,37.6,
36.9,37.3,28.2,23.4,33.7,29.2,41.7,22.6,40.4,
34.4,30.4,14.9,51.8,33.8,37.9,29.5,42.4,36.6,
47.4)

susceptible <- c(38.4,32.9,48.5,20.9,11.6,
22.3,30.2,33.4,26.7,39.0,12.8,14.6,12.2,23.1,
29.4,16.0,20.1,23.3,22.9,22.5,15.1,31.0,16.9,
16.1,10.8)

allvalues <- c(resistant,control,susceptible)

groups <- c(rep(1,25),rep(2,25),rep(3,25))

oneway.test(allvalues ~ groups)
```

then we can implement this test. Now these commands are the strangest set we ever have to use to do a statistical test in this booklet, but they are not too strange when we look closely. We end up with *allvalues* being a list of all our 75 fecundity measures. We need some way to tell ***R*** which particular group (resistant, control or susceptible) a given measure belongs to; and our list groups does that. The expression we use to define *groups* tells ***R*** to make a list that starts with repeating “1” 25 times, then repeating “2” 25 times, then repeating “3” 25 times. So it is a list that is 75 long. The first 25 items on the list have the same value (“1”) and that relates to the first 25 values on our list *allvalues* belonging to

the same group ("resistant"). In a similar way every value in the list *groups* can tell ***R*** which of our three groups any measurement in *allvalues* belongs to. Our last command implemented the statistical test using all the data in our list *allvalues* but splitting that list into different samples according to the values in groups. When we type in that code, we get the following response back from ***R***.

One-way analysis of means (not assuming equal variances)

data: allvalues and groups

F = 8.2744, num df = 2.000, denom df = 47.564, p-value = 0.0008247

The null-hypothesis under test here is that all the means are the same. The small p-value gives us cause to reject that null-hypothesis, and we would then use three different t-tests to work out which means are different from which other means.

Again, if you don't like ***R***, then ANOVA is available in Excel, it is called "Anova: single factor"; and calculators are available online for the Kruskal Wallis test. For comparing across three groups we like to use http://vassarstats.net/kw3.html.

You can use this approach in theory for comparing any number of groups, however imagine you had five different groups and ANOVA suggested a difference between at least some means. In fact this would require 10 different t-tests for you to compare every group with every other group. You probably should not do all the t-tests. All the t-tests will be tedious for you to perform and for someone reading your report to wade through all your results; but more importantly such multiple testing carries a risk. Statistical tests are good but not perfect, so sometimes they will give you a low p-value and make you think there is a difference between two groups when there really isn't one. Remember that the chance of this happening in any one test is pretty low, so in general we don't worry about it; but when we end up doing ten or more similar tests on the same experiment then we might need to worry about all the small chances adding up. So if you are working with more than three groups, think about not comparing every group. Rather, on the basis of initial graphing of your data and thinking about scientific hypotheses that you feel are most important, select a smaller number of comparisons to make. As a rule of thumb try not to make more comparisons that you have groups.

2.4 COMPARING THE MEAN OR MEDIAN OF A DISTRIBUTION AGAINST A SPECIFIED VALUE

Imagine that you read in a book that, based on our understanding of their physiology, healthy fruit fly females should produce 40 eggs a day on average. We could use our control group to test this prediction. The mean and median values for our sample of 25 control individuals were 33.4 and 34.4 eggs per day respectively. These values seem a little lower than 40, but are they sufficiently lower than expected to claim that we have evidence that our fruit flies are producing fewer eggs than expected for healthy individuals? We can use the Wilcoxon and t-test to explore this. If we type the following into ***R***:

```
control <- c(35.4,27.4,19.3,41.8,20.3,37.6,36.9,
37.3,28.2,23.4,33.7,29.2,41.7,22.6,40.4,34.4,
30.4,14.9,51.8,33.8,37.9,29.5,42.4,36.6,47.4)

wilcox.test(control, mu = 40)

t.test(control, mu = 40)
```

Where we use 'mu' to specify the mean or median we want to test. Then we get the following output:

Wilcoxon signed rank test with continuity correction

data: control

V = 43.5, p-value = 0.00143

alternative hypothesis: true location is not equal to 40

Warning message:

In wilcox.test.default(control, mu = 40):

cannot compute exact p-value with ties

One Sample t-test

data: control

t = -3.7061, df = 24, p-value = 0.001103

alternative hypothesis: true mean is not equal to 40

95 percent confidence interval:

29.68092 37.06308

sample estimates:

mean of x

33.372

The Wilcoxon test (which ***R*** calls by the alternate name 'Wilcoxon signed rank test') produces a p-value related to the null hypothesis that the median value is equal to 40, the small p-value gives us reason to reject this; suggesting that the median is significantly lower than 40. The t-test examines the null hypothesis that the mean value is 40, and again we have reason to reject this based on the p-value. Hence we have reason to believe that our control fruit flies are producing fewer eggs than predicted by previous work. This result demonstrates the importance of having a control in your experiment rather than depending on results obtained in another experiment. It might be the case that in our experiment variation in other factors, such as temperature or the diet of the fruit flies for example, may have had an effect on fecundity.

This approach can be used to test the mean or median of any sample against a specified value. The specified value might be that observed in a previous experiment, or a theoretical prediction. Remember that strictly speaking you should only use the t-test if the distribution of the sample looks reasonably close to a normal distribution. If you don't want to use ***R*** your best bet is to type "one sample t-test calculator" or "one sample Wilcoxon test calculator" into your favourite search engine.

2.5 CHAPTER CONCLUSION

In this Chapter we have described various ways to use null hypothesis statistical testing. An important thing to remember is that the p-value of a statistical test allows us to infer whether a null hypothesis is supported by our data or not. The null hypothesis is often that there is no effect (for example that there is no difference between two groups). If we feel that our data suggests that the null hypothesis can be rejected on the basis of a statistical test, the statistical test does not tell us what the direction and size of the effect is. In our example above our testing might tell us that the null hypothesis that fecundity is the same in the *control* and *resistant* groups can be rejected. However, we have to turn to our preliminary investigation of the data to describe the nature of the effect, in this case we can conclude on the basis of the sample means that it appears that selection for resistance reduces fecundity from around 33 eggs per day to around 25 eggs per day.

So far we have considered the situation where we only measure one trait on each individual in our sample, but quite a lot of scientific questions require us to measure two traits on each sample and look at how they are associated; we will explore how to do that in the next Chapter.

Looking for a relationship between two measured variables

3

This Chapter deals with looking for a relationship between two traits that are quantitative variables measured for each individual in a sample. If one of your variables is qualitative or if the number of values is less than five, then refer to Chapter 2. If both your variables are qualitative or have less than five values then refer to Chapter 4.

In this Chapter we will show you how to:

- *draw and interpret a scatter plot of the two traits (section 3.1)*
- *use a statistical test to measure the strength of association between the two traits (section 3.2)*
- *produce and interpret a line of best fit by simple linear regression (section 3.3)*
- *investigate a non-linear association (section 3.4).*

A common situation in biology is where we have measured two different traits on each individual in our sample and we want to understand how those two traits are related. For example, we might be interested in how running speed is related to age across a sample of school teachers; or how the heights of tomato plants are related to the mass of fruit that we harvest from each of them; or how the size of different fields are related to the bird diversity that we record in each. This Chapter will introduce the tools needed to explore this situation.

Notice that here we are *measuring* two variables. There is not a variable set by the experimenter (the independent variable) and one that is measured (the dependent variable) as there is in a controlled experiment to investigate cause and effect. Rather we are exploring if there is a correlation or association between the two variables. If there is a relationship, we cannot say for sure which variable is affecting the other or indeed if there is a third variable that is affecting them both. Although as a result of our correlation study we may be able to suggest a hypothesis concerning the two variables that can be tested by experiment.

Notice also that we are interested in two measured traits, each of which could take on broad range of possible values. Imagine we did ask a sample of school teachers to run 100 m and we recorded their times with a stopwatch. We record with accuracy of perhaps a second, but we would expect times to vary between perhaps 15 seconds and 40 seconds. So for one of the traits (running speed) we would expect a broad range of possible values. Since teachers might be as young as 25 or as old as 65, if each provides you with their age to the nearest year then for the second trait (age) we would again expect a broad range of possible values. In this situation, the methods described in this Chapter should be an effective way to investigate the data. However, if we felt it was potentially embarrassing to ask the teachers to reveal their exact age, we might not ask them this but instead categorize each teacher as either “younger”, “middle-aged” or “older” based on our perception of their appearance. In this case age no longer takes on a broad range of possible values but is restricted to three possible categories; we can still explore whether there is a difference in characteristic running speeds between the three categories - but now the methods introduced in Chapter 2 would be the best approach to use. As a general rule of thumb, you should switch to using the methods introduced in Chapter 2 if one of the traits that you measure on each individual in the sample has less than five values. If both traits are restricted to such a small number of values then we will discuss how to handle data of that sort in Chapter 4.

However, let us now focus on the situation where both measured traits are quantitative variables that can take on a broad range of numerical values. As always we first encourage you present your data graphically.

3.1. HOW TO DRAW AND INTERPRET SCATTER PLOTS

In Chapter 1 we looked at the heights in cm of thirty 11-year old girls, these heights were 135, 146, 153, 154, 139, 131, 149, 137, 143, 146, 141, 136, 154, 151, 155, 133, 149, 141, 164, 146, 149, 147, 152, 140, 143, 148, 149, 141, 137 and 135. Imagine that we now recorded the mass of each of these girls to the nearest kilogram and got the values for each individual girl (in the same order as the heights above) as 26, 33, 55, 50, 32, 25, 44, 31, 36, 35, 28, 28, 36, 48, 36, 31, 34, 32, 47, 37, 46, 36, 47, 33, 42, 32, 32, 29, 34 and 30. We might expect that taller girls are also heavier; one way to explore this would be to plot the data as a scatter plot. Here we select one trait to be our x-axis and

the other as our y-axis (there is no convention as to which trait should be on which axis as we do not have a dependent and an independent variable) and plot a point for each individual in the sample; for our data this is shown in Figure 3.1.

Notice our graph has an informative title and labels for both axes that explain the units. It does seem from visual inspection of the graph that as a generality taller girls are heavier; although we note that this is only a general trend and clearly factors other than height influence mass, so it would be possible to find two girls in our sample where the taller one is lighter. We can see from our graph why we use the term scatter plot. If there was a strong correlation between the two traits, the symbols on the graph would be close to a straight line (a linear association). In our case all the girls on our sample do not fall on a single line, but are scattered around such a line.

It seems clear with our data that there really is a trend for taller girls to be heavier. Another way to say this is that there is a positive association between our two variables. You can also have negative associations, where higher values of one trait are associated with lower values of the other. For example, if we plotted body mass and litter size for a sample of different mammal species we might expect to find a negative association where those species with high body mass produce generally fewer offspring at one time.

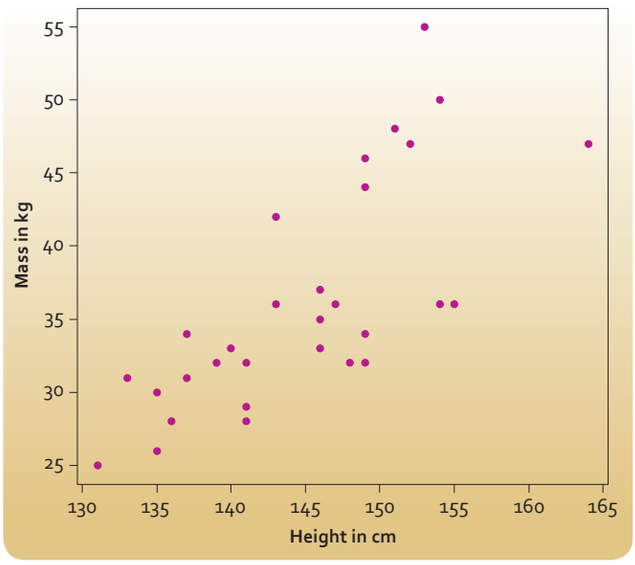


Figure 3.1 *- Scatterplot of the height and mass of a sample of 30 11-year old girls from Rottenrow Primary School.*

For other data-sets two people might disagree on whether they can see evidence of a positive or negative trend in the data. It would be useful to have some type of objective measure of association.

3.2 TESTING FOR LINEAR ASSOCIATION (PEARSON CORRELATION)

If we have two traits measured on each of a sample of individuals then we can use *Pearson's product moment correlation coefficient* to measure the strength of association between those two traits. It is very easy to implement this in ***R*** through the *cor.test* function, for our data we simply type

```
heights <- c(135,146,153,154,139,131,149,
137,143,146, 141,136,154,151,155,133,149,
141,164,146,149,147,152,140,143,148,149,
141,137,135)

masses <- c(26,33,55,50,32,25,44,31,36,35,
28,28,36,48,36,31,34,32,47,37,46,36,47,33,
42,32,32,29,34,30)

cor.test(heights,masses)
```

and ***R*** will respond with

```
Pearson's product-moment correlation

data: heights and masses

t = 5.8631, df = 28, p-value = 2.647e-06

alternative hypothesis: true correlation is

not equal to 0

95 percent confidence interval:

0.5215983 0.8699620

sample estimates:

    cor

0.7423653
```

The most important number in this output is the last one (approximately 0.74), this is the estimate of *Pearson's product moment correlation coefficient* for our data. Since this measure has a long name, it is sometimes shorted to *Pearson's r* or even just *r*. *Pearson's r* will always take a value between -1 and 1. In our case it is a positive value, indicating that the data are suggestive of a positive association. The closer the value is to either 1 or -1 the closer the

points on our scatter plot would be to a single straight line. If *r* were 1 then all the individuals in our sample would fall perfectly on a single straight line.

Imagine if we also measured leg length of each girl and found that the r value between height and leg length was higher (say 0.85), than between height and mass (0.74). What could we conclude? This would suggest that height is a better predictor of leg length than of mass. That is, if we know an 11-year old girl's height we would be more confident about our ability to use this knowledge to predict her leg length than her mass. Another way to think of this is that the scatter plot for mass against height would show more scatter around a straight line than one of leg length against height. We will talk about such predictions in the next section, but before we do that, we note that cor.test also offers us a p-value. This p-value is associated with the null hypothesis of no association (in our case that a girl's height and mass are entirely unrelated). The p-value is very low (2.647×10^{-6}), so our data gives us grounds to reject this null hypothesis and conclude that the two traits are related. Our scatter plot tells us that the relationship is likely to be a positive association - and this is in line with our expectation based on our general knowledge.

3.3 PRODUCING AND INTERPRETING A LINE OF BEST FIT (LINEAR REGRESSION)

We have talked in previous sections about how closely a scatter of data points might correspond to a single straight line, but it would be useful to know which particular straight line best represents the data. There is a statistical technique called *simple linear regression* that can estimate this line (called the *line of best fit*). This is easily implemented in ***R*** if we simply type **lm(masses ~ heights)** then we get the following output

```
lm(formula = masses ~ heights)

Coefficients:

(Intercept)      heights

 -71.3706        0.7427
```

Before we turn to the output, let's demystify the input: **lm** simply calls a function in ***R***, and we need to tell that function first the variable to have on the y-axis (*masses* in our case) then the variable to have on the x axis (*heights* in our case), and we separate these variables with the ~ symbol. To specify the formula for a straight line we need to know the intercept (the value of the y-axis when the value of the x-axis is zero) and the gradient of the line, and the output gives us these (approximately -71 and 0.74 respectively). This means that if M is the mass in kg and H is the height in cm, then the straight line that best approximates our data has the formula

$$M = -71 + 0.74H$$

(We round these values to two significant digits rather than use the very precise values given by ***R*** to avoid anyone thinking that we were much more precise in how we took our measurements that we really were - see Section 5.1 for more discussion on this.)

To see if this equation seems plausible, we could use it to estimate the mass of a girl of height 150 cm say. Substituting H = 150 into the equation, gives a predicted mass of 40 kg. Looking at our data this seems plausible. We can add our line of best fit to our scatter plot.

The intercept (-71 in our case) is often where the line cuts the y-axis, but you can see in our scatter plot (Figure 3.2) that the line cuts the y-axis at 24. This is because we started our x-axis at 130 rather than zero to avoid lots of blank space on our graph. If we had drawn the x-axis starting from zero, the intercept of the line of best fit on the y-axis would be seen to be at -71. What we would expect if there truly is a straight line relationship between the two variables

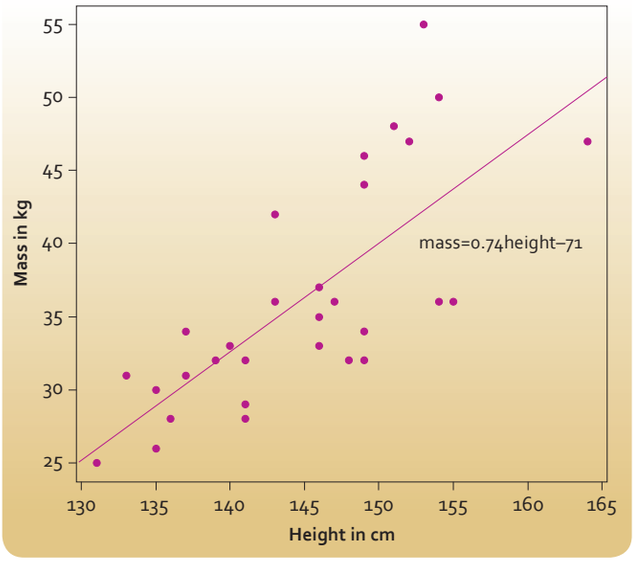


Figure 3.2 *- Scatter plot of the height and mass of a sample of 30 11-year old girls from Rottenrow Primary School, with added line of best fit.*

is that we could pick any section of the line and there would be roughly as many data-points above that segment of the line as below it. This seems pretty much true for our data so we can conclude that there is generally a linear association between height and mass in our sample of 11-year-old girls - we can describe that association by finding the line of best fit, and we can describe the strength of the association (how close data points lie to the line of best fit) using Pearson's *r*. The line of best fit also allows us to predict one variable if we know the other (the process of interpolation). But remember that we don't expect all individuals to lie exactly on the line. Our best guess for the mass of a girl of height 150 cm is 40 kg, but this doesn't mean that we expect all girls of that height to be of that mass. What that does mean is that if we sampled a few girls of height 150 cm we would expect them to have a range of masses, but the mean of their masses would be close to 40 kg.

Notice that we can choose which variable to have on the x-axis and which to have on the y-axis. It does not matter which you pick, but if there is one variable you are interested in predicting on the basis of the other then the one you are interested in predicting should go on the y-axis.

Image: Stockimages/FreeDigitalPhotos.net

In fact, *Pearson's r* and the line of best fit are linked. If we multiply *r* by itself, then r^2 is called the *coefficient of determination* for the line of best fit. This will be a number between zero and one, the closer this value is to one, the closer the points on our scatter point will be to all falling on the line of best fit. The higher the coefficient of determination the better the x-variable is as a predictor of the y-variable. For example in our case the coefficient of determination is 0.55 (0.74 x 0.74 = 0.55). There is variation between 11 year old girls in their mass and this variation can be described by its variance (as discussed in section 1.3). If the coefficient of determination of our line of best fit is 0.55, then 55% of the variance between girls in mass can be explained by variance between girls in their height. So variation between girls in their height can only explain about half of the variation between girls in mass. This tells us that girls are not all exactly the same shape, so height is a useful but not perfect predictor of mass; something that agrees with our intuition.

As usual, you don't have to use ***R***. If you like Excel then the CORREL function will calculate *Pearson's r* for you, and the Regression function in the *Analysis Toolpak* will do simple linear regression. Otherwise, just type "correlation coefficient calculator" or "line of best fit calculator" into your favourite search engine.

3.4 DESCRIBING NON-LINEAR ASSOCIATION

Not all relationships between two variables will be linear. In maths, physics and chemistry, data are often plotted on a graph to illustrate a known linear relationship or to determine the formula for a linear relationship; hence why in these situations a straight line (often a line of best fit) is drawn through the plotted points. In biology, experiments where there may be naturally occurring variation in the data and where uncontrolled confounding variables may also be present, things are much more uncertain. This is why once the data are plotted we should test for correlation using *Pearson's r* and, if appropriate, use simple linear regression to determine a line of best fit. If these techniques do not give a clear result then it is safer to join the plotted points with a series of straight lines to visualise any trends or patterns. Where you have a graph where you have 'joined the dots,' you cannot estimate values between the points from the graph (interpolation) or beyond the measured values (extrapolation). However if your plotted data has sections which you think show a linear effect we give you some advice overleaf on how you could proceed.

Imagine plotting the heights of humans as a function of age for people aged between 5 and 25. You would probably expect that perhaps up until age 15 there is a steady rise in height with increasing age, but from 15 to 25 there is little change in height with age. In this case we would expect that a single linear relationship would not offer a good description of the data. If on plotting your data you find such a non-linear effect then our advice would be to use your plot of the data to split your data up into sub-ranges and analyse these separately; so in the case above we might split our dataset into those ages 15 and under and those aged over 15 and explore those two parts of our sample separately.

You will sometimes encounter even more complex patterns of association. Imagine plotting data on how often per year people were seen by a general practitioner as a function of age. Your expectation might be that frequency of visits to the GP will be high in the first years of life, but decline to lower levels that remain relatively constant through ages from around 15 - 45 years before starting to climb again in later life. Our advice on tackling such complex patterns is once again to first of all plot the data and use the visual appearance of the plot to allow you to identify sub ranges of the data to which the methods proposed here could usefully be applied.

3.5 CHAPTER CONCLUSION

In this Chapter we have focused on situations where for each experimental subject in our sample we have two measured traits each of which could take on a broad range of possible values. We recommended at the start of this Chapter that you should switch to using the methods introduced in Chapter 2 if one of the traits that you measure on each individual in the sample is restricted to four or fewer possible values. In this next Chapter we will discuss how to proceed if both traits are restricted to a small number of values.

4 Dealing with categorical data

This Chapter deals with qualitative (categorical) variables. Qualitative variables provide categorical data where individuals are assigned to separate categories.

In this Chapter we will show you how to:

- *present a sample of categorical data in a table and bar chart (section 4.1)*
- *use a statistical test to examine if the distribution of the numbers of individuals in each category of a sample is evenly spread (section 4.2)*
- *use a statistical test to examine if the distribution of the numbers of individuals in each category of a sample is similar to a theoretical distribution (section 4.3)*
- *use a statistical test to examine if the distribution of the numbers of individuals in each category of two samples differ (sections 4.4 and 4.5).*

Experiments generally involve study of a sample of individuals. Up until now we have been interested in how to handle data where you take measurements on each individual in your sample to obtain a number. This might be the weights of different dogs, or the numbers of leaves on different seedlings, or the IQ scores of different people. Instead of such measurement data, you sometimes gather categorical data, where for every individual in your sample you see which level of a category they fall into. For example, if we trap a sample of field mice; we might categorise each individual in terms of their sex into one of two levels (male or female). For each mouse, we do not obtain a measurement (i.e. a number) but rather a categorisation as either male or female. Here we are going to explore how you can make the most of such data. If you caught the mice and sexed them (a categorical variable) and measured their mass (a continuous variable) and wanted to test whether the sexes differed in their average mass then the methods in Chapter 2 should help you.

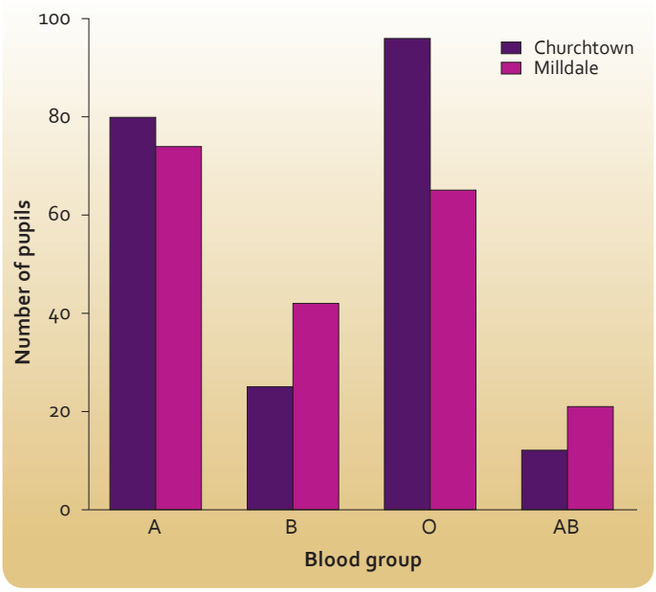


Table 4.1 *- ABO blood groups for samples of children at two schools.*

4.1 PRESENTING CATEGORICAL DATA

Categorical data lends itself to presentation in a table. Imagine that you measure the blood type (A, B, O or AB) of samples of children in each of two schools (called Churchtown and Milldale). You might present the results as shown in table 4.1 below.

School	***Blood Type***			
	A	**B**	**O**	**AB**
Churchtown	80	25	98	12
Milldale	75	42	65	21

Table 4.1 *- ABO blood types of samples of children taken from two schools.*

It is always a good idea to tabulate the data, as this tells the reader the exact values, but it might be easier for the reader to see trends in the data if you also present them graphically in a bar chart, as in Figure 4.1.

Often we process categorical data to show the proportions in each category. We can do this either by expressing the number in each category as a percentage of the whole sample (or as a decimal fraction of 1.0), as shown in table 4.2.

School	***Blood type***			
	A (%)	**B (%)**	**O (%)**	**AB** (%)
Churchtown	37	12	46	6
Milldale	37	21	32	10

Table 4.2 *- Percentages of ABO blood types of samples of children taken from two schools.*

Looking at these three ways of expressing the data we can see that across both schools A and O tend to be the most common types, with B being less common and AB being the least common of all. There appears to be a difference between the two schools as well, with numbers of A individuals being broadly similar between schools, but Milldale seems to have a lower fraction of O individuals with proportionately more B and AB individuals.

Note that (in this simple case) by expressing the results as percentages (table 4.2) or as a bar chart (Figure 4.1) we do not learn much more than by looking at the raw data (table 4.1). That is why we should consider processing the results using a statistical test. This will provide us with more information than simply carrying out an arithmetic percentage calculation for each category.

Before we leave presentational matters, note that it is often useful to provide row and column totals in tables of collected data.

School	***Blood type***				
	A	B	O	AB	Totals
Churchtown	80	25	98	12	215
Milldale	75	42	65	21	203
Totals	155	67	163	33	418

Table 4.3 *- ABO blood types of samples of children taken from two schools (with row and column totals).*

4.2 COMPARING AGAINST A UNIFORM DISTRIBUTION

A common question to ask is whether your sample of individuals is spread evenly across the different possible levels of a category (a uniform distribution). The chi-squared test is a very flexible statistical test for use with categorical data, and can explore this question. If in the population of all Scottish school children there were equal numbers of individuals with all four blood groups, then we would expect approximately equal numbers in our total sample drawn from these two schools. The randomness associated with sampling means that we should not expect exactly equal numbers. With a large sample size of 418 we would not expect large deviations away from uniformity (i.e. from all the counts across all the categories being the same). Specifically, if it were true that all blood types were equally as common in the Scottish school population then we should expect approximately 105 individuals in each group (418 ÷ 4 = 104.5). This expectation that all four groups would be the same size is our null hypothesis. Our data suggests that this null hypothesis is unlikely to be correct, since the A and O groups seem to be a lot more common and the B and AB groups a lot less common than expected under this null hypothesis. We can test this using a chi-squared test. For our example this is very easy to do in ***R***, we simply type

```
chisq.test(c(155,67,163,33))
```

and ***R*** carries out all the calculations for us and responds with

```
Chi-squared test for given probabilities

data:  c(155, 67, 163, 33)

X-squared = 119.5311, df = 3, p-value < 2.2e-16
```

Remember the important thing for us is whether the p-value is less than 0.05. In this case the number is so small (2.2×10^{-16} - which ***R*** displays as "2.2e-16") that ***R*** has given it in scientific notation. This is a tiny number, much less than 0.05, so we are safe to conclude that we seem to have evidence that the null hypothesis that all four blood types are equally common in the wider population can be rejected on the basis of our data. ***R*** has also reported the chi-squared statistic (119.5311), and the degrees of freedom (3). You don't need to know the specifics, but these values were used to calculate the p-value. It's good practice to quote them alongside your p-value , so in this case you might write a conclusion along the lines below.

*"If the null hypothesis were true that all blood types were equally as common in the Scottish school population then we should expect uniformity of approximately 105 individuals in each group (418 ÷ 4 = 104.5) in this sample. The data in table 4.3 suggests that this null hypothesis is unlikely to be correct, since the A and O groups seem to be a lot more common and the B and AB groups a lot less common than expected if this null hypothesis were true. This was confirmed by using a chi-squared test performed in **R** which gave the results that chi-squared was 119.5311, degrees of freedom were 3, and the p value was 2.2×10^{-16}. This p-value is much less than 0.05, suggesting that there is statistically significant evidence for deviation from uniformity across the four blood types."*

4

Extra Notes on using the Chi-squared test

1) Chi is a letter in the Greek alphabet, and so you will sometimes see chi-squared written as χ^2.
2) If you don't want to use ***R***, you can find websites to do a chi-squared test for you. The type of test we performed above was called a "goodness of fit" test since we tested how well our data fitted the null hypothesis of equal numbers across all the categories. If you type "goodness of fit calculator" into your favourite search engine then you will find a number of suitable sites. One that we find easy to use is http://graphpad.com/quickcalcs/chisquared2/.
3) These websites will ask you to input both observed and expected values, the observed values are those we actually observed in our experiment (155, 67, 163 & 33 in the case above); the expected values are those we would have expected if the null hypothesis of uniformity were true (104.5,104.5,104.5,104.5 in the case above).
4) Don't worry if a website gives slightly different values for χ^2 and the p-value to those given by ***R*** for the example above, there are a number of very slightly different ways of doing the chi-squared test.
5) Chi-squared tests can be unreliable if your sample sizes are too small, so aim to have all your observed values greater than 5.
6) There is no logical order to our categories, so it doesn't matter to the statistical test what order you give the observed values, the answer the chi-squared test gives will be exactly the same.

4.3 COMPARING AGAINST A NON-UNIFORM DISTRIBUTION

Sometimes we might want to compare our pattern of observed counts across levels of a category with a more complex theoretical prediction than uniformity. If we return to our blood type example, then there have been many studies of the UK distribution of blood types, since such information is vital to hospitals that need to have the appropriate amounts of different types of blood available for transfusions. These studies suggest the following distribution across blood types: A (38%) B (10%) O (47%) AB (5%). For our total sample size of 418, this distribution suggests we should expect to have observed the following numbers: A (159), B (42), O (196) and AB (21). We in fact observed A (155), B (66), O (163) and AB (33). Our observations, are pretty similar to the predictions, although the number of O individuals is a little low and the numbers of B and AB individuals a little high. But is this deviation sufficient to suggest that our sample is significantly different from the theoretical distribution based on previous studies? Again, we turn to a chi-squared test to help. If we enter the following code into ***R*** then it will do the test for us:

```
chisq.test(c(155,67,163,33), p=c(38/100,
10/100,47/100,5/100))
```

R gives the following output

```
Chi-squared test for given probabilities

data:  c(155, 67, 163, 33)

X-squared = 27.9892, df = 3, p-value = 3.651e-06
```

The p-value here is very small (3.651×10^{-6} - which ***R*** writes as "3.651e-06"). From our argument before about how to interpret p-values, this suggests that our pooled samples from the two schools do seem to deviate statistically significantly from the distribution of blood types expected on the basis of previous studies of the UK population as a whole.

You can see that previously when testing against a uniform distribution, we only had to supply ***R*** with a list of the values that we observed in our study. Now when we are testing against some other distribution, we need to tell ***R*** what that distribution is, and we do this by presenting the distribution as a list of fractions. Those fractions should add up to one, and should be in the same order as the observed values: so the observed value of 155 and the fraction 38/100 both relate to blood type A.

Again if you don't want to use ***R*** the same websites that we recommended in the last section will work in this case, although they will often ask you to input your specified distribution as a list of expected values rather than a list of fractions, but it is easy to work out these, as we did above.

4.4 COMPARING SEVERAL DISTRIBUTIONS

Another question you could ask in our blood-test study is whether there is a difference in the distribution of blood types between the two schools. Once again the chi-squared test can do this for us. This question involves comparing two observed distributions to see if they seem to differ in any fundamental way. This involves looking for an *interaction* (sometimes called a *contingency*)

between blood type and school. Essentially we are asking if our expectation of a randomly selected individual's chance of being of a given blood type should change if we are given the additional information of what school they attend. Other ways to say this are:

is the distribution of blood types *contingent* on school,

or

is there an *interaction* between blood type and school?

For this reason, scientists often call tables like table 4.1 a *contingency* table. We will need to build such a table in ***R*** and this is done using the following command:

```
data <- rbind(c(80,25,98,12),c(75,42,65,21))
```

This does look weird but we are listing the two rows of the table in the way we have always done: a list separated by commas and enclosed in brackets with the letter "c" at the front. We ask ***R*** to bind those rows into a table, and to call that table "data". When we enter this into ***R***, then ***R*** doesn't seem to do anything, but if we then type chisq.test(data), then it will carry out the test on this table and gives us the output

```
Pearson's Chi-squared test

data:  data

X-squared = 13.2767, df = 3, p-value = 0.004075
```

From this we can conclude that there is a statistically significant difference between the schools in the distribution of pupils across the four blood types.
If we inspect the table or figure to help us infer the difference between the two schools; it seems that Milldale has somewhat higher fractions of B and AB individuals and a lower fraction of O individuals.
We can compare any number of distributions this way. Imagine we visited a third school and recorded 56 A individuals, 23 Bs, 62 Os and 12 ABs, then we can simply compare across all three schools by typing

```
moredata <- rbind(c(80,25,98,12),c(75,42,65,
21),c(56,23,62,12))

chisq.test(moredata).
```

If this was significant that would tell us that there was a difference between schools, but the test does not tell us what that difference is. That is, it does not tell us which particular school is different from which particular other school or schools, and what that difference is. We just have to infer what that difference is likely to be from looking at the numbers in a table and/or figure.

If you don't want to use ***R*** then you can find ways to analyse contingency tables on the internet. Just type "contingency table calculator" into your favourite search engine. A good one is http://vassarstats.net/newcs.html

We are going to leave this blood-test example now, but let's not forget that we are biologists first, not statisticians. We have found that these schools differ from each other in terms of their distribution of blood types, and the two schools together seem to deviate from the wider UK population. How could we explain that?

The most likely explanation lies in that fact that different regions of the world have different distributions of blood types. What we have noticed is that Milldale school children have a higher prevalence of the B and AB types. These types are proportionally more common in the Indian subcontinent, so our guess is that the explanation for our results comes from recent immigration to the UK. It is likely that Milldale school in particular has a relatively high proportion of pupils whose parents or grandparents were born in the Indian subcontinent. If you want to read more about the ABO blood groups and how they vary around the world, then a really good website is http://www.blood.co.uk/about-blood/blood-around-the-world/.

Just to introduce you to nomenclature that you might see in textbooks; table 1 has 2 rows of data and 4 columns, so you would find it referred to as a 2 x 4 contingency table. You could imagine redesigning that table so the columns where in a different order; or so it had 2 columns and 4 rows; neither of those changes would make any difference to the outcome of any chi-squared tests you carry out on that contingency table.

4

4.5 TESTING FOR AN INTERACTION

In a sense we don't need this section at all, since when we compared two different distributions before we were really testing for an interaction. However, interactions are so fundamental to so many scientific investigations that we think it's worth showing you another example.

Imagine you wanted to investigate whether there was a sex bias in susceptibility to infection by ticks in field mice. This is a question about an interaction, because you are asking *is there an interaction between sex and susceptibility* to ticks. Another way to express this is: you are asking whether your expectation of the likelihood of an individual mouse having ticks would change if you were given information on the sex of the mouse. If your expectation would change then this suggests that there is an interaction between tick infection status and sex; if your expectation would not change then this suggests that the two factors are *independent* and there is no interaction.

Imagine that you go out and collect the data in Table 4.4.

Sex	***Tick status***	
	Ticks	*No ticks*
Male	17	14
Female	25	7

Table 4.4 *- The sex and tick infestation status of 63 trapped field mice.*

Visual inspection of this data might make you suspect there is an interaction: 55% of males have ticks compared to 78% of females. Following the methodology of the last section, we could use ***R*** to test this statistically using the commands:

```
data <- rbind(c(17,14),c(25,7))

chisq.test(data)
```

This gives the following ***R*** output:

```
Pearson's Chi-squared test with Yates' continuity correction

data:  data

X-squared = 2.8658, df = 1, p-value = 0.09048
```

The p-value is greater than 0.05, and so our data do not provide sufficient evidence to reject the null hypothesis of no interaction. That is, we have no reason to be confident that the sexes of field mice differ in their prevalence of ticks.

4.6 CHAPTER CONCLUSION

One last note: as mentioned earlier, the chi-squared test is unreliable for small sample sizes. This is particularly true for 2x2 contingency tables, so do ensure all the values in your table are above 5 before testing such tables statistically.

We hope you find the chi-squared test an easy-to-use and effective tool whenever you have categorical data.

Presenting and interpreting numerical values

5

This Chapter deals with some general advice on presenting numerical data when you write up reports and on interpreting numerical values given by other people.

In this Chapter we will show you how to:

- *present numerical values consistent with the accuracy of a measurement instrument, with appropriate units and in an ordered way (section 5.1)*
- *make the best use of text, tables and graphs to present numerical data (section 5.2)*
- *interpret standard error and confidence intervals to compare how close the mean value from a sample is to the true mean of the population (section 5.3).*

5.1 REPORTING NUMERICAL VALUES

Report your results within the limits of sensitivity of your measurement device (its resolution). Every instrument you use has a level of sensitivity, so state what it is. For example, "Crabs were weighed on a laboratory balance (Rossiter and sons Ltd; Weightmaster 5, +/- 0.05 g)." Report your results within this level of resolution and make sure any calculated values are in agreement with this precision; for example "Mean crab weight was 120.1 g". Since we can only measure to the nearest 0.05 g, it would introduce a spurious level of precision to quote a mean of "120.11".

If we were combining data, say for example we were going to weigh the crabs again after a period of time and calculate their change in weight; it is good practice to retain an additional figure to reduce accumulated rounding errors. However the final figure should be stated within the limits of the measuring device; "Crab mean weight gain was 0.2 g".

The same principle applies when converting units. If you read in a book that a dinosaur was "about 20 feet high", but you want to convert this into metres, you should convert it to "about 6 m" not "about 6.096 m". It is true that a foot converts to 0.3048 m, but we should reflect the level of uncertainty that existed in the original description.

Always give units and be consistent in your units. Most things that we measure have units, always quote those units every time and be consistent in your units. If you are reporting masses don't flip between giving values in grams and kilograms; use the same unit every time "*Median female mass was 99.5 g (inter-quartile range 89.2 g - 120.3 g); median male mass was 116.4 g (inter-quartile range 105.0 g - 135.9 g).*"

Be consistent in ordering your results to make life easy for your reader. If you are comparing males and females across several analyses and several graphs, always discuss the males then the females in the same order in the text and draw them in the same order on graphs. Similarly if you are discussing different traits of apples, oranges and pears, always compare them in the same order for each different trait. All you are doing is making life as easy as possible for the reader, to minimise the risk of them misunderstanding you.

5.2 USING TEXT, TABLES AND GRAPHS TO PRESENT NUMERICAL VALUES CLEARLY

Avoid giving too many numbers in an abstract. You will often have to provide a summary or abstract at the start of a scientific report, for example the Advanced Higher Project Report. This is the first thing that people read after your title. The reader is looking for a broad overview of the contents of your report; they don't want to get bogged down in detail. Hence, don't go into fine detail like sample sizes and p-values. However, it is appropriate to give your conclusions quantitatively in general terms. Hence you might write "*the application of nitrogen fertiliser increased yield typically by around 70% although there was variation between varieties*" rather than a very detailed (but harder to remember) description like "*fertiliser resulted in a mean increase in yield of 66.3% (range 13.4-130.2%)*".

Make sure your results section isn't just a random collection of numbers. Your results section should have a clear narrative thread through it so the reader can understand not just what information you are giving them but why you are giving them that information. One way to achieve this is to finish your introduction with a clear list of the specific issues that you want to investigate. For example, you might state that you "*want to investigate (i) the size range of a particular species of locally occurring crab, (ii) whether the two sexes differ in size, and (iii) how sex and size influence diet choice*".

You could usefully use these three key issues as three subheadings of the results section, to introduce a clear structure and remind the reader what your main aims are. Also, you should strive to write sentences in your results section that summarise your results, even if these sentences involve a lot of

5

numbers. This shows the reader that you can 'see' your results and help them to do so too; see some examples below:

We measured the mass of 124 crabs ranging from 101.1 g to 150.7 g (mean = 110.6 g, standard deviation 7.3 g). See Appendix A for a histogram of these raw values.

Median female mass was 99.5 g (inter-quartile range 89.2 g – 120.3 g); median male mass was 116.4 g (inter-quartile range 105.0 g – 135.9 g).

Because the two samples we wish to compare did not follow a normal distribution, we adopted a Mann-Whitney U-test rather than a t-test. This suggested a significant difference between the two distributions (N1 = 60, N2 = 63; U = 27, P = 0.008).

Think about whether you should provide data in text form, in a table or in a graph. Ask yourself what you want the reader to do. If the reader needs to consult specific values then present those values in a table; but if what you want them to do is see trends or look for relationships then a graph will work better. You should never have a table with four or fewer values in it, such a small amount of information can more directly be given to the reader in a single sentence in the text, rather than asking them to stop their reading and consult a table.

Checklist of some things to consider about a table

- Have you given a sufficiently full and informative title that the table can be understood without reference to the text?
- If relevant, have you given sample sizes?
- Are your row and column titles sufficiently informative?
- Have you included units of measurement?
- Have you ordered the rows and columns in a helpful fashion and is that ordering consistent across tables (if you want the reader to compare values in two columns try and place those columns side by side)?
- Can you remove clutter and redundancy (e.g. if all the columns are in the same units then you could give this unit in the title rather than repeating it in every column)?

Compare the two tables below; hopefully you can see that the second is much preferable.

Diet	***Heart rate (beats/sec) (SD)***		
	1st	*2nd*	*3rd*
A	51 (2.3)	49 (2.2)	48 (2.2)
B	49 (2.5)	49 (2.7)	47 (3.1)
C	48 (3.1)	47 (2.3)	45 (2.3)
D	53 (2.3)	52 (2.1)	52. (2.4)

Table 5.1 *- Results of the diet and heart rate experiment.*

Diet	***Heart rate (beats/sec) (SD)***		
	1st test (7 days)	*2nd test (14 days)*	*3rd test (28 days)*
Diet A (no caffeine)	51 (2.3)	49 (2.2)	48 (2.2)
Diet B (no alcohol)	49 (2.5)	49 (2.7)	47 (3.1)
Diet C (no caffeine or alcohol)	48 (3.1)	47 (2.3)	45 (2.3)
Diet D (control)	53 (2.3)	52 (2.1)	52. (2.4)

Table 5.1 *- Resting heart rate (mean (standard deviation)) measured at three time points after individuals were randomised to one of 4 dietary regimes*.*

** Normal diet (Group D), normal diet but avoiding caffeine (Group A), normal diet but avoiding alcohol (Group B), normal diet but avoiding both caffeine and alcohol (Group C). There were 12 individuals in each group, except for the last measurement for Group D where N = 11.*

Checklist of some things to consider about a graph

- Have you given a sufficiently full and informative title that the graph can be understood without reference to the text?
- If relevant, have you given sample sizes?
- Have you labelled both axes in a full and useful way?
- Have you included units in your axes labels where appropriate?
- Have you ordered elements in the graph in a helpful fashion and is that ordering consistent across graphs?
- If you have multiple data types in your graph (e.g. circles for males and crosses for females on a scatterplot) then make sure this is explained in the title or in a legend and make sure you are consistent across graphs).

5.3 USING STANDARD ERROR AND CONFIDENCE INTERVALS TO DESCRIBE HOW CONFIDENT YOU ARE THAT THE MEAN FROM YOUR SAMPLE IS CLOSE TO THE MEAN FROM THE POPULATION YOU SAMPLED

5.3.1 What are standard errors and confidence intervals?

Imagine you wanted to estimate the number of nettle plants in a local field. It seems too daunting to try and count them all. You take a square wooden quadrat frame of size 1 m and throw this randomly 40 times within the field, every time it lands you count the number of nettle plants inside the frame; from this you calculate an estimate of mean nettle density (plants/m^2). From Google Earth or an Ordnance Survey map you are able to estimate the area of the field in square metres, which you multiply by the mean nettle density estimate to come up with an estimated number of nettles in the field. This is a reasonable but relatively crude measure of the real total number of nettles. Imagine you perform this calculation and come up with the estimate 18 319. We recommend that you present this calculation, but then add a sentence saying something like "bearing in mind the sources of imprecision in our method, we conclude that there are around 18 000 nettles in the field." If you present only the number 18 319 then the reader might assume that you believe you know the exact number, and that it is 18 319, and not 18 321 for example. But given the limitations of your method, you would not be surprised if the actual number was 18 321, but you might be surprised if the number was as big as 19 455. Hence, by saying you estimate the number to be around 18 000, you are suggesting that you would not be surprised if the actual number were somewhere between 17 500 and 18 500. You are showing the reader that you have an appreciation for the likely level of precision of your techniques.

Another way to cope with explaining the likely precision of your estimated value is to offer the reader a range of values that you think the true value is likely to lie within. This is called a confidence interval. Imagine in the case above you feel that your greatest source of error will lie in your estimate of the density of nettles rather than your estimate of the area of the field. You can quantify this uncertainty rather easily since you have 40 estimates of density, from which you worked out a mean value. You could also work out a standard deviation (see Chapter 1) as a measure of the variability across your 40 estimates. From this it is easy to work out something called the *standard error*, because this is just the standard deviation divided by the square root of the sample size (square root of 40 in our case). We can use the mean and the standard error to estimate something called the 95% confidence interval. This involves two numbers either side of the mean which we expect will bracket the true value on 95% of occasions. These two numbers are simply the mean minus twice the standard error and the mean plus twice the standard error. We would take these two values for density and multiply them by our estimate of area to report something like "*we estimate that there are around 18 600 nettles in the field (95% confidence interval 17 700 - 18 900)*". This tells the reader that they should be pretty confident the true number of nettles lies between 17 700 and 18 900. Notice that there is no such thing as a 100% confidence interval, we can never be certain unless we actually go to the trouble of counting all those nettles; when we sample we have to live with a little uncertainty. But since standard error reduces with increasing sample size, the more we sample the greater we can expect our precision to be. Lastly note that *standard error* and *confidence interval* are sometimes abbreviated to *SE* and *CI* respectively.

5

5.3.2 When are these measures useful?

In Chapter 1 we discussed a range of different ways that you could characterise the spread of values in a data-set: the *standard deviation*, the *inter-quartile range* and the *range*. In this Chapter we have introduced two other measures that are related to this spread of values: the *standard error* and the *confidence interval*. When presenting your results, you have a broad choice of ways to describe the spread of values. The key distinction is to remember that the *standard deviation*, *inter-quartile range* and *range* describe the spread of values in our sample; whereas *standard error* and *confidence interval* describe the consequences of that spread for our confidence in how precisely the mean of our sample of individuals approximates the true mean value of the population that we sampled. You can imagine that if all 11-year-old Scottish girls were very similar in height then the mean height of a sample of 50 girls would be very similar to the mean value we would get if we measured all the girls in Scotland. In fact, there is quite a bit of variation between girls, and that will imply that our mean based on the sample is less likely to be very close to the true mean; but we can use the variation in our sample (together with the sample size) to estimate how much this difference might be: and that difference is encapsulated in the *standard error* or *confidence interval*. So decide whether you are interested in talking about variation in the sample, or interested in talking about how that variation affects the precision of our sample mean when deciding which of these measures of spread to use.

All of these measures of spread can be represented in the same way on bar graphs using error bars, as shown in Figure 5.1.

The error bars here show the values one standard deviation above and below the mean value. Since error bars look the same way regardless of whether you use them to show standard deviation, standard error or confidence interval, it is essential that you state somewhere (most obviously in the title of your figure) which you are using. Since standard error is the standard deviation divided by the square root of the sample size, the error bars will look smaller if you use standard error rather than standard deviation, but that is not a good reason for opting to use standard error; your aim is not to show the reader as small error bars as possible, but to use error bars to explain some scientific point to the reader.

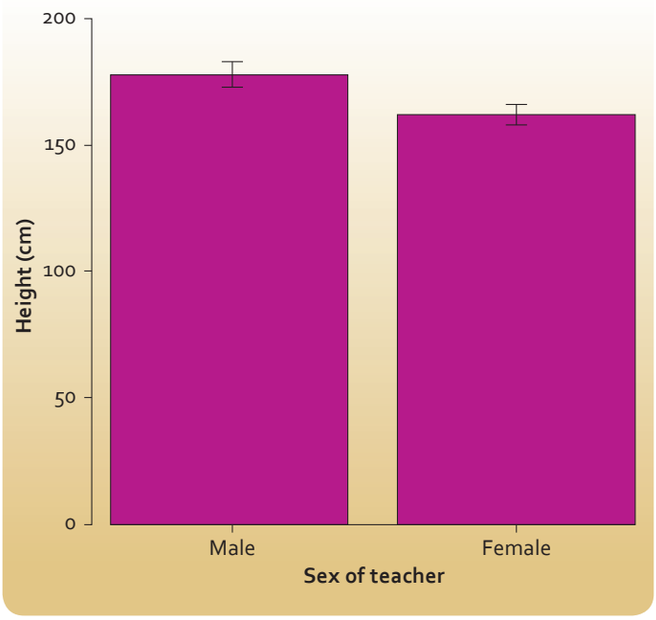


Figure 5.1 *- Mean and standard deviation of the heights of male and female teachers at Bogstonehall Secondary School.*

5.3.3 Comparing two confidence intervals

In Figure 5.1, unsurprisingly, we find that men in our sample have a higher mean height than females; but we also see that there is variation in both sexes. So it is not clear to us whether the greater height of men in our sample is statistically significant or not. The best way to explore that is with a statistical test (as described in Chapter 2), however if you plot confidence intervals as error bars then there are some simple rules that will give you a good guide to what the statistical test is likely to say. We described above how to estimate a 95% confidence interval from the mean, standard deviation and sample size. For our example of male and female teachers, if we measured their sprint speeds this might look like the summary shown in Figure 5.2.

Here we can see that the 95% confidence intervals do not overlap (the lowest value of one is higher than the highest value of the other). When this happens it is very likely that a statistical test will suggest that there is a statistically significant difference between males and females. We cannot be absolutely certain, as our method of calculating confidence intervals by doubling the standard error is only an approximation, but it is a good guide.

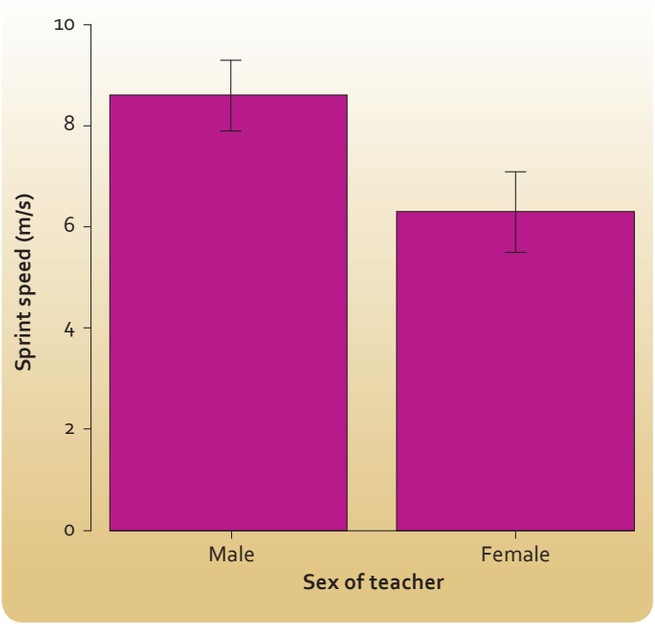


Figure 5.2 *- Mean and 95% confidence interval of the sprint speeds of male and female teachers at Bogstonehall Secondary School.*

If the 95% confidence limits overlap a little bit then it is still likely that the statistical test will suggest a statistically significant difference; but if they overlap by quite a lot then it likely will not. We need a rule of thumb for "overlap a little". If both sample sizes are at least 10, and one confidence interval isn't more than twice as long as the other, then the following rule works quite well.

If the amount of overlap is less than a quarter of the length of the smaller confidence interval, then the two means are probably significantly different.

So if in the case above we had a different group of 23 men and 34 women and the confidence intervals were 9.1-7.7 m/s for men and 8.1-6.6 m/s for women; then we can use the rule of thumb because the sample sizes are big enough and the lengths of the confidence intervals (1.4 and 1.5 m/s) are sufficiently similar. The size of the overlap is 8.1-7.7 = 0.4 m/s. This overlap is 29% of the length of the smallest 95% confidence interval, so our best guess is that the p value in a statistical test will be just a bit bigger than 0.05, from which we would conclude that we do not have statistically significant evidence of a difference in this case between men and women.

You might wonder what use this estimation method is: why not just do the statistical test? We think it is useful in two ways, sometimes when reading someone else's report they will not have provided the test, or provided all the data to let you carry out the test yourself; but if they give you a bar chart with error bars then this method offers you a good approximation to the statistical test. Secondly, in your own work if you get a result you were not expecting from a statistical test comparing two means then this method offers you a quick check. If this method suggests that the result of your statistical test really is surprising then you could go back and check that you have input all your data correctly and performed the test correctly.

5.4 OUR FINAL THOUGHTS

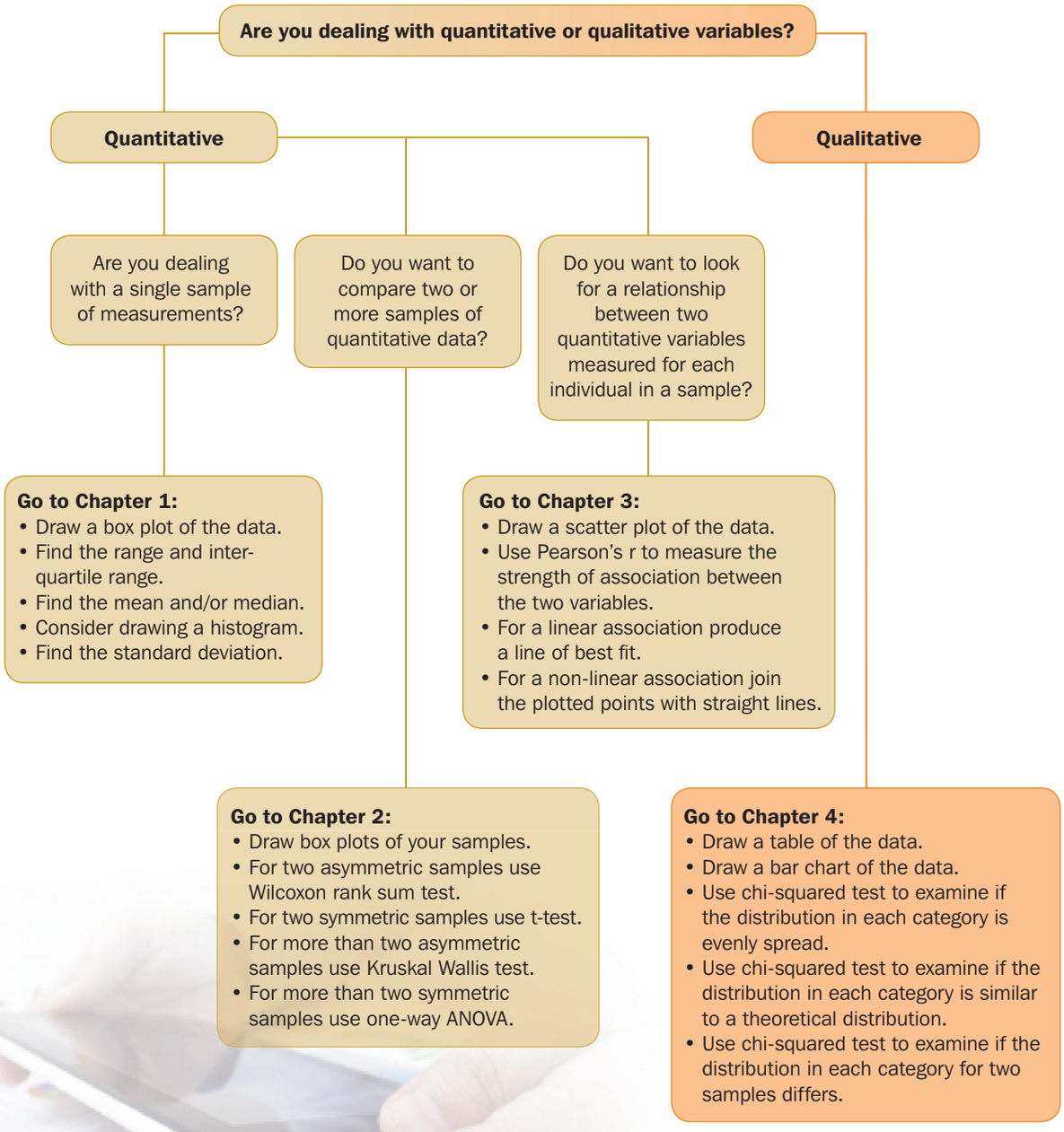
As we said at the start, making the most of your data once you have collected it (using summary measures, tables, graphs and statistical tests) should not be a daunting part of a biology experiment or a research project. We really hope this booklet helps you increase your confidence with statistics, and helps you make the very best of your hard-won data. Good luck!

6 Appendices

APPENDIX 1

Statistical test finder

The flow chart below summarises the information in the 'Choosing your statistical test' section of the Introduction and the contents of Chapters 1, 2, 3 and 4.



6

APPENDIX 2

Summary of statistical skills from school courses in mathematics

Broad General Education
- Rounding numbers.
- Use scientific notation.
- Use fractions, decimal fractions and percentages.
- Use proportions and ratios.
- Use understanding of bias and sample size to evaluate data.
- Evaluate raw and graphical data.
- Find the mean, median and mode of a set of numbers and decide which is the most appropriate to use.
- Display data in tables, charts, diagrams and graphs.
- Use appropriate tables, charts, diagrams and graphs to display discrete, continuous or grouped data.

National 4
- Constructing a frequency table with class intervals from raw data.
- Determine mean, median and mode of a data set.
- Display discrete, continuous and grouped data in an appropriate way.
- Represent raw data in a pie chart.
- Construct a scatter graph.
- Draw a best-fitting straight line.
- Rounding numbers.
- Calculate percentages.
- Calculate percentage increase and decrease.
- Calculate ratio and direct proportion.
- Extract and interpret data from tables, bar and pie charts, scatter and line graphs, stem and leaf diagrams.
- Make and explain decisions based on the interpretation of data.

National 5
- Rounding to a given number of significant figures.
- Compare data sets using interquartile range and standard deviation.
- Determine the equation of a best-fitting straight line on a scatter graph and use it to estimate *y* given *x*.

Statistics stand-alone Unit at SCQF level 6
- Types of data, random sampling, outliers.
- Interpretation of pie charts, bar charts, stem and leaf diagrams, box plots, frequency tables, contingency tables and histograms.
- Interpretation of histograms.
- Mean, median, standard deviation and interquartile range.
- Correlation and linear regression.
- 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.
- Undertake a correlation and regression analysis.
- Undertake a data analysis.

APPENDIX 3

Further reading

If you would like to delve a little deeper into the ideas here then there are some books we can recommend.

Asking Questions in Biology: A Guide to Hypothesis Testing, Experimental Design and Presentation in Practical Work and Research Projects

By Francis Gilbert, Peter McGregor, Chris Barnard. Published by Pearson Education.

The fourth edition was published in 2011, but if you can get an earlier edition second hand those are wonderful too. This book guides you through the whole process of coming up with an interesting research idea, how you can collect data to explore the question you are interested in, how to analyse that data, and how to write up the analysis.

Biomeasurement: a student's guide to biological statistics

By Dawn Hawkins. Published by Oxford University Press.

The third edition was published in 2014, but if you can get an earlier edition second hand then those would be great too. This book will introduce you to more sophisticated statistical techniques that we introduce in this booklet, but this is done in a clear way that should be accessible by later-years school pupils.

Practical Statistics for Field Biology

By Jim Fowler, Lou Cohen and Phil Jarvis. Published by John Wiley and Sons.

This is very much like the book by Hawkins but is smaller. This makes it cheaper although it does not cover the diversity of techniques covered by Hawkins. It will still let you explore statistics in more depth than we have here, and does so in a fantastically clear style.

SSERC (Scottish Schools Education Research Centre), 2 Pitreavie Court, South Pitreavie Business Park, Dunfermline KY11 8UB.
Telephone 01383 626 070, fax 01383 842 793, e-mail sts@sserc.org.uk, website www.sserc.org.uk

SSERC is a Company Limited by Guarantee (Scottish Company No. SC131509) and a registered educational charity (Scottish Charity No. SCO17884) Registered Office – 5th Floor, Quartermile Two, 2 Lister Square, Edinburgh EH3 9GL.

ISBN 978-0-9531776-9-1.