Chapter 3

Collecting, presenting and summarising data

Data are the key to many important management decisions. Is a new product selling well? Do potential customers like the new advertising campaign? Should we launch a new product? These are all questions that can be answered with data. We begin this part of the course with some basic methods of collecting, representing and describing data. We will start off by we looking at the different kinds of data that exist, how we might obtain these data, and some basic methods for presenting them. But first – some important definitions...

3.1 Important definitions

For the work we will look at over the next six months or so, you *must* be familiar with the following words and phrases, and you *must* understand what they mean!

The quantities measured in a study are called *random variables* and a particular outcome is called an *observation*. A collection of observations is the *data*. The collection of all possible outcomes is the *population*.If we were interested in the height of people doing Accounting & Finance courses at Newcastle, that would be our random variable; a particular person's height would be the observation and if we measured everyone doing ACC1012, those would be our data, which would form a *sample* from the population of all students registered on Accounting & Finance degrees.

In practice it is difficult to observe whole populations, unless we are interested in a very limited population, e.g. the students taking ACC1012. In reality we usually observe a subset of the population; we will come back to sampling later in Section 3.2.

Once we have our data, it is important to understand what *type* it is – so we can figure out exactly what to do with it. You should refer to Section 2.1 of the summer revision booklet for full details – the diagram overleaf provides a useful summary – you may want to annotate this!

3.2 Sampling

We can rarely observe the whole population. Instead, we observe some sub–set of this, that is, the *sample*. The difficulty is in obtaining a *representative* sample. For example, if you were to ask the people leaving a gym if they exercised this would produce a *biased* sample and would not be representative of the population as a whole. The importance of obtaining a representative sample cannot be stressed too highly. As we will see in Semester 2, we use the data from our samples in order to make *inferences* about the population and these inferences influence the decision making process.

There are three general forms of sampling techniques.

1. *Random sampling* – where the members of the sample are chosen by some random mechanism.
2. *Quasi–random sampling* – where the mechanism for choosing the sample is only partly random.
3. *Non–random sampling* – where the sample is specifically selected.

3.2.1 Simple Random Sampling

If we had a population of 200 students we could put all their names into a hat and draw out 20 names as our sample. Each name has an equally likely chance of being drawn and so the sample is completely random. Furthermore, each possible sample of 20 has an equal chance of being selected. In reality, the drawing of the names would be done by a computer and the population and samples would be considerably larger. The disadvantages of this method are that we often do not have a complete list of the population. For example, if you were surveying the market for some new software, the population would be everybody with a compatible computer. It would be almost impossible to obtain this information. Not all elements of the population are equally accessible and hence you could waste time trying to obtain data from people who are unwilling to provide it. Thirdly, it is possible that, purely by chance, you could pick an unrepresentative sample, either over– or under–representing elements of the population.

3.2.2 Stratified Sampling

This is a form of random sample where clearly defined groups, or *strata*, exist within the population, for example males and females, working or not working, age groups etc. If we know the overall proportion of the population that falls into each of these groups, we can randomly sample from each of the groups and then adjust the results according to the known proportions. For example, assume that the population is 55% female and 45% male and we wanted a sample of 1000. We could first decide to have 550 females and 450 males in our sample. We would then pick the members of our sample from their respective groups randomly. We do not have to make the numbers in the samples proportional to the numbers in the strata because we could adjust the results but sampling within each stratum ensures that that stratum is properly represented in our results and gives us more precise information about the population as a whole. Such sampling should generally reflect the major groupings within the population.

The disadvantages are that we need clear information on the size and composition of each group or stratum, which can be difficult to obtain; and as with simple random sampling, We still need to know the entire population so as to sample from it.

3.2.3 Systematic Sampling

This is a form of quasi–random sampling which can be used where the population is clearly structured. For example, if you were interested in obtaining a 10% sample from a batch of components being manufactured, you would select the first component at random; after that, you pick every tenth item to come off the production line. The simplicity of selection makes this a particularly easy sampling scheme to implement, especially in a production setting. The disadvantages of this method are that it is not random and if there is a pattern in the process it may be possible to obtain a biased sample. It is only really applicable to structured populations.

3.2.4 Multi–stage Sampling

This is another form of quasi–random sampling. These types of sampling schemes are common where the population is spread over a wide geographic area which might be difficult or expensive to sample from. Multi–stage sampling works, for example, by dividing the area into geographically distinct smaller areas, randomly selecting one (or more) of these areas and then sampling, whether by random, stratified or systematic sampling schemes within these areas. For example, if we were interested in sampling school children, we might take a random (or stratified) sample of education authorities, then, within each selected authority, a random (or stratified) sample of schools, then, within each selected school, a random (or stratified) sample of pupils. This is likely to save time and cost less than sampling from the whole population. The sample can be biased if the stages are not carefully selected. Indeed, the whole scheme needs to be carefully thought through and designed to be truly representative.

3.2.5 Cluster Sampling

This is a method of non–random sampling. For example, a geographic area is sub–divided into clusters and *all* the members of a particular cluster are then surveyed. This differs from multi–stage sampling covered in Section 3.2.4 where the members of the cluster were sampled randomly. Here, no random sampling occurs. The advantage of this method is that, because the sampling takes place in a concentrated area, it is relatively inexpensive to perform.

The very fact that small clusters are picked to allow an entire cluster to be surveyed introduces the strong possibility of bias within the sample. If you were interested in the take up of organic foods and were sampling via the cluster method you could easily get biased results; if, for example, you picked an economically deprived area, the proportion of those surveyed that ate organically might be very low, while if you picked a middle class suburb the proportion is likely to be higher than the overall population.

3.2.6 Judgemental sampling

Here, the person interested in obtaining the data decides whom they are going to ask. This can provide a coherent and focused sample by choosing people with experience and relevant knowledge to provide their opinions. For example, the head of a service department might suggest particular clients to survey based on his judgement. They might be people he believes will be honest or have strong opinions. This methodology is non–random and relies on the judgement of the person making the choice. Hence, it cannot be guaranteed to be representative. It is prone to bias.

3.2.7 Accessibility sampling

Here, only the most easily accessible individuals are sampled. This is clearly prone to bias and only has convenience and cheapness in its favour. For example, a sample of grain taken from the top of a silo might be quite unrepresentative of the silo as a whole in terms of moisture content.

3.2.8 Quota Sampling

This method is similar to stratified sampling but uses judgemental (or some other) sampling rather than random sampling within groups. We would classify the population by any set of criteria we choose to sample individuals and stop when we have reached our quota. For example, if we were interested in the purchasing habits of 18–23 year old male students, we would stop likely candidates in the street; if they matched the requirements we would ask our questions until we had reached our quota of 50 such students. This type of sampling can lead to very accurate results as it is specifically targeted, which saves time and expense.

The accurate identification of the appropriate quotas can be problematic. This method is highly reliant on the individual interviewer selecting people to fill the quota. If this is done poorly bias can be introduced into the sample.

3.2.9 Sample Size

When considering data collection, it is important to ensure that the sample contains a sufficient number of members of the population for adequate analysis to take place. Larger samples will generally give more precise information about the population. Unfortunately, in reality, issues of expense and time tend to limit the size of the sample it is possible to take. For example, national opinion polls often rely on samples in the region of just 1000.

3.3 Frequency Tables

Once we have collected our data, often the first stage of any analysis is to present them in a simple and easily understood way. Tables are perhaps the simplest means of presenting data. There are many types of tables. For example, we have all seen tables listing sales of cars by type, or exchange rates, or the financial performance of companies. These types of tables can be very informative. However, they can also be difficult to interpret, especially those which contain vast amounts of data.

Frequency tables are amongst the most commonly–used tables and are perhaps the most easily understood. They can be used with continuous, discrete, categorical and ordinal data. Frequency tables have uses in some of the techniques we will later on in this chapter.

3.3.1 Frequency tables for categorical data

The following table presents the modes of transport used daily by 30 students to get to and from University (survey date: 3rd August 2012).

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

The table obviously contains much information. However, it is difficult to see which method of transport is the most widely used. One obvious next step would be to count the number of students using each mode of transport:

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	30

This gives us a much clearer picture of the methods of transport used. Also of interest might be the *relative* frequency of each of the modes of transport. The relative frequency is simply the frequency expressed as a proportion of the total number of students surveyed. If this is given as a percentage, as here, this is known as the *percentage relative frequency*.

Mode	Frequency	Relative Frequency (%)
Car	10	33.3
Walk	7	23.4
Bike	4	13.3
Bus	4	13.3
Metro	4	13.3
Train	1	3.4
Total	30	100

3.3.2 Frequency tables for count data

Example 3.1

The following table shows the raw data for car sales at a new car showroom over a two week period in July 2012.

Date	Cars Sold	Date	Cars Sold
01/07/12	9	08/07/12	10
02/07/12	8	09/07/12	5
03/07/12	6	10/07/12	8
04/07/12	7	11/07/12	4
05/07/12	7	12/07/12	6
06/07/12	10	13/07/12	8
07/07/12	11	14/07/12	9

Present these data in a relative frequency table by number of days on which different numbers of cars were sold.

Cars Sold	Tally	Frequency	Relative Frequency %
Totals			

3.3.3 Frequency tables for continuous data

With discrete data, and especially with small data sets, it is easy to count the quantities in the defined categories. With continuous data this is not possible. Strictly speaking, no two observations are precisely the same. With such observations we group the data together. For example, the following data set represents the service time in seconds for callers to a credit card call centre.

214.8412	220.6484	216.7294	195.1217	211.4795
195.8980	201.1724	185.8529	183.4600	178.8625
196.3321	199.7596	206.7053	203.8093	203.1321
200.8080	201.3215	205.6930	181.6718	201.7461
180.2062	193.3125	188.2127	199.9597	204.7813
198.3838	193.1742	204.0352	197.2206	193.5201
205.5048	217.5945	208.8684	197.7658	212.3491
209.9000	197.6215	204.9101	203.1654	192.9706
208.9901	202.0090	195.0241	192.7098	219.8277
208.8920	200.7965	191.9784	188.8587	206.8912

To produce a continuous data frequency table we first need to divide the range of the variable into smaller ranges called *class intervals*. The class intervals should, between them, cover every possible value. There should be no gaps between the intervals. One way to ensure this is to include the boundary value as the smallest value in the next class above. This can be written as, for example, $20 \leq \text{obs} < 30$. This means we include all observations (represented by "obs") within this class interval that have a value of at least 20 up to values just below 30.

Some things to think about:

- Often for simplicity we would write the class intervals up to the number of decimal places in the data and avoid using the inequalities; for example, 20 up to 29.999 if we were working to 3 decimal places.
- We need to include the full range of data in our table and so we need to identify the minimum and maximum points (sometimes our last class might be "greater than such and such").
- The class interval width should be a convenient number – for example 5, 10, or 100, depending on the data. Obviously we do not want so many classes that each one has only one or two observations in it.
- The appropriate number of classes will vary from data set to data set; however, with simple examples that you would work through by hand, it is unlikely that you would have more than ten to fifteen classes.

Example 3.2

Create a frequency table for the call centre data. Also, find the relative frequencies in each class interval.

Class Interval	Tally	Frequency	Relative Frequency %
Totals			

3.4 Graphical methods for presenting data

We have looked at ways of collecting data and then collating them into tables. Frequency tables are useful methods of presenting data; they do, however, have their limitations. With large amounts of data graphical presentation methods are often clearer to understand. Here, we look at methods for producing graphical representations of data of the types we have seen previously.

3.4.1 Stem and Leaf plots

Stem and leaf plots are a quick and easy way of representing data graphically. They can be used with both discrete and continuous data. You should refer to Section 2.3.1 of the summer revision booklet for more details about these plots.

Example: Percentage returns on a share

The following numbers show the percentage returns on an ordinary share for 23 consecutive months:

0.2	–2.1	1.0	0.1	–0.5	2.4	–2.3	1.5	1.2	–0.6	2.4	–1.2
1.7	–1.3	–1.2	0.9	0.5	0.1	–0.1	0.3	–0.4	0.5	0.9	

Here, the largest value is 2.4 and the smallest –2.3, and we have lots of decimal values in between. Thus, it seems sensible here to have a stem unit of 1 and a leaf unit of 0.1. A stem and leaf diagram for this set of returns then might look like:

Stem	Leaf
–2	3 1
–1	3 2 2
–0	6 5 4 1
0	1 1 2 3 5 5 9 9
1	0 2 5 7
2	4 4

$n = 23$, stem unit $= 1$, leaf unit $= 0.1$.

Example: Production line data

Consider the following data on lengths of items on a production line (in cm):

2.97	3.81	2.54	2.01	3.49	3.09	1.99	2.64	2.31	2.22

The stem and leaf plot for this is shown overleaf. Notice that all figures have been rounded down, or *cut*, to one decimal place.

1	9
2	0 2 3 5 6 9
3	0 4 8

$n = 10$, stem unit = 1 cm, leaf unit = 0.1 cm.

Why do you think we *cut* the extra digits?

Example 3.3

The observations in the table below are the recorded time it takes to get through to an operator at a telephone call centre (in seconds). Construct a stem–and–leaf plot for these data, and comment.

54	56	50	67	55	38	49	45	39	50
45	51	47	53	29	42	44	61	51	50
30	39	65	54	44	54	72	65	58	62

Stem **Leaf**

$n =$ stem unit = leaf unit =

Example 3.4

The stem and leaf plot below represents the marks on a test for 52 students. Comment on the distribution of these marks.

1	4										
1	5	7	7								
2	1	1	2	3							
2	5	5	6	7	8	8					
3	2	3	3	3	4	4	4				
3	5	5	6	7	7	8	8	9	9	9	9
4	0	0	1	1	1	2	2	4	4	4	
4	5	7	7	8	8	8	9				
5	0	0	0								

$n = 52$, stem unit $= 10$, leaf unit $= 1$.

3.4.2 Bar Charts

Bar charts are a commonly–used and clear way of presenting categorical data or any ungrouped discrete frequency observations. See Section 2.3.2 of the summer revision booklet for more details.

3.4.3 Pie charts

Pie charts are simple diagrams for displaying categorical or grouped data. These charts are commonly used within industry to communicate simple ideas, for example market share. They are used to show the proportions of a whole. They are best used when there are only a handful of categories to display. See Section 2.3.3 of the summer revision booklet for more details.

3.4.4 Histograms

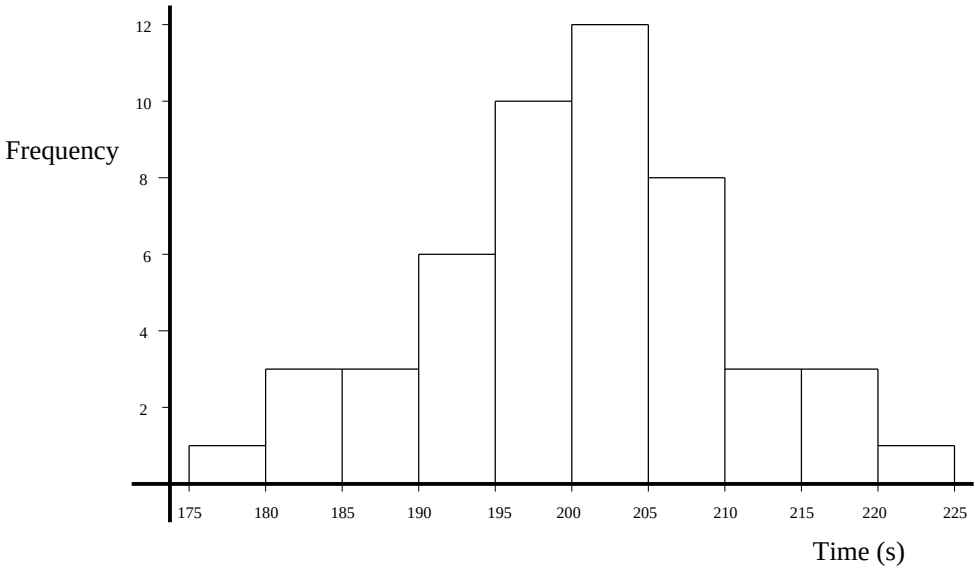
Bar charts have their limitations; for example, they cannot be used to present continuous data. When dealing with continuous random variables a different kind of graph is required. This is called a *histogram*. At first sight these look similar to bar charts. There are, however, two critical differences:

- the horizontal (x-axis) is a continuous scale. As a result of this there are *no gaps between the bars* (unless there are no observations within a class interval);
- the height of the rectangle is only proportional to the frequency if the class intervals are all equal. With histograms it is the *area* of the rectangle that is proportional to their frequency.

The frequency table for the data on service times for a telephone call centre (Section 3.3.3) was

Service time	**Frequency**
$175 \leq$ time < 180	1
$180 \leq$ time < 185	3
$185 \leq$ time < 190	3
$190 \leq$ time < 195	6
$195 \leq$ time < 200	10
$200 \leq$ time < 205	12
$205 \leq$ time < 210	8
$210 \leq$ time < 215	3
$215 \leq$ time < 220	3
$220 \leq$ time < 225	1
Total	50

Notice that all the class intervals are the same width, and so the histogram for these data is:

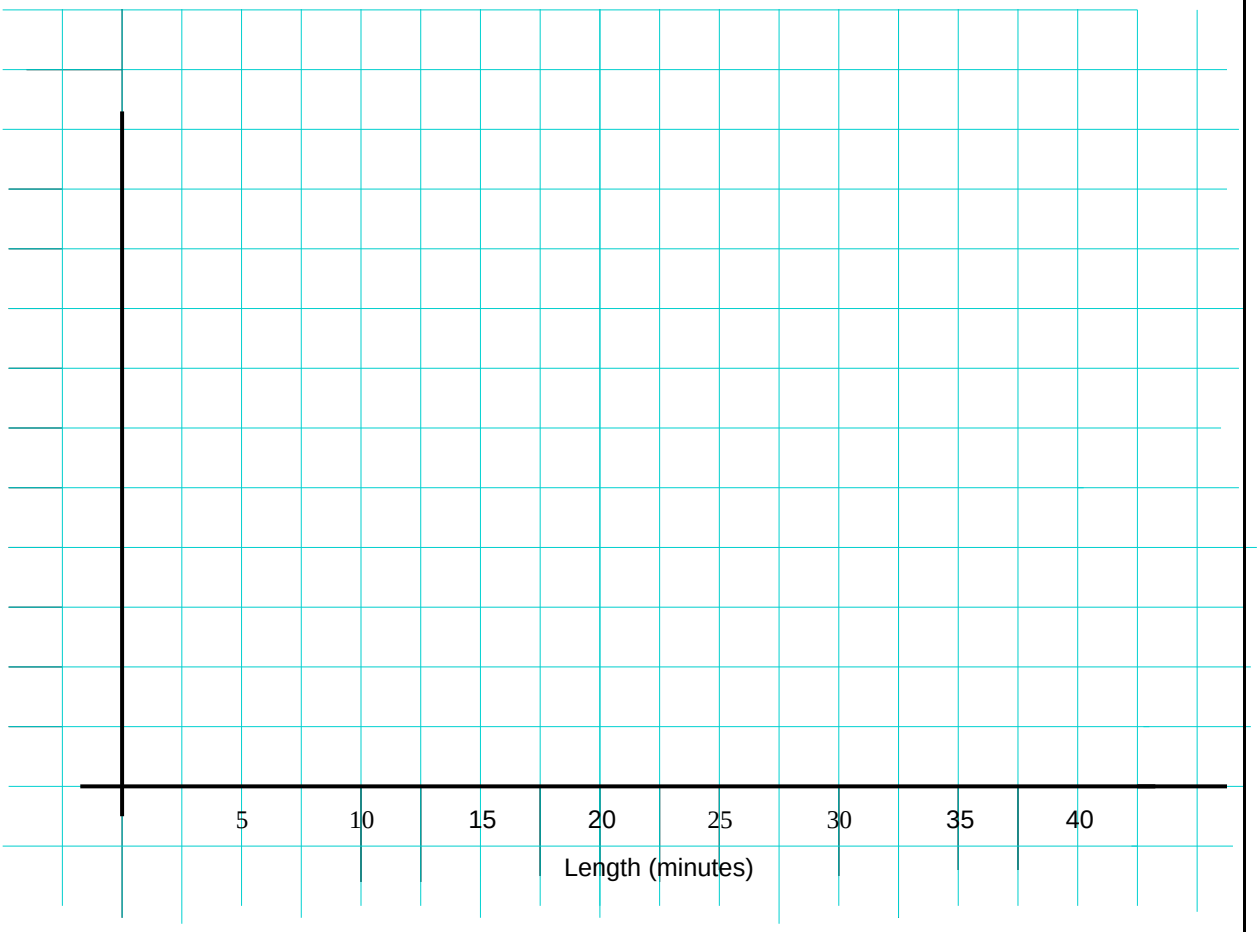


Example 3.5

The *Holiday Hypermarket* travel agency received 64 telephone calls yesterday morning. The table below gives information of the lengths, in minutes, of these telephone calls.

Length (x) minutes	**Frequency**	**Frequency density**
$0 \leq x < 5$	4	0.8
$5 \leq x < 15$	10	1.0
$15 \leq x < 30$	24	
$30 \leq x < 40$	20	
$40 \leq x < 45$	6	

Complete this table, and construct a histogram for these data. What is the *modal class* here?

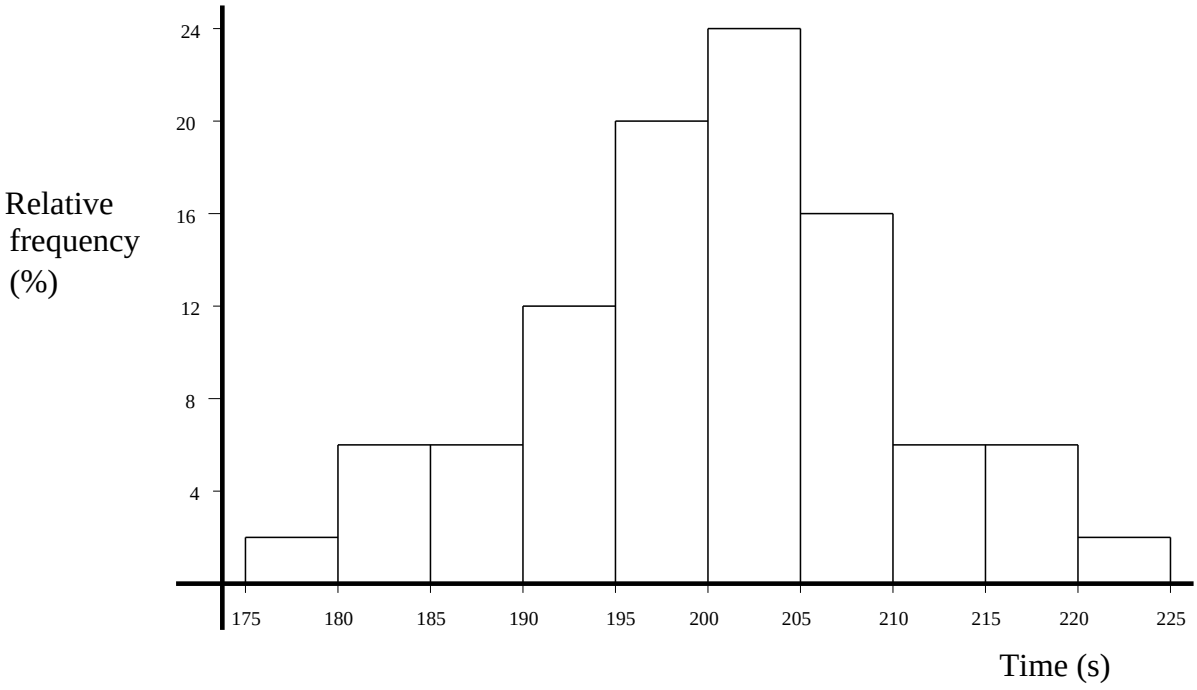


3.4.5 Percentage Relative Frequency Histograms

When we produced frequency tables in Section 3.3, we included a column for *percentage relative frequency*. This contained values for the frequency of each group, relative to the overall sample size, expressed as a percentage. For example, a percentage relative frequency table for the data on service time (in seconds) for calls to a credit card service centre is:

Service time	**Frequency**	**Relative Frequency (%)**
$175 \leq \text{time} < 180$	1	2
$180 \leq \text{time} < 185$	3	6
$185 \leq \text{time} < 190$	3	6
$190 \leq \text{time} < 195$	6	12
$195 \leq \text{time} < 200$	10	20
$200 \leq \text{time} < 205$	12	24
$205 \leq \text{time} < 210$	8	16
$210 \leq \text{time} < 215$	3	6
$215 \leq \text{time} < 220$	3	6
$220 \leq \text{time} < 225$	1	2
Totals	50	100

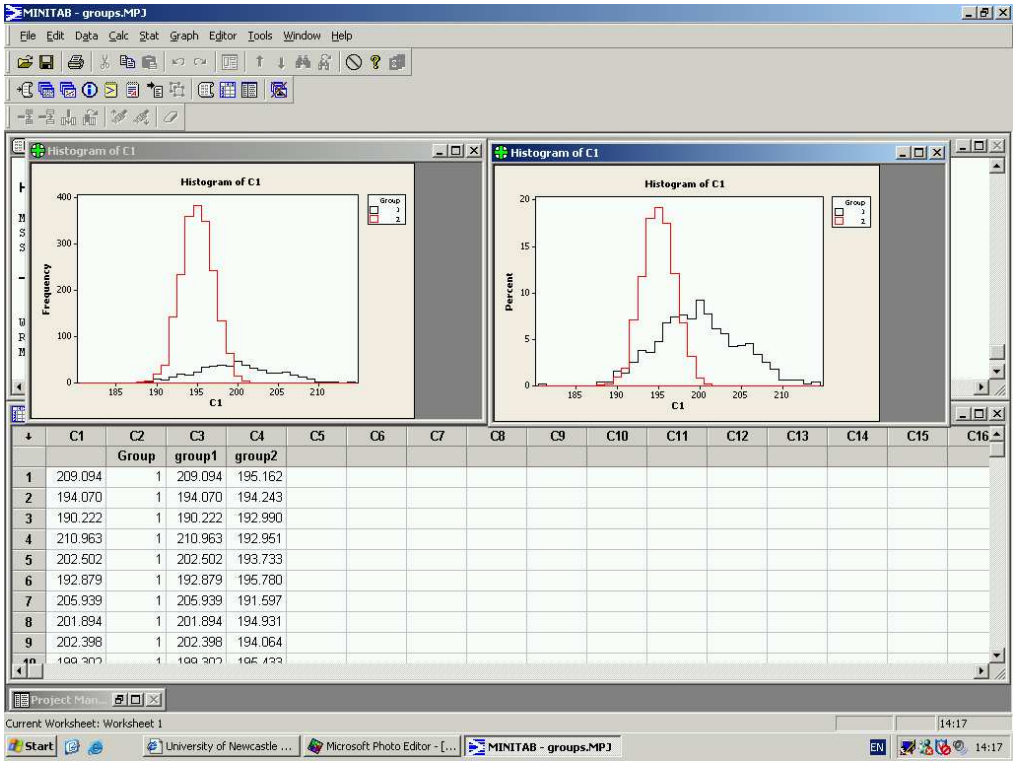
You can plot these data like an ordinary histogram, or, instead of using *frequency/frequency density* on the vertical axis (y-axis), you could use the *percentage relative frequency/percentage relative frequency density*.



Note that the y-axis now contains the relative percentages rather than the frequencies. You might well ask "why would we want to do this?".

These percentage relative frequency histograms are useful when comparing two samples that have different numbers of observations. If one sample were larger than the other then a frequency histogram would show a difference simply because of the larger number of observations. Looking at percentages removes this difference and enables us to look at *relative* differences.

For example, in the following graph (produced in the computer package `Minitab` – see Semester 2) there are data from two groups and four times as many data points for one group as the other. The left–hand plot shows an ordinary histogram and it is clear that the comparison between groups is masked by the quite different sample sizes. The right–hand plot shows a histogram based on (percentage) relative frequencies and this enables a much more direct comparison of the distributions in the two groups.



	C1	C2	C3	C4
		Group	group1	group2
1	209.094	1	209.094	195.162
2	194.070	1	194.070	194.243
3	190.222	1	190.222	192.990
4	210.963	1	210.963	192.951
5	202.502	1	202.502	193.733
6	192.879	1	192.879	195.780
7	205.939	1	205.939	191.597
8	201.894	1	201.894	194.931
9	202.398	1	202.398	194.064

Overlaying histograms on the same graph can sometimes not produce such a clear picture, particularly if the values in both groups are close or overlap one another significantly.

3.4.6 Relative Frequency Polygons

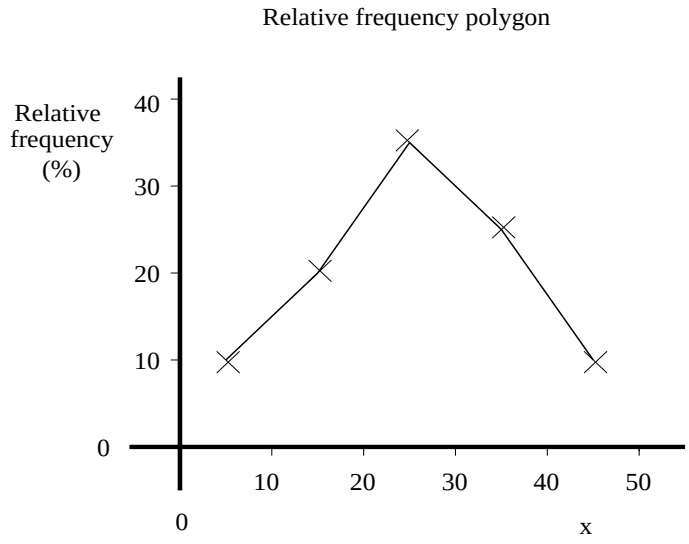
These are a natural extension of the relative frequency histogram. They differ in that, rather than drawing bars, each class is represented by one point and these are joined together by straight lines. The method is similar to that for producing a histogram:

1. Produce a percentage relative frequency table.
2. Draw the axes
   - The x-axis needs to contain the full range of the classes used.
   - The y-axis needs to range from 0 to the maximum percentage relative frequency.
3. Plot points: pick the mid point of the class interval on the x-axis and go up until you reach the appropriate percentage value on the y-axis and mark the point. Do this for each class.
4. Join adjacent points together with straight lines.

The relative frequency polygon is exactly the same as the relative frequency histogram, but instead of having bars we join the mid–points of the top of each bar with a straight line. Consider the following simple example.

Class Interval	Mid Point	% Relative Frequency
$0 \leq x < 10$	5	10
$10 \leq x < 20$	15	20
$20 \leq x < 30$	25	35
$30 \leq x < 40$	35	25
$40 \leq x < 50$	45	10

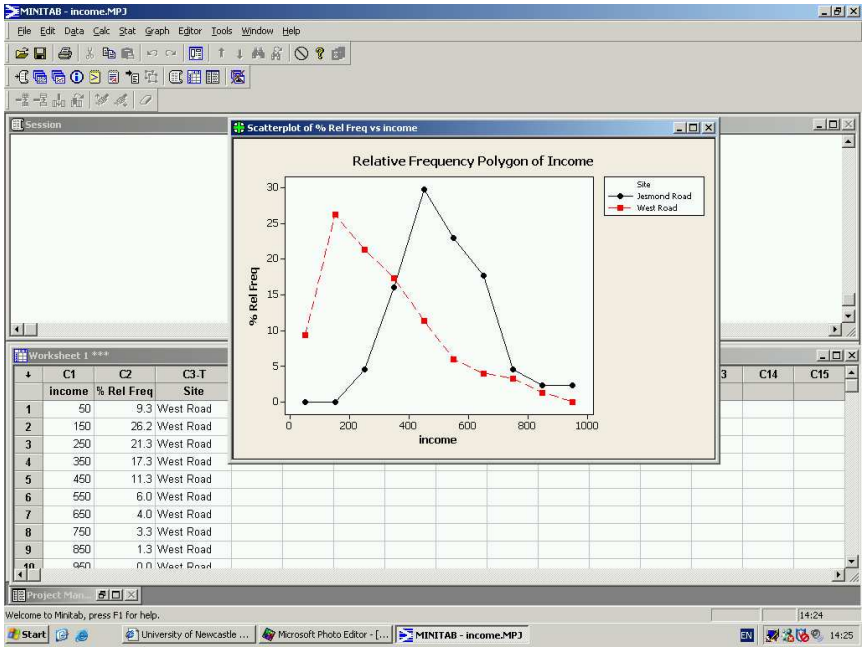
We can draw this easily by hand:



These percentage relative frequency polygons are very useful for comparing two or more samples – we can easily "overlay" many relative frequency polygons, but overlaying the corresponding histograms could get really messy! Consider the following data on gross weekly income (in £) collected from two sites in Newcastle. Let us suppose that many more responses were collected in Jesmond so that a direct comparison of the frequencies using a standard histogram is not appropriate. Instead we use *relative* frequencies.

Weekly Income (£)	**West Road** (%)	**Jesmond Road** (%)
$0 \leq$ income < 100	9.3	0.0
$100 \leq$ income < 200	26.2	0.0
$200 \leq$ income < 300	21.3	4.5
$300 \leq$ income < 400	17.3	16.0
$400 \leq$ income < 500	11.3	29.7
$500 \leq$ income < 600	6.0	22.9
$600 \leq$ income < 700	4.0	17.7
$700 \leq$ income < 800	3.3	4.6
$800 \leq$ income < 900	1.3	2.3
$900 \leq$ income < 1000	0.0	2.3

The computer package `Minitab` (see Semester 2) was used to produce the following plot of the percentage relative frequency polygons for the two groups.



We can clearly see the differences between the two samples. The line connecting the boxes represents the data from West Road and the line connecting the circles represents those for Jesmond Road. The distribution of incomes on West Road is skewed towards lower values, whilst those on Jesmond Road are more symmetric. The graph clearly shows that income in the Jesmond Road area is higher than that in the West Road area.

3.4.7 Cumulative Frequency Polygons (Ogives)

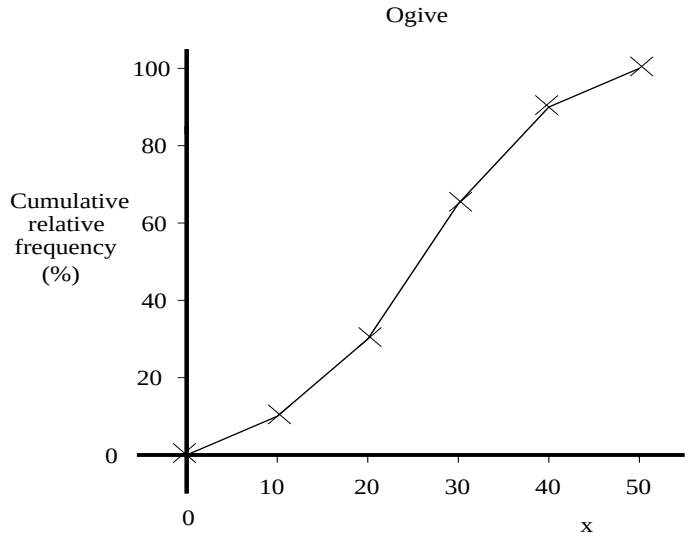
Cumulative percentage relative frequency is also a useful tool. The cumulative percentage relative frequency is simply the sum of the percentage relative frequencies at the end of each class interval (i.e. we add the frequencies up as we go along). Consider the example from the previous section:

Class Interval	% Relative Frequency	Cumulative % Relative Frequency
$0 \leq x < 10$	10	10
$10 \leq x < 20$	20	30
$20 \leq x < 30$	35	65
$30 \leq x < 40$	25	90
$40 \leq x < 50$	10	100

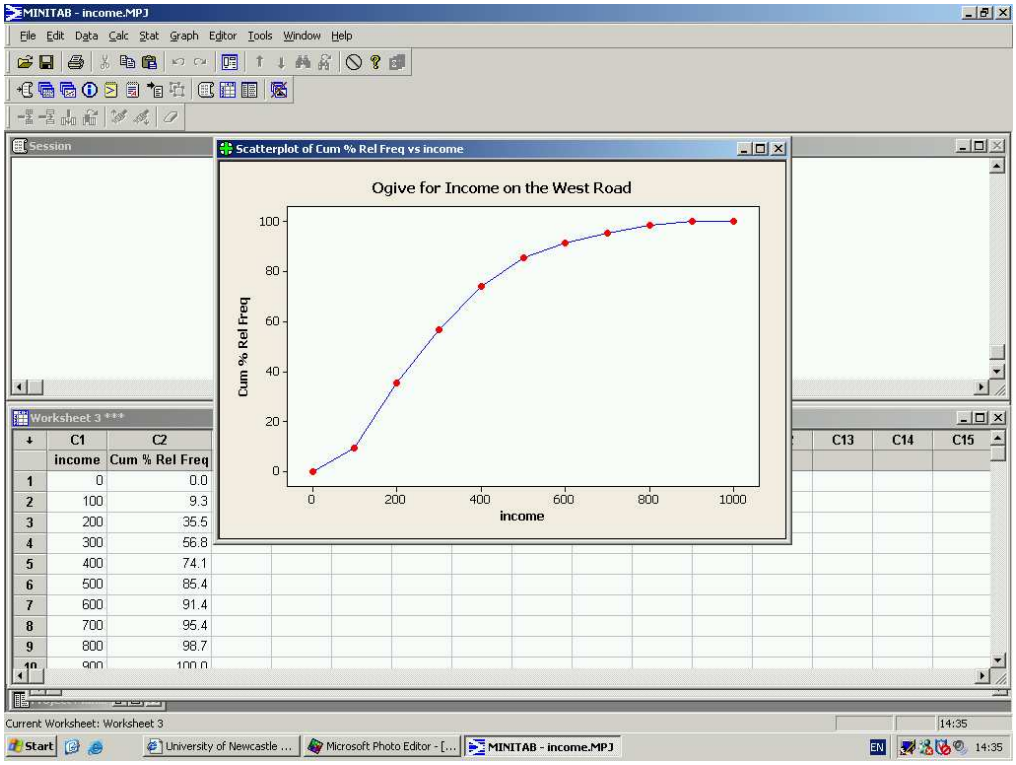
At the upper limit of the first class the cumulative % relative frequency is simply the % relative frequency in the first class, i.e. 10. However, at the end of the second class, at 20, the cumulative % relative frequency is $10 + 20 = 30$. The cumulative % relative frequency at the end of the last class must be 100.

The corresponding graph, or *ogive*, is simple to produce by hand:

1. Draw the axes.
2. Label the x-axis with the full range of the data and the y-axis from 0 to 100%.
3. Plot the cumulative % relative frequency at the *end point* of each class.
4. Join adjacent points, starting at 0% at the lowest class boundary.



For example, `Minitab` was used to produce the ogive below for the income data from the West Road survey:

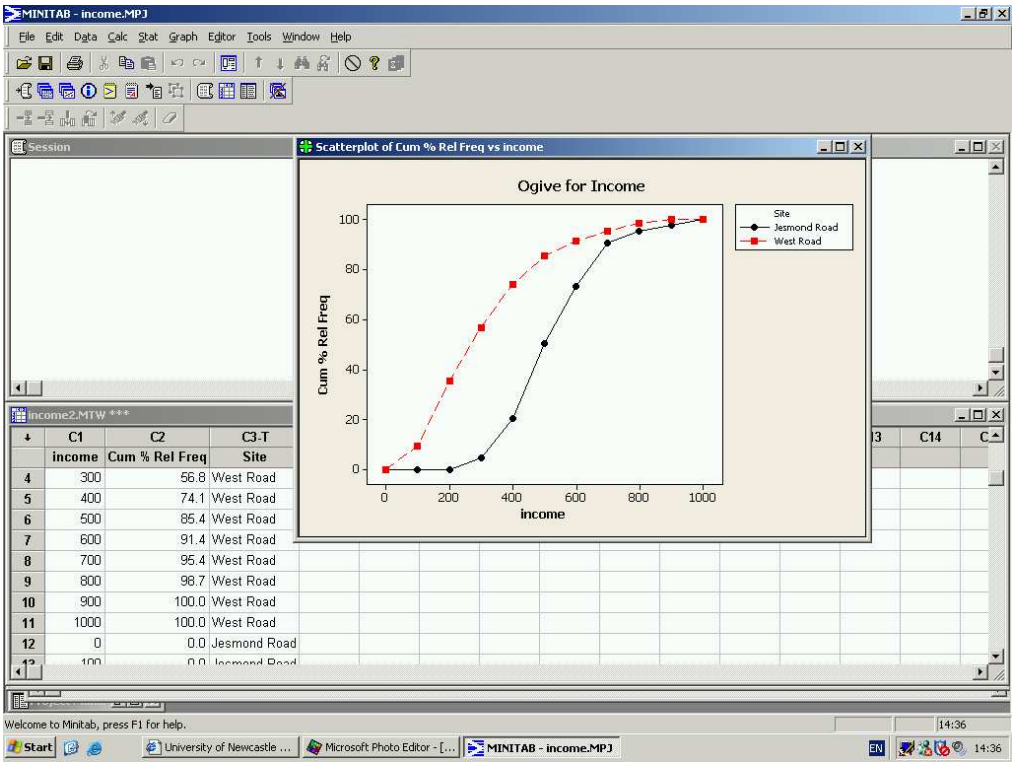


This graph instantly tells you many things. To see what percentage of respondents earn less than £x per week:

1. Find x on the x-axis and draw a line up from this value until you reach the ogive;
2. From this point trace across to the y-axis;
3. Read the percentage from the y-axis.

If we wanted to know what percentage of respondents in the survey in West Road earn less than £250 per week, we simply find £250 on the x-axis, trace up to the ogive and then trace across to the y-axis and we can read a figure of about 47%. The process obviously works in reverse. If we wanted to know what level of income 50% of respondents earned, we would trace across from 50% to the ogive and then down to the x-axis and read a value of about £300.

Ogives can also be used for comparison purposes. The following plot contains the ogives for the income data at both the West Road and Jesmond Road sites.



	C1	C2	C3-T
	income	Cum % Rel Freq	Site
4	300	56.8	West Road
5	400	74.1	West Road
6	500	85.4	West Road
7	600	91.4	West Road
8	700	95.4	West Road
9	800	98.7	West Road
10	900	100.0	West Road
11	1000	100.0	West Road
12	0	0.0	Jesmond Road

It clearly shows the ogive for Jesmond Road is shifted to the right of that for West Road. This tells us that the surveyed incomes are higher on Jesmond Road. We can compare the percentages of people earning different income levels between the two sites quickly and easily.

This technique can also be used to great effect for examining the changes before and after the introduction of a marketing strategy. For example, daily sales figures of a product for a period before and after an advertising campaign might be plotted. Here, a comparison of the two ogives can be used to help assess whether or not the campaign has been successful.

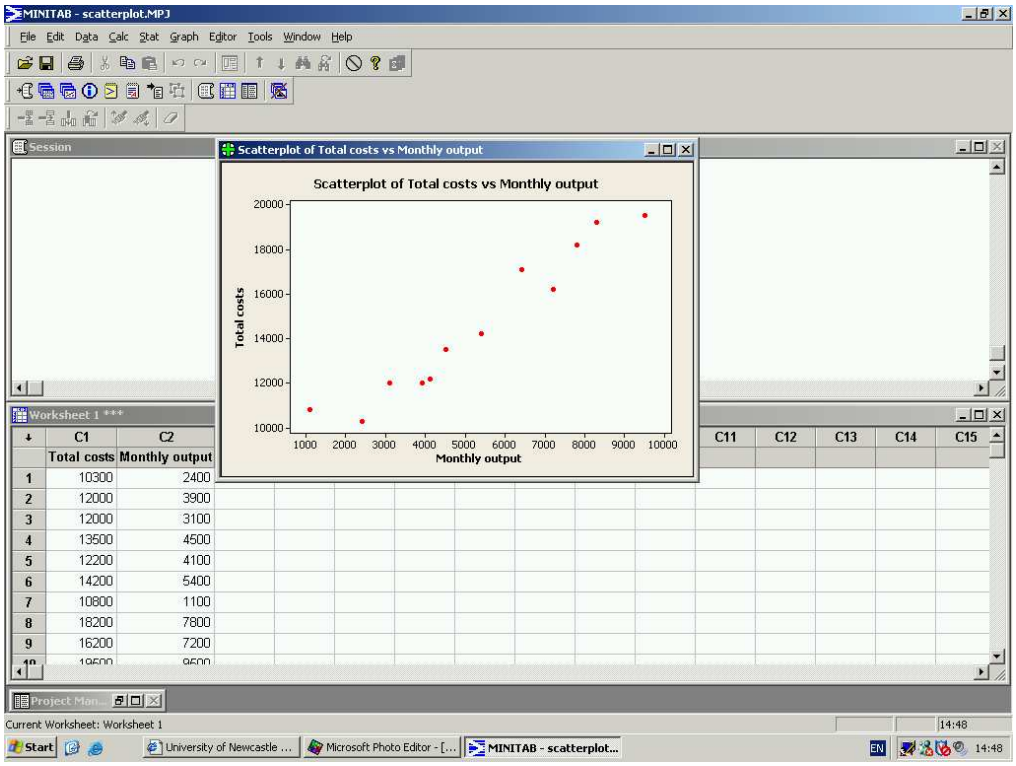
3.4.8 Scatter Plots

Scatter plots are used to plot two variables which you believe might be related, for example, height and weight, advertising expenditure and sales, or age of machinery and maintenance costs.

Consider the following data for monthly output (y thousand units) and total costs (x thousand pounds) at a factory.

x	10.3	12	12	13.5	12.2	14.2	10.8	18.2	16.2	19.5	17.1	19.2
y	2.4	3.9	3.1	4.5	4.1	5.4	1.1	7.8	7.2	9.5	6.4	8.3

If you were interested in the relationship between the cost of production and the number of units produced you could easily plot this by hand. Here, we have used `Minitab` to produce the scatterplot (see Semester 2).



The plot highlights a clear relationship between the two variables: the more units made, the more it costs in total. This relationship is shown on the graph by the upwards trend within the data – as monthly output increases so too does total cost. A downwards sloping trend would imply that as output increased, total costs declined, an unlikely scenario. This type of plot is the first stage of a more sophisticated analysis which we will develop in Semester 2 of this course.

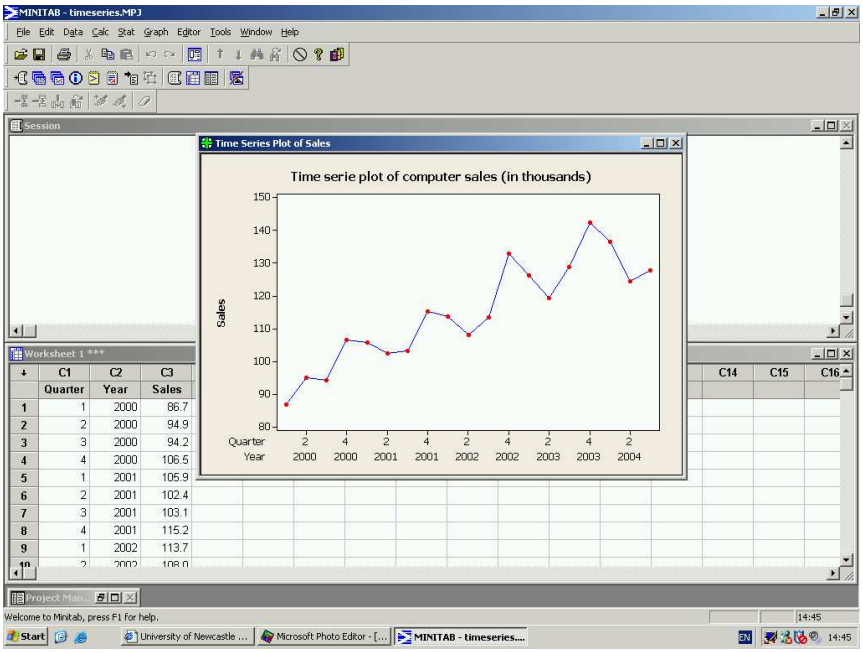
3.4.9 Time Series Plots

So far we have only considered data where we can (at least for some purposes) ignore the order in which the data come. Not all data are like this. One exception is the case of time series data, that is, data collected over time. Examples include monthly sales of a product, the price of a share at the end of each day or the air temperature at midday each day. Such data can be plotted by using a scatter plot, but with *time* as the (horizontal) x-axis, and where the points are connected by lines.

Consider the following data on the number of computers sold (in thousands) by quarter (January-March, April-June, July-September, October-December) at a large warehouse outlet.

Year (Quarter)	2000 (Q1)	2000 (Q2)	2000 (Q3)	2000 (Q4)	2001 (Q1)	2001 (Q2)	2001 (Q3)	2001 (Q4)
Units sold	86.7	94.9	94.2	106.5	105.9	102.4	103.1	115.2
Year (Quarter)	2002 (Q1)	2002 (Q2)	2002 (Q3)	2002 (Q4)	2003 (Q1)	2003 (Q2)	2003 (Q3)	2003 (Q4)
Units sold	113.7	108.0	113.5	132.9	126.3	119.4	128.9	142.3
Year (Quarter)	2004 (Q1)	2004 (Q2)	2004 (Q3)					
Units sold	136.4	124.6	127.9					

The time series plot, as produced in `Minitab`, is shown below (see Semester 2 for more details):



The plot clearly shows us two things: firstly, that there is an upwards trend to the data, and secondly that there is some regular variation around this trend. We will come back to more sophisticated techniques for analysing time series data later in the course.

3.5 Numerical summaries of data

So far we have only considered graphical methods for presenting data. These are always useful starting points. As we shall see, however, for many purposes we might also require *numerical* methods for summarising data. Before we introduce some ways of summarising data numerically, let us first think about some notation.

3.5.1 Mathematical notation

Before we can talk more about numerical techniques we first need to define some basic notation. This will allow us to generalise all situations with a simple shorthand.

Very often in statistics we replace actual numbers with letters in order to be able to write general formulae. We generally use a single upper case letter to represent our random variable and the lower case to represent sample data, with subscripts to distinguish individual observations in the sample. Amongst the most common letters to use is x, although y and z are frequently used as well. For example, suppose we ask a random sample of three people how many mobile phone calls they made yesterday. We might get the following data: 1, 5, 7. If we take another sample we will most likely get different data, say 2, 0, 3. Using algebra we can represent the general case as x_1, x_2, x_3:

1st sample	1	5	7
2nd sample	2	0	3
typical sample	x_1	x_2	x_3

This can be generalised further by referring to the random variable *as a whole* as X and the ith observation in the sample as x_i. Hence, in the first sample above, the second observation is $x_2 = 5$ whilst in the second sample it is $x_2 = 0$. The letters i and j are most commonly used as the index numbers for the subscripts.

The total number of observations in a sample is usually referred to by the letter n. Hence in our simple example above $n = 3$.

The next important piece of notation to introduce is the symbol $\sum$. This is the upper case of the Greek letter *sigma*, pronounced "sigma". It is used to represent the phrase "sum the values". This symbol is used as follows:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

This notation is used to represent the sum of all the values in our data (from the first $i = 1$ to the last $i = n$), and is often abbreviated to $\sum x$ when we sum over all the data in our sample.

3.5.2 Measures of Location

These are also referred to as measures of *centrality* or, more commonly, *averages*. In general terms, they tell us the value of a "typical" observation. There are three measures which are commonly used: the *mean*, the *median*, and the *mode*. We will consider these in turn.

The Arithmetic Mean

The arithmetic mean is perhaps the most commonly used measure of location. We often refer to it as the average or just the mean. The arithmetic mean is calculated by simply adding all our data together and dividing by the number of data we have. So if our data were 10, 12, and 14, then our mean would be

$$\frac{10+12+14}{3} = \frac{36}{3} = 12.$$

We denote the mean of our sample, or sample mean, using the notation $\bar{x}$ ("x bar"). In general, the mean is calculated using the formula

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

or equivalently as

$$\bar{x} = \frac{\sum x}{n}.$$

For small data sets this is easy to calculate by hand, though this is simplified by using the statistics mode on a University approved calculator. This will be shown to you in the workshops.

Sometimes we might not have the raw data; instead, the data might be available in the form of a table. It is still possible to calculate the mean from such data. Let us first consider the case where we have some ungrouped discrete data. Previously we have seen the data:

Date	Cars Sold	Date	Cars Sold
01/07/12	9	08/07/12	10
02/07/12	8	09/07/12	5
03/07/12	6	10/07/12	8
04/07/12	7	11/07/12	4
05/07/12	7	12/07/12	6
06/07/12	10	13/07/12	8
07/07/12	11	14/07/12	9

The mean number of cars sold per day is

$$\bar{x} = \frac{9+8+\ldots+8+9}{14} = 7.71.$$

These data can be presented as the frequency table:

Cars Sold $(x_{(j)})$	**Frequency** (f_j)
4	1
5	1
6	2
7	2
8	3
9	2
10	2
11	1
Total (n)	14

The sample mean can be calculated from these data as

$$\begin{aligned}\bar{x} &= \frac{4+5+\overbrace{6+6}^{\times 2}+\overbrace{7+7}^{\times 2}+\overbrace{8+8+8}^{\times 3}+\overbrace{9+9}^{\times 2}+\overbrace{10+10}^{\times 2}+11}{14} \\ &= \frac{(\mathbf{4}\times 1)+(\mathbf{5}\times 1)+(\mathbf{6}\times 2)+(\mathbf{7}\times 2)+(\mathbf{8}\times 3)+(\mathbf{9}\times 2)+(\mathbf{10}\times 2)+(\mathbf{11}\times 1)}{14} \\ &= 7.71.\end{aligned}$$

We can express this calculation of the sample mean from discrete tabulated data as

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_{(j)} f_j.$$

Here the different values of X which occur in the data are $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. In the example $x_{(1)} = 4,\ x_{(2)} = 5, \ldots, x_{(k)} = 11$ and $k = 8$.

If we only have grouped frequency data, it is still possible to *approximate* the value of the sample mean. Consider the following (ordered) data:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

The sample mean of these data is 9.73. Grouping these data into a frequency table gives:

Class Interval	**mid–point** (m_j)	**Frequency** (f_j)
$8.0 \le x < 8.5$	8.25	1
$8.5 \le x < 9.0$	8.75	1
$9.0 \le x < 9.5$	9.25	5
$9.5 \le x < 10.0$	9.75	7
$10.0 \le x < 10.5$	10.25	2
$10.5 \le x < 11.0$	10.75	3
$11.0 \le x < 11.5$	11.25	1
Total (n)		20

When the raw data are not available, we don't know where each observation lies in each interval. The best we can do is to assume that all the values in each interval lie at the central value of the interval, that is, at its mid–point. Therefore, the (approximate) sample mean is calculated using the the frequencies (f_j) and the mid–points (m_j) as

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} f_j m_j.$$

For the grouped data above, we obtain

$$\bar{x} = \frac{1}{20}\left(1 \times 8.25 + 1 \times 8.75 + \cdots + 3 \times 10.75 + 1 \times 11.25\right) = 9.775.$$

This value is fairly close to the correct sample mean and is a reasonable approximation given the partial information we have in the table.

For large samples with narrow intervals, this approximate value will be very close to the correct sample mean (calculated using the raw data).

The Median

The median is occasionally used instead of the mean, particularly when the data have an asymmetric profile (as indicated by a histogram or stem–and–leaf plot – see earlier). The median is the middle value of the observations when they are listed in ascending order. It is straightforward to determine the median for small data sets, particularly via a stem and leaf plot. For larger data sets, the calculation is more easily done using `Minitab` (see Semester 2).

The median is that value that has half the observations above it and half below. If the sample size (n) is an odd number, we have:

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \text{ largest observation.}$$

For example, if our data were 2, 3, 3, 5, 6, 7, 9, then the sample size ($n = 7$) is an odd number and therefore the median is the

$$\frac{7+1}{2} = 4^{th} \text{ largest observation,}$$

that is, the median is the fourth largest (or smallest) ranked observation: for these data the median = 5.

If the sample size (n) is an even number the process is slightly more complicated:

$$\text{median} = \text{average of the } \left(\frac{n}{2}\right)^{th} \text{ and the } \left(\frac{n}{2}+1\right)^{th} \text{ largest observations.}$$

For example, if our data were 2, 3, 3, 5, 6, 7, 9, 10 then the sample size ($n = 8$) is an

even number and therefore

$$\begin{aligned}\text{median} &= \text{average of the } \left(\frac{8}{2}\right)^{th} \text{ and the } \left(\frac{8}{2}+1\right)^{th} \text{ largest observations} \\ &= \frac{5+6}{2} \\ &= 5.5.\end{aligned}$$

It is possible to estimate the median value from an ogive as it is half way through the ordered data and hence is at the 50% level of the cumulative frequency. The accuracy of this estimate will depend on the accuracy of the ogive drawn.

The Mode

This is the final measure of location we will look at. It is the value of the random variable in the sample which occurs with the highest frequency. It is usually found by inspection. For discrete data this is easy. The mode is simply the most common value. So, on a bar chart, it would be the category with the highest bar. For example, consider the following data: 2, 2, 2, 3, 3, 4, 5. Quite obviously the mode is 2 as it occurs most often. We often talk about modes in terms of categorical data. For example, in a survey of 12 students, 4 said they read the *Metro* newspaper, 5 said they read *The Sun* and 3 said they read *The Times*. Thus, the mode is *The Sun*, as it is the most popular newspaper. It is possible to refer to modal classes with grouped data. This is simply the class with the greatest frequency of observations. For example, the model class of

Class	**Frequency**
$10 \leq x < 20$	10
$20 \leq x < 30$	15
$30 \leq x < 40$	30

is obviously $30 \leq x < 40$. It is not possible to put a single value on the mode with such continuous data. However, the modal class might tell you much about the data. Modal classes are also obvious from histograms, being the highest peaked bar. Of course, if we change the class boundaries, the position of the modal class may change.

3.5.3 Measures of Spread

A measure of location is insufficient in itself to summarise data as it only describes the value of a typical outcome and not how much variation there is in the data. For example, the two datasets 6, 22, 38 and 21, 22, 23 both have the same mean (22) and the same median (22). However the first set of data ranges considerably from this value while the second stays very close. They are quite clearly very different data sets. The mean or the median does not fully represent the data. There are three basic measures of spread which we will consider: the *range*, the *inter–quartile range* and the *sample variance*.

The Range

This is the simplest measure of spread. It is simply the difference between the largest and smallest observations. In our simple example above the range for the first set of numbers is $38 - 6 = 32$ and for the second set it is $23 - 21 = 2$. These clearly describe very different data sets. The first set has a much wider range than the second.

There are two problems with the range as a measure of spread. When calculating the range you are looking at the two most extreme points in the data, and hence the value of the range can be unduly influenced by one particularly large or small value, known as an *outlier*. The second problem is that the range is only really suitable for comparing (roughly) equally sized samples as it is more likely that large samples contain the extreme values of a population.

The Inter–Quartile Range

The inter–quartile range describes the range of the middle half of the data and so is less prone to the influence of the extreme values.

To calculate the inter–quartile range (IQR) we simply divide the the ordered data into four quarters. The three values that split the data into these quarters are called the *quartiles*. The first quartile (*lower quartile*, $Q1$) has 25% of the data below it; the second quartile (*median*, $Q2$) has 50% of the data below it; and the third quartile (*upper quartile*, $Q3$) has 75% of the data below it. We already know how to find the median, and the other quartiles are calculated as follows:

$$Q1 = \frac{(n+1)}{4}\text{th smallest observation}$$
$$Q3 = \frac{3(n+1)}{4}\text{th smallest observation.}$$

Just as with the median, these quartiles might not correspond to actual observations. For example, in a dataset with $n = 20$ values, the lower quartile is the $5\frac{1}{4}$th largest observation, that is, a quarter of the way between the 5th and 6th largest observations. This calculation is essentially the same process we used when calculating the median. Consider again the data:

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Here the 5th and 6th smallest observations are 9.2 and 9.3 respectively. Therefore, the lower quartile is $Q1 = 9.225$. Similarly the upper quartile is the $15\frac{3}{4}$ smallest observation, that is, three quarters of the way between 10.3 and 10.4; so $Q3 = 10.375$.

The inter–quartile range is simply the difference between the upper and lower quartiles, that is

$$IQR = Q3 - Q1$$

which for these data is $IQR = 10.375 - 9.225 = 1.15$.

The interquartile range can also be *estimated* from the ogives in a similar manner to the median. Simply draw the ogive and then read off the values for 75% and 25% and calculate the difference between them. This is especially useful if you only have grouped data. Again the accuracy depends on the quality of your graph.

The inter–quartile range is useful as it allows us to make comparisons between the ranges of two data sets, without the problems caused by outliers or uneven sample sizes.

The Sample Variance and Standard Deviation

The *sample variance* is the standard measure of spread used in statistics. It is usually denoted by s^2 and is simply the "average" of the squared distances of the observations from the sample mean. Strictly speaking, the sample variance measures deviation about a value calculated from the data (the sample mean) and so we use an $n-1$ divisor rather than n. That is, we use the formula

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}.$$

We can rewrite this using more condensed mathematical notation and simplify this to

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

or equivalently as

$$s^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \right\}.$$

Note that the notation x_i^2 represents the squared value of the observation x_i. That is, $x_i^2 = (x_i)^2$.

The *sample standard deviation* s is the positive square root of the sample variance. This quantity is often used in preference to the sample variance as it has the same units as the original data and so is perhaps easier to understand.

If this appears complicated, don't worry, as most basic calculators will give the sample standard deviation when in Statistics mode. Note that the correct standard deviation is given by the s or σ_{n-1} button on the calculator and **not** the σ or σ_n buttons.

A different calculation is needed when the data are given in the form of a grouped frequency table with frequencies (f_i) in intervals with mid–points (m_i). First the sample mean $\bar{x}$ is approximated (as described earlier) and then the sample variance is approximated as

$$s^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^{k} f_i m_i^2 - n(\bar{x})^2 \right\}.$$

Example 3.6

Consider again the data

8.4	8.7	9.0	9.0	9.2	9.3	9.3	9.5	9.6	9.6
9.6	9.7	9.7	9.9	10.3	10.4	10.5	10.7	10.8	11.4

Calculate the sample variance and hence the sample standard deviation.

3.5.4 Box–and–whisker plots

Box and whisker plots are another graphical method for displaying data and are particularly useful in highlighting differences between groups, for example, different spending patterns between males and females or comparing pricing within designated market segments. These plots use some of the key summary statistics we have looked at earlier – the quartiles – and also the maximum and minimum observations.

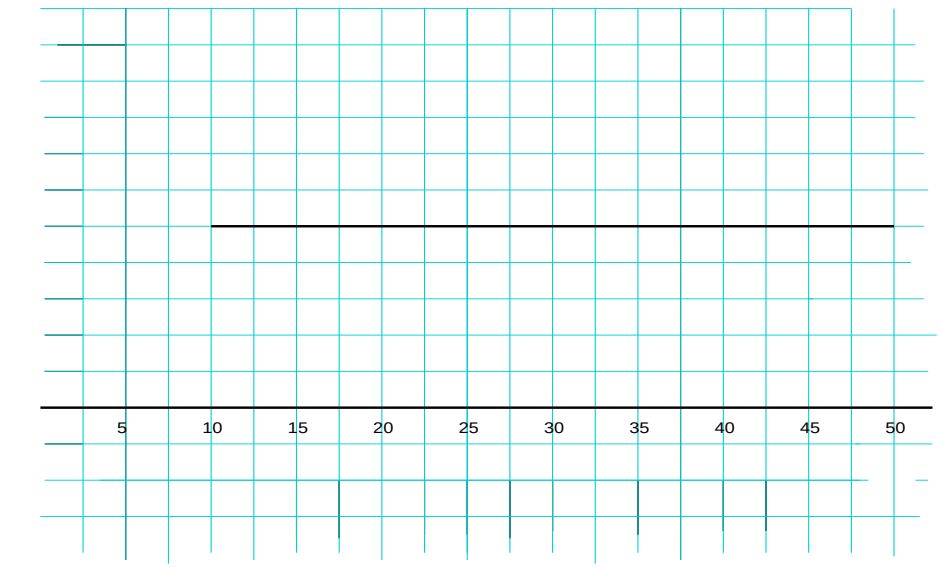
The plot is constructed as follows. After laying out an x–axis for the full range of the data, a rectangle is drawn with ends at the the upper and lower quartiles. The rectangle is split into two at the median. This is the "box". Finally, lines are drawn from the box to the minimum and maximum values – these are the "whiskers".

Example 3.7

Suppose that, from our data, we obtain the following summary statistics:

Minimum	$min = 10$
Lower quartile	$Q1 = 40$
Median	$Q2 = 43$
Upper quartile	$Q3 = 45$
Maximum	$max = 50$

Complete the associated box–and–whisker plot, and comment.



3.6 Chapter 3 practice questions

1. Identify the type of data described in each of the following examples:

(a) An opinion poll was taken asking people which party they would vote for in a general election.

(b) In a steel production process the temperature of the molten steel is measured and recorded every 60 seconds.

(c) A market researcher stops you in Northumberland Street and asks you to rate between 1 (disagree strongly) and 5 (agree strongly) your response to opinions presented to you.

(d) The hourly number of units produced by a beer bottling plant is recorded.

2. Reem AlBarri, Andrew Alcock and Sarah Koullapi all work as business analysts for a leading clothing manufacturer. The design team want to know how successful a new pair of jeans will be, and so they ask the analysts to conduct some research.

(a) Andrew is lazy. He goes home that night and asks his four flatmates to give the new jeans a score from 1–5, and then reports his findings back to the design team the next day.

What sort of sampling scheme has Andrew adopted, and why might it be flawed?

(b) Reem thinks she will try to collect a simple random sample of people who might wear the jeans; she will then try to elicit their opinions. Why might a simple random sample be difficult to obtain here?

(c) Sarah is more thoughtful than both Reem and Andrew. She thinks about the target population for the new jeans: females in the age group 21–30. She then spends a full day walking up and down Northumberland Street, stopping people he thinks fulfil this criteria and asking them to decide whether or not they would buy the new jeans.

What sort of sampling scheme has Sarah adopted?

3. Clark Taylor is the managing director of *Taylor's Taytos*, an Irish potato company supplying potatoes to a leading manufacturer of crisps. The following table shows the weight (in kilograms) of sacks of potatoes leaving her factory yesterday:

10.04	9.38	9.75	11.23	10.80	10.45	9.89
12.48	9.77	10.46	10.05	10.32	10.66	10.50

Display these data in a stem–and–leaf plot. Note the number of decimal places and adjust accordingly. Clearly state both the stem and leaf units.

4. (a) There are 200 students taking ACC1012: 120 female students and 80 male students. We are interested in how often ACC1012 students use their mobile telephones to make calls. Rather than survey all 200 students, we randomly sample 30 female students and 20 male students, and ask them to record the number of calls they make, from their mobile phone, over a one month period; the results are shown below.

98	99	99	100	100
101	100	104	97	101
102	100	99	101	99
100	96	99	101	99
99	98	95	99	99
97	101	100	101	101
103	102	96	98	103
98	100	102	99	101
98	99	100	98	99
102	98	99	99	97

(i) What form of sampling technique has been used to select these 50 students? Explain your answer, and give one advantage and one disadvantage of this type of sampling. Is this form of sampling truly random, quasi–random or non–random?

(ii) Put these data into a relative frequency table, and comment.

(b) From a complete list of all 200 ACC1012 students, one name is picked at random: Samuel James Paul Ribchester. Last month, Samuel made 50 phone calls from his mobile phone.

(i) What sampling technique was used to select Samuel?

(ii) Comment on Samuel's telephone usage relative to the rest of the class.

(iii) The following data are the recorded length (in seconds) for the 50 calls Jake made last month (written in ascending order). Construct a frequency table appropriate for these data, and use this to construct a histogram.

257.7226	259.6408	263.5242	267.5344	267.9781
270.7399	278.3108	281.1613	281.4837	283.6594
285.9805	286.6464	286.9626	289.5667	292.0031
292.0917	292.6725	293.2735	293.4027	293.9145
293.9364	298.4445	299.6535	300.1725	300.5140
301.6963	302.4314	302.5770	303.3484	303.9191
304.1124	304.6044	305.4378	306.5106	306.9344
310.2583	310.9137	311.5926	312.7291	312.9645
313.9611	314.8500	314.8501	317.9180	320.7182
323.7993	326.9056	327.7353	337.5806	346.4497

5. Consider the following data for daily sales at a small record shop, before and after a local radio advertising campaign.

Daily Sales	**Before**	**After**
$1000 \leq$ sales < 2000	10	7
$2000 \leq$ sales < 3000	30	10
$3000 \leq$ sales < 4000	40	25
$4000 \leq$ sales < 5000	20	35
$5000 \leq$ sales < 6000	15	37
$6000 \leq$ sales < 7000	12	40
$7000 \leq$ sales < 8000	10	20
$8000 \leq$ sales < 9000	8	10
$9000 \leq$ sales < 10000	0	5
Totals	145	189

(a) Calculate the percentage relative frequency for before and after.

(b) Plot the relative frequency polygons for both on one graph and comment.

6. A market researcher asked 650 students what their favourite daily newspaper was. The results are summarised in the frequency table below. Represent these data in an appropriate graphical manner.

The Times	140
The Sun	200
The Sport	50
The Guardian	120
The Financial Times	20
The Mirror	80
The Daily Mail	10
The Independent	30

7. Recall the data from question 3 on the weight (in kg) sacks of potatoes leaving *Taylor's Taytos* potato farm. This was a subset of observations from the following larger sample:

10.4	10.0	9.3	11.3	9.6
11.2	10.5	8.5	10.4	8.2
9.3	9.6	10.3	10.0	11.5
11.3	10.8	8.9	10.0	9.5
10.0	11.3	11.0	9.7	10.6
9.9	10.2	10.6	10.2	8.1
8.7	9.4	10.9	10.0	9.9
9.2	11.6	9.6	9.5	10.4
10.6	8.8	10.1	10.3	9.7
10.7	10.6	12.8	10.6	10.2

(i) Calculate the mean of the data.

(ii) Put the data in a grouped frequency table.

(iii) Estimate the sample mean from the grouped frequency table. Why is this an *estimate* of the sample mean?

(iv) Calculate the median of the data.

(v) Find the modal class.

(vi) Calculate the range of the data.

(vii) Calculate the inter–quartile range.

(viii) Calculate the sample standard deviation.

8. The following observations are the number of passenger planes landing on runway East 2 at New York's JFK airport, per hour, over a ten hour period on Friday 23rd December 2011:

30 28 27 33 35 32 25 27 30 1

(a) Estimate the average number of passenger planes arriving, per hour, using (i) the mean and (ii) the median.

(b) Summarise the spread of these data by calculating (i) the range; (ii) the inter–quartile range and (iii) the standard deviation.

(c) Runway East 2 was closed for most of the final hour during the observation period due to a snowstorm. Given this information, comment on the dataset above and suggest which of your summaries of location and spread would be the most suitable. Which would be the *least* suitable? Why?

(d) Draw a box and whisker plot for these data.

9. Chloe collected the following data on the weight, in grams, of "large" chocolate chip cookies produced by Millie's Cookie Company.

27.1 22.4 26.5 23.4 25.6 26.3 51.3 24.9 26.0 25.4

To summarise, Chloe was going to calculate the mean and standard deviation for this sample. However, her friend Mark warned her that the mean and standard deviation might be inappropriate measures of location and spread for these data.

(a) Do you agree with Mark? If so, why?

(b) Mark suggested the *geometric mean* as an alternative to the standard sample mean, given by

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i},$$

where the notation $\prod$ represents the *product*, as opposed to $\sum$ which represents the *sum*.

(i) Calculate the geometric mean for this dataset.

(ii) Do you think Mark was right to suggest the geometric mean as an alternative measure of average? Explain your answer.

(c) Calculate measures of location and spread that you feel are more suitable.