MAS187/AEF258

University of Newcastle upon Tyne

2005-6

Contents

Chapter 1

Collecting and Presenting Data

1.1 Introduction

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

1.1.1 Examples

Sizing Clothes Most clothing now comes in essentially standard sizes but from where do these standards come? By sampling from the general population as a whole, standards can be set around the most common sizes. We can not say that an individual is exactly a standard size. However we can say that they will probably fall within a range either side of a standard.

Car Maintenance If we were buying a new car, it would be useful to know how much it was going to cost to run it over the next three years. Obviously we can not predict this exactly as each individual car and each user will be slightly different. Collecting data from people who have bought similar cars will give us some idea of the distribution of costs over the population of car buyers, which in turn provides us with information as to the likely cost of running the car.

1.1.2 Definitions

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 management courses at Newcastle, that would be our random variable, a particular person’s height would be the observation and if we measured everyone doing MAS187, those would be our data, which form a sample from the population of all students registered with the School of Management.

In practice it is difficult to observe whole populations, unless we are interested in a very limited population, e.g. the students taking MAS187. In reality we are usually observe a subset of the population, we will come back to sampling later in section 1.2.

Variables are of two types, *qualitative* and *quantitative*. Qualitative variables have non-numeric outcomes. They are usually *categorical*. Examples of qualitative variable include sex of a person or animal, colour of a car, mode of transport, football team supported. Quantitative variables have numeric outcomes with a natural ordering. Examples include people's height, time to failure of a component, number of defective components in a batch.

Quantitative variables are usually of one of two types: *discrete* or *continuous*. Discrete random variables can only take a sequence of distinct values which are usually the integers, although not necessarily so. Discrete variables are countable, for example the number of defective pieces in a manufacturing batch, the number of people in a tutorial group, or a person's shoe size. There are other kinds of discrete data. *Ordinal* data are data are ordered but which are not really numbers in the usual sense. For example, if you are asked to rank a response to a question between 1 and 10, from strongly disagree to strongly agree, an answer of 8 obviously indicates stronger agreement than one of 4, but not necessarily twice as strong in any meaningful sense.

Continuous variables can take any value over some continuous scale. Simple examples include height, weight, time taken to be served in a bank queue or the fuel consumption of a car. The important thing to note about continuous data is that, no matter how small an interval we consider, it is always possible (in theory, at least) to make an observation in the interval by using sufficiently precise measurement. We can measure to differing degrees of accuracy using different equipment but we could never say absolutely precisely how much someone weighs. Continuous variables are often expressed up to a number of significant digits and could appear to be discrete. It is the underlying variable which defines their status and not the form in which they are expressed.

1.1.3 Surveys

Surveys or questionnaires are often used to gain insight into the impact of many management decisions. For example, the market prospects of a new product or customer views on the impact and potential of the new technologies. When preparing a survey there are many key questions to consider:

- what is the purpose of the survey?
- what is the target population?
- is there a list of the target population?
- how can bias be avoided?
- how accurate does the survey have to be?
- what resources are available for conducting the survey?

- how are the data to be collected?

There are many ways of collecting survey information. Each has its advantages and disadvantages regarding the cost of implementation, the response rate (of successfully completed questionnaires), the speed with which the survey can be completed and the quality and accuracy of the information collected. The three main ways are:

Postal questionnaire – low cost, low response rate, slow turn around time, low quality information

Telephone interviewing – moderate cost, moderate response rate, fast turn around time, good quality information (?)

Face-to-face interviewing – high cost, high response rate, fast turn around time, high quality information

1.2 Sampling

We can rarely observe the whole population. Instead we observe some sub-set of this called the sample. The difficulty is in obtaining a *representative* sample. For example if you were to ask the people leaving a gym if they took exercise this would produce a *biased* sample and would not be representative of the population as a whole. The importance of obtaining a representative sample can not be stressed too highly. As we will see later 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.

Random sampling where the members of the sample are chosen by some random mechanism.

Quasi-random sampling where the mechanism for choosing the sample is only partly random.

Non-random sampling where the sample is specifically selected rather than randomly selected.

1.2.1 Simple Random Sampling

This method is the simplest to understand. 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 find this information out. Not all elements of the population are equally accessible and hence you could waste time and money 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. Using our software example you could pick by chance only companies that have recently updated their software and hence would not be interested in your new package.

1.2.2 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, if we assume that the population is 55% female and 45% male and we wanted a sample of 1000. We would 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 can 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. We still need to know the entire population so as to sample from it.

1.2.3 Systematic Sampling

This is a form of quasi-random sampling which can be used where the population is clearly structured. For example 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. This 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.

1.2.4 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 areas which might be difficult or expensive to sample from. Multi-stage sampling works, for example, by dividing the area into geographically distinct area, randomly selecting one of these areas and then sampling, whether by random, stratified or systematic sampling schemes within this area. 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.

1.2.5 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 the cluster are then surveyed. This differs from multi-stage sampling covered in section 1.2.4 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 the 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. (Technically, this is not strictly *bias* but *innefficiency* but, for now, it should be clear that there is a problem).

1.2.6 Judgemental sampling

This is an entirely non-random method of sampling. 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 and hence it can not be guaranteed to be representative. It is prone to bias.

1.2.7 Accessibility sampling

Here 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.

1.2.8 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.

1.2.9 Sample Size

When considering collecting data, 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, questions 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 1000.

1.3 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 common tables used and 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 see in the next lecture.

1.3.1 Frequency Tables

The following table presents the modes of transport used daily by 30 students to get to and from University.

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

The data presented in the tables above are, of course, categorical. However other forms of data can also be presented in frequency tables. The following table shows the raw data for car sales at a new car showroom over a two week period in July.

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

Presenting these data in a relative frequency table by number of days on which numbers of cars were sold, we get the following table:

Cars Sold	Tally	Frequency	Relative Frequency %
Totals			

1.3.2 Continuous Data Frequency Tables

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. 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 also 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”). Thirdly the class interval width should be a convenient number, for example 5, 10, 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. Bearing this in mind, let us create a frequency table for these data. As with discrete data frequency tables, we might also be interested in the percentage relative frequency of each class. This is simply calculated by taking the number in the class, dividing it by the total number in the sample and then multiplying this by 100% to obtain a percentage.

The data above give the following frequency table.

Class Interval	**Tally**	**Frequency**	**Relative Frequency %**
Totals			

1.4 Exercises 1

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

1. An opinion poll was taken asking people which party they would vote for in a general election.
2. In a steel production process the temperature of the molten steel is measured and recorded every 60 seconds.
3. 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.
4. The hourly number of units produced by a beer bottling plant is recorded.

The following table includes data for the number of telephone call made by 50 students in a month.

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

Put these data into a relative frequency table.

The following data are the recorded length (in seconds) of 50 mobile phone calls made by one student. Construct a frequency table appropriate for these data.

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

Chapter 2

Graphical methods for presenting data

2.1 Introduction

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 presenting graphical representations of data of the types we have seen previously.

2.2 Stem and Leaf plots

Stem and leaf plots are a quick easy and way of graphically representing data. They can be used with both discrete and continuous data. The method for creating a stem and leaf plot is similar to that for creating a grouped frequency table. The first stage, as with grouped frequency tables, is to decide on a reasonable number of intervals which span the range of data. The interval widths for a stem and leaf plot must be equal. Because of the way the plot works it is best to make the width either an integer power of 10 (e.g. $1 = 10^0$ or $10 = 10^1$ or $100 = 10^2$ or $1000 = 10^3$ or …or $0.1 = 10^{-1}$ or $0.01 = 10^{-2}$ or $0.001 = 10^{-3}$ or $\ldots$) or 2 or 5 times a power of 10 (e.g. $20 = 2 \times 10^1$ or $0.05 = 5 \times 10^{-2}$). We can use 2 or 5 because these are factors of 10. We are not free as a result of this condition to choose the boundaries of the intervals. Once we have decided on our class intervals we can construct the stem and leaf plot. This is perhaps best described by demonstration.

Consider the following data, $11, 12, 9, 15, 21, 25, 19, 8$. The first step is to decide on a interval widths which can be the same as the *stem unit*. One obvious choice would be 10s. This would make the *leaf unit* 1. The stem and leaf plot is constructed as below.

0	8	9		
1	1	2	5	9
2	1	5		
Stem	**Leaf**			

$$n = 8, \quad \text{stem unit} = 10, \quad \text{leaf unit} = 1.$$

You can clearly see where the data have been put. The stem units are to the left of the vertical line, while the leaves are to the right. So, for example, our first observation, 11, is made up of a stem unit of one 10 and a leaf unit one 1.

As an example where the interval width is not a power of 10, consider the following observations

$$17, 18, 15, 14, 12, 19, 20, 21, 24, 15.$$

If you were to choose 10 as the stem unit and 1 as the leaf unit, the stem and leaf plot would look like

$$\begin{array}{l}
n = 10 \\
\begin{array}{r|lllllll}
1 & 2 & 4 & 5 & 5 & 7 & 8 & 9 \\
2 & 0 & 1 & 4 & & & &
\end{array} \\
\text{Stem unit} = 10, \quad \text{Leaf unit} = 1.
\end{array}$$

Here the interval width is 20.

There is not much of a visible pattern in the data in this plot. If we choose $5 \times 10^0 = 5$ units as our interval width, the stem unit remaining as 10's, again with 1 as our leaf unit, the stem and leaf plot would look as follows.

$$\begin{array}{l}
n = 10 \\
\begin{array}{r|lllll}
1 & 2 & 4 & & & \\
1 & 5 & 5 & 7 & 8 & 9 \\
2 & 0 & 1 & 4 & &
\end{array} \\
\text{Stem unit} = 10, \quad \text{Leaf unit} = 1.
\end{array}$$

Changing the interval width like this produces a plot which starts to show some sort of pattern in the data. Graphical presentations are intended to draw out such patterns.

Let us work through the following example. The observations in the table below are the recorded time in seconds it takes to get through to an operator at a telephone call centre.

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

$n =$

Stem **Leaf**

Stem unit = Leaf unit =

If there is more than one significant figure in the data, the extra digits are cut not rounded to the nearest value, that is to say 2.97 would become 2.9. To illustrate this, consider the following data on lengths (in cm) of items on a production line:

$$2.97, 3.81, 2.54, 2.01, 3.49, 3.09, 1.99, 2.64, 2.31, 2.22.$$

The stem and leaf plot for this is as follows.

$n = 10$

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

Stem unit $= 1$ cm, Leaf unit $= 0.1$ cm.

Here the interval width is $5 \times 10^{-1} = 0.5$. This allows for greater clarity in the plot.

2.2.1 Using Minitab

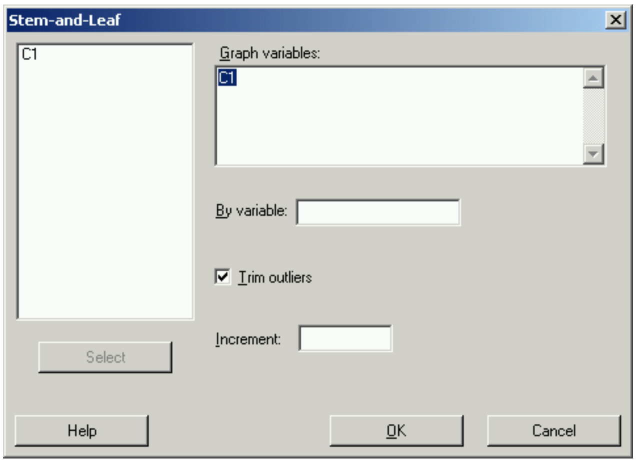
With the small data sets we have seen so far, it is obviously relatively easy to create the stem and leaf plots by hand. With larger data sets this would be more problematic and certainly time consuming. Fortunately there are computer packages that will create these plots for us. **MINITAB** is one such package and we will be using this as an example of what it is possible to achieve using computers. **MINITAB** is an application found on university PC clusters and is run by clicking on

```
Start > All Programs > Statistical Software > Minitab 14 > Minitab 14
```

You will see two windows: a session window and a worksheet. Data are entered into columns labelled C1, C2, C3 etc in the worksheet. Suppose C1 contains some data. To obtain a stem and leaf plot of these data you would need to do the following:

```
Graph > Stem-and-Leaf...
```

This brings up the window below. You need to type in C1 under `Variable` and click `OK`. If you want you can choose the stem unit by entering a value in `Increment` first, otherwise the programme selects this for you.



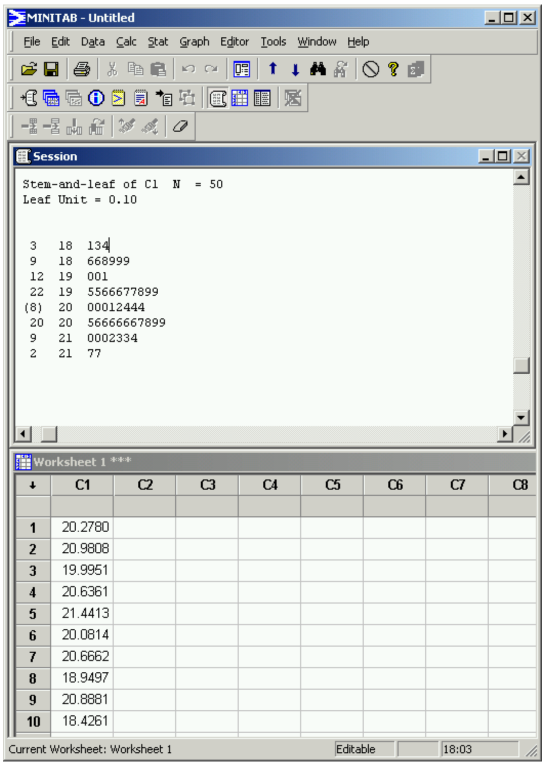
Graph variables: C1

By variable:

☑ Trim outliers

Increment:

This creates a stem and leaf plot in the session window:



```
Stem-and-leaf of C1  N  = 50
Leaf Unit = 0.10


  3   18  134
  9   18  668999
 12   19  001
 22   19  5566677899
(8)   20  00012444
 20   20  56666667899
  9   21  0002334
  2   21  77
```

↓	C1	C2	C3	C4	C5	C6	C7	C8
1	20.2780							
2	20.9808							
3	19.9951							
4	20.6361							
5	21.4413							
6	20.0814							
7	20.6662							
8	18.9497							
9	20.8881							
10	18.4261							

Current Worksheet: Worksheet 1 | Editable | 18:03

It is easy to see some of the advantages of a graphical presentation of data. For example, here you can clearly see that the data are centered around a value in the low 200's and fall away on either side. From stem and leaf plots we can quickly and easily tell if the data are symmetric or asymmetric. We can see whether there are any *outliers*, that is, observations which are either much larger or much smaller than is typical of the data. We could perhaps even tell whether the data are *multi-modal*. That is to say, whether there are two or more peaks on the graph with a gap between them. If so this might suggest that the sample might contain data from two or more groups.

2.3 Bar Charts

Bar Charts are common and clear ways of presenting categorical data or any ungrouped discrete frequency observations. As with stem and leaf plots, various computer packages allow you to produce these with relative ease. First let us work through the process of producing these by hand. This will enable you to get a clear idea of how these charts are constructed.

Constructing a bar chart is a 5 step process:

1. First decide what goes on each axis of the chart. By convention the variable being measured goes on the horizontal (x-axis) and the frequency goes on the vertical (y-axis).

2. Next decide on a numeric scale for the frequency axis. This axis represents the frequency in each category by its height. It must start at zero and include the largest frequency. It is common to extend the axis slightly above the largest value so you are not drawing to the edge of the graph.

3. Having decided on a range for the frequency axis we need to decide on a suitable number scale to label this axis. This should have sensible values, for example, $0, 1, 2, \ldots,$ or $0, 10, 20 \ldots,$ or other such values as make sense given the data.

4. Draw the axes and label them appropriately.

5. Draw a bar for each category. When drawing the bars it is essential to ensure the following:

    - the width of each bar is the same;
    - the bars are separated from each other by equally sized gaps.

Recall the example on students' modes of transport:

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 first logical step is again to put these into a frequency table, giving

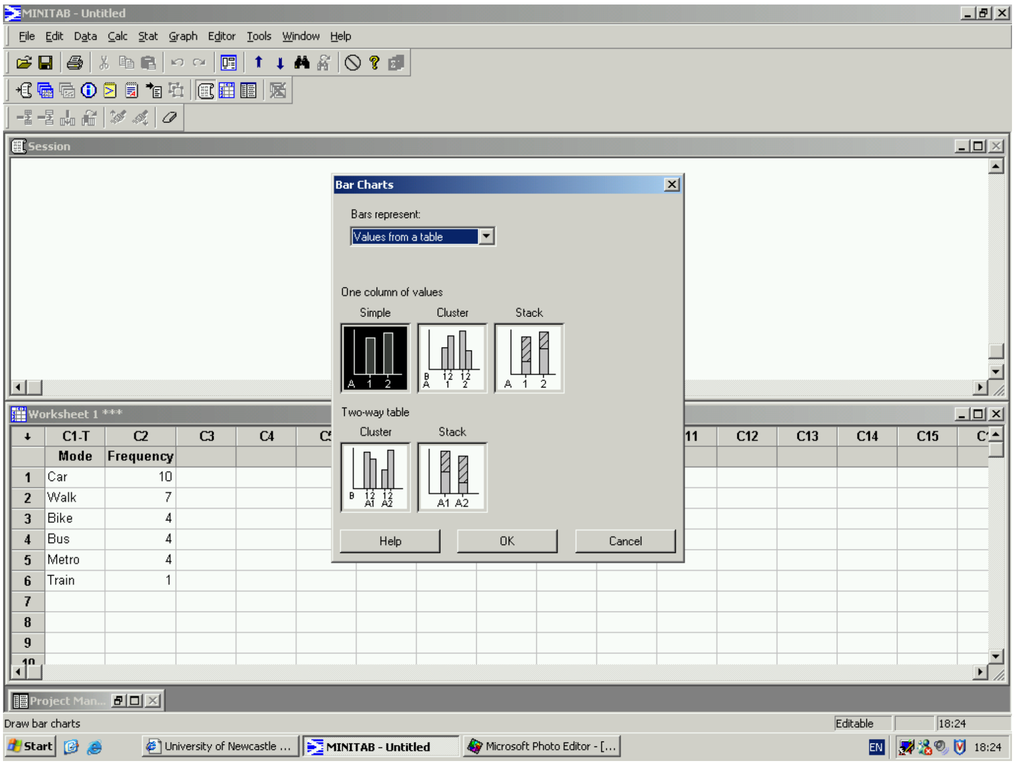
Mode	**Frequency**
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	30

We can then present this information as a bar chart:

Such graphs are easily drawn using **MINITAB**:

1. First enter the data in the worksheet, either in summary format or as raw data, with column C1 containing the categories and the (raw or frequency) counts in column C2.

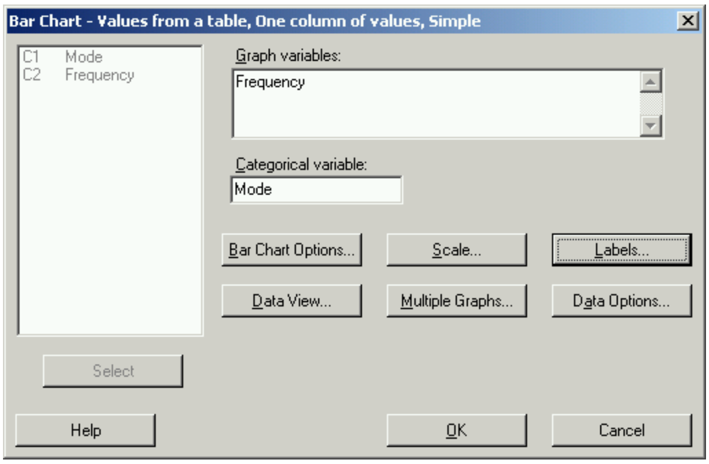
2. `Graph > Bar Chart...`



Bars represent: Values from a table

↓	C1-T	C2
	Mode	Frequency
1	Car	10
2	Walk	7
3	Bike	4
4	Bus	4
5	Metro	4
6	Train	1

3. Select the appropriate data format (raw data or tabulated data), the columns containing the data, and the graph format



Graph variables: Frequency

Categorical variable: Mode

Note: options exist to configure the graph e.g. `Label` can be used to give the graph a title.

4. When ready click on **OK**

This procedure produces the chart

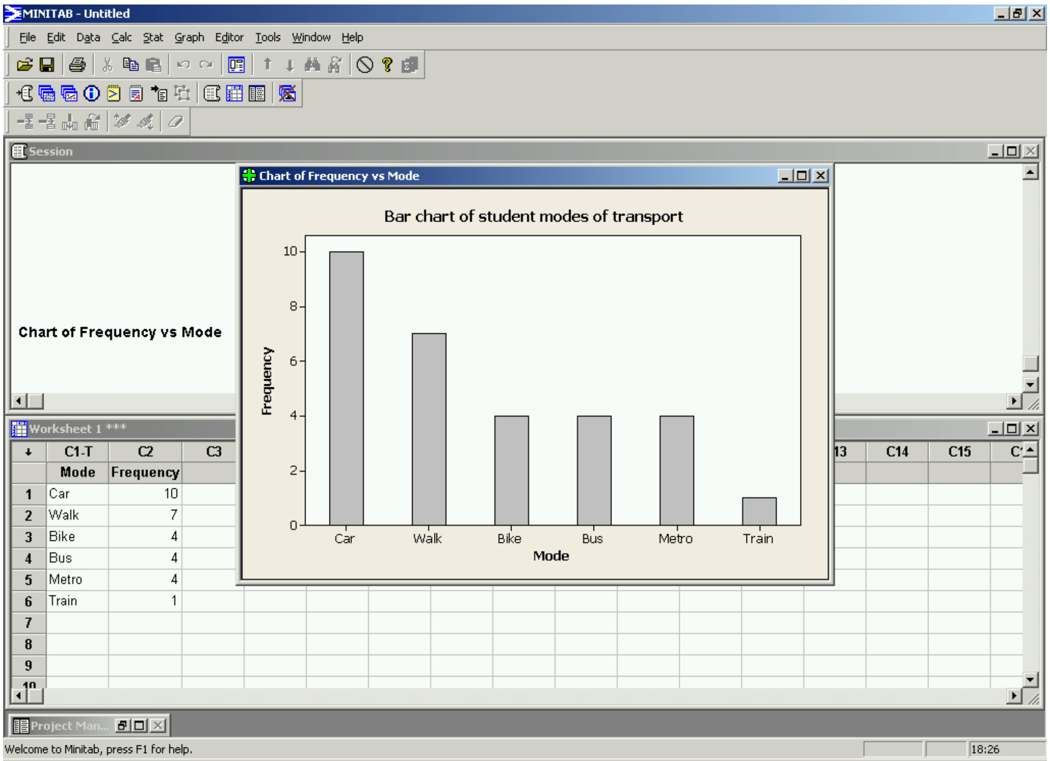

MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Chart of Frequency vs Mode
Bar chart of student modes of transport
Frequency
0
2
4
6
8
10
Car
Walk
Bike
Bus
Metro
Train
Mode
Worksheet 1 ***
C3
13
C14
C15
Project Man...
18:26

Chart of Frequency vs Mode

↓	C1-T	C2
	Mode	Frequency
1	Car	10
2	Walk	7
3	Bike	4
4	Bus	4
5	Metro	4
6	Train	1
7		
8		
9		

Welcome to Minitab, press F1 for help.

This bar chart clearly shows that the most popular mode of transport is the car and that the metro, walking and cycling are all equally popular (in our small sample). Bar charts provide a simple method of quickly spotting simple patterns of popularity within a discrete data set.

2.4 Multiple Bar Charts

The data below gives the daily sales of CDs (in £) by music type for an independent retailer.

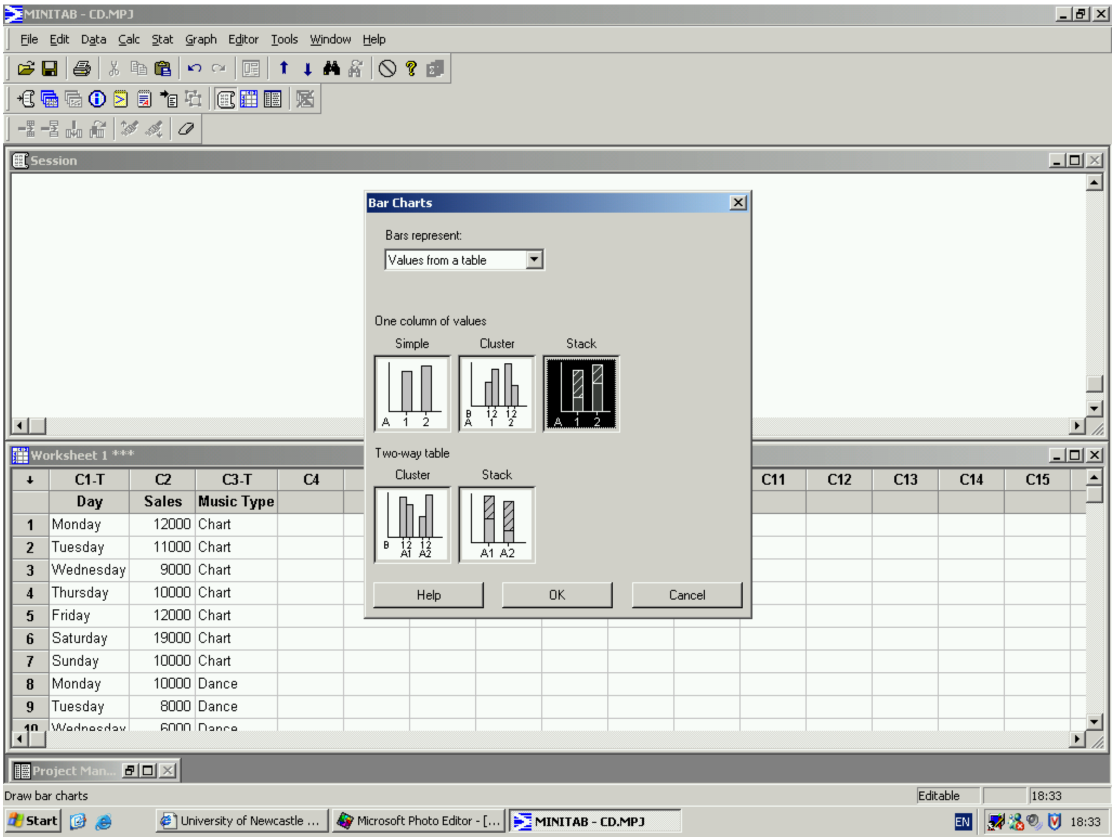
Day	Chart	Dance	Rest	Total
Monday	12000	10000	2700	24700
Tuesday	11000	8000	3000	22000
Wednesday	9000	6000	2000	17000
Thursday	10000	5000	2500	17500
Friday	12000	11000	3000	26000
Saturday	19000	12000	4000	35000
Sunday	10000	8000	2000	20000
Total	83000	60000	19200	162200

Bar charts could be drawn of total sales per music type in the week, or of total daily sales. It might be interesting to see daily sales broken down into music types. This can be done in a similar manner to the bar charts produced previously. The only difference is that the height of the bars is dictated by the total daily sales, and each bar has segments representing each music type. This is done in **MINITAB** as follows:

1. Enter the data into the worksheet, the types of music in columns and the days as rows.

2. `Graph > Bar Chart...`

3. Select the appropriate data format and the `Stack` graph format.

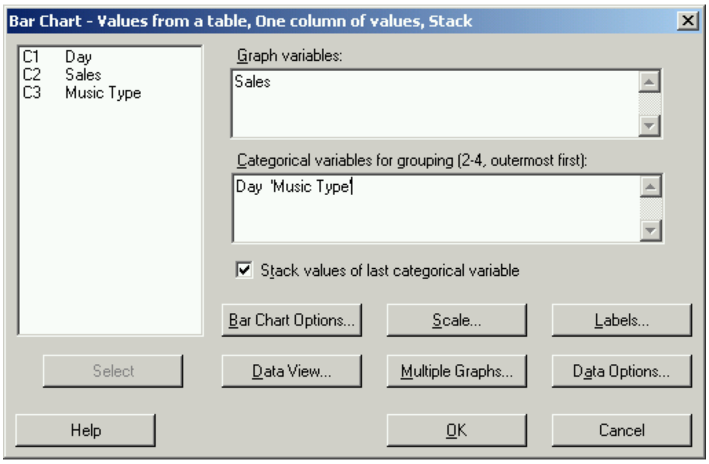


Bars represent: Values from a table

	C1-T	C2	C3-T	C4	C11	C12	C13	C14	C15
	Day	Sales	Music Type						
1	Monday	12000	Chart						
2	Tuesday	11000	Chart						
3	Wednesday	9000	Chart						
4	Thursday	10000	Chart						
5	Friday	12000	Chart						
6	Saturday	19000	Chart						
7	Sunday	10000	Chart						
8	Monday	10000	Dance						
9	Tuesday	8000	Dance						
10	Wednesday	6000	Dance						

4. Click `OK`

5. Enter the column containing the `Sales` data under `Graph variables` and the `Day` and `Music Type` in the grouping dialogue box



Bar Chart - Values from a table, One column of values, Stack

C1 Day
C2 Sales
C3 Music Type

Graph variables: Sales

Categorical variables for grouping (2-4, outermost first): Day 'Music Type'

☑ Stack values of last categorical variable

6. Click `OK`.

The MINITAB worksheet and chart this procedure produces are

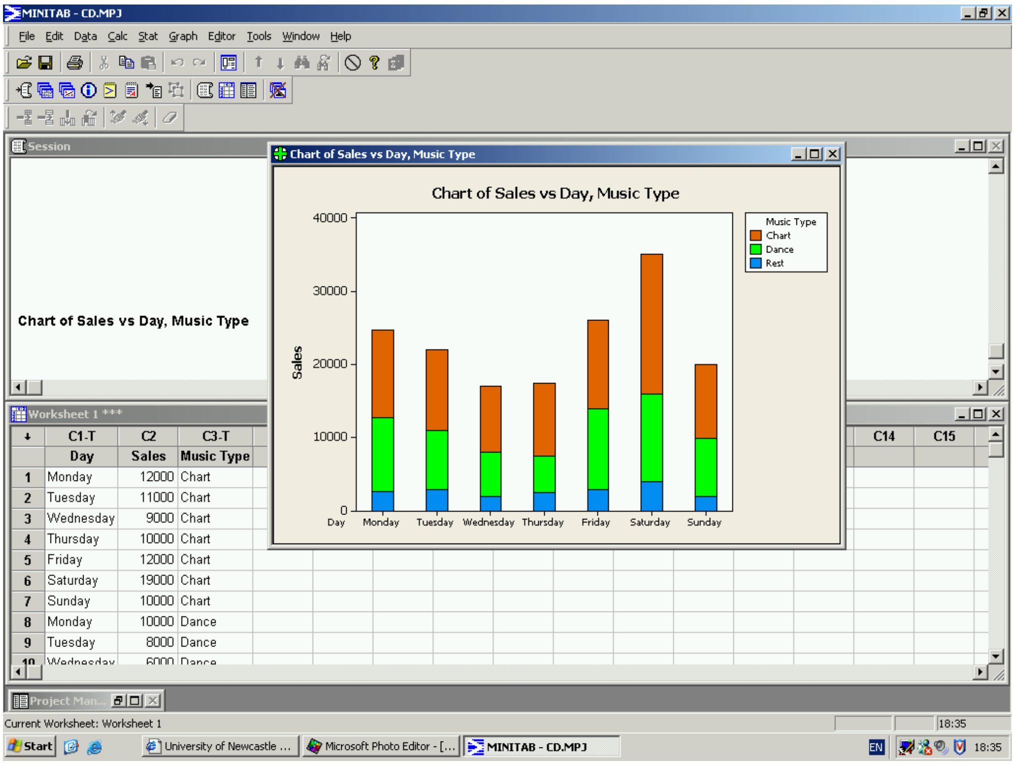


Chart of Sales vs Day, Music Type

↓	C1-T	C2	C3-T
	Day	Sales	Music Type
1	Monday	12000	Chart
2	Tuesday	11000	Chart
3	Wednesday	9000	Chart
4	Thursday	10000	Chart
5	Friday	12000	Chart
6	Saturday	19000	Chart
7	Sunday	10000	Chart
8	Monday	10000	Dance
9	Tuesday	8000	Dance
10	Wednesday	6000	Dance

Current Worksheet: Worksheet 1

These types of charts are particular good for presenting such financial information or illustrating any breakdown of data over time, for example, the number of new cars sold by month and model.

2.5 Histograms

Bar charts have their limitations. They can not be used to present data on continuous variables. When dealing with continuous 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.

Initially we will only consider histograms with equal class intervals. Those with uneven class intervals require more careful thought.

Producing a histogram, is much like producing a bar chart and in many respects can be considered to be the next stage after producing a grouped frequency table. In reality it is often best to produce a frequency table first which collects all the data together in an ordered format. Once we have the frequency table, the process is very similar to drawing a bar chart.

1. Find the maximum frequency and draw the vertical (y-axis) from zero to this value, including a sensible numeric scale.

2. The range of the horizontal (x-axis) needs to include not only the full range of observations but also the full range of the class intervals from the frequency table.

3. Draw a bar for each group in your frequency table. These should be the same width and touch each other (unless there are no data in one particular class).

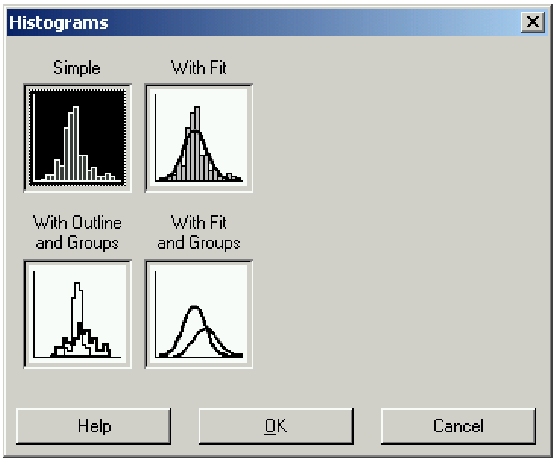
The frequency table for the data on service times for a telephone call centre was

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

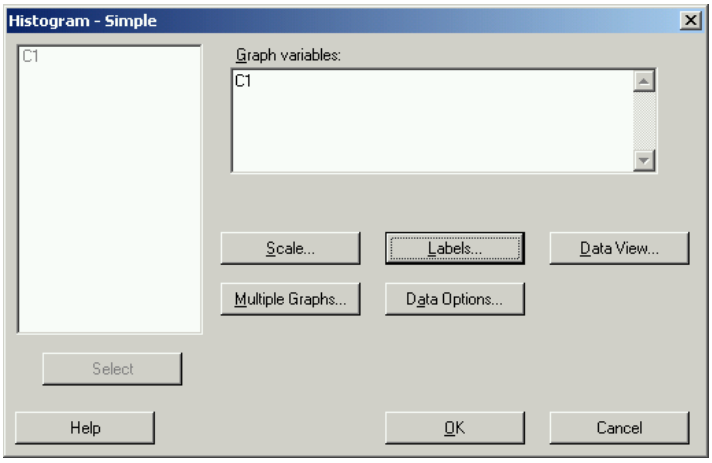
The histogram for these data is

Normally, as with stem and leaf plots and bar charts we would get **MINITAB** to do this for us.

1. Enter the data in column C1 of the worksheet. For illustrative purposes I have randomly generated 500 observations in this column.
2. `Graph > Histogram...`
3. Select the `Simple` graph format



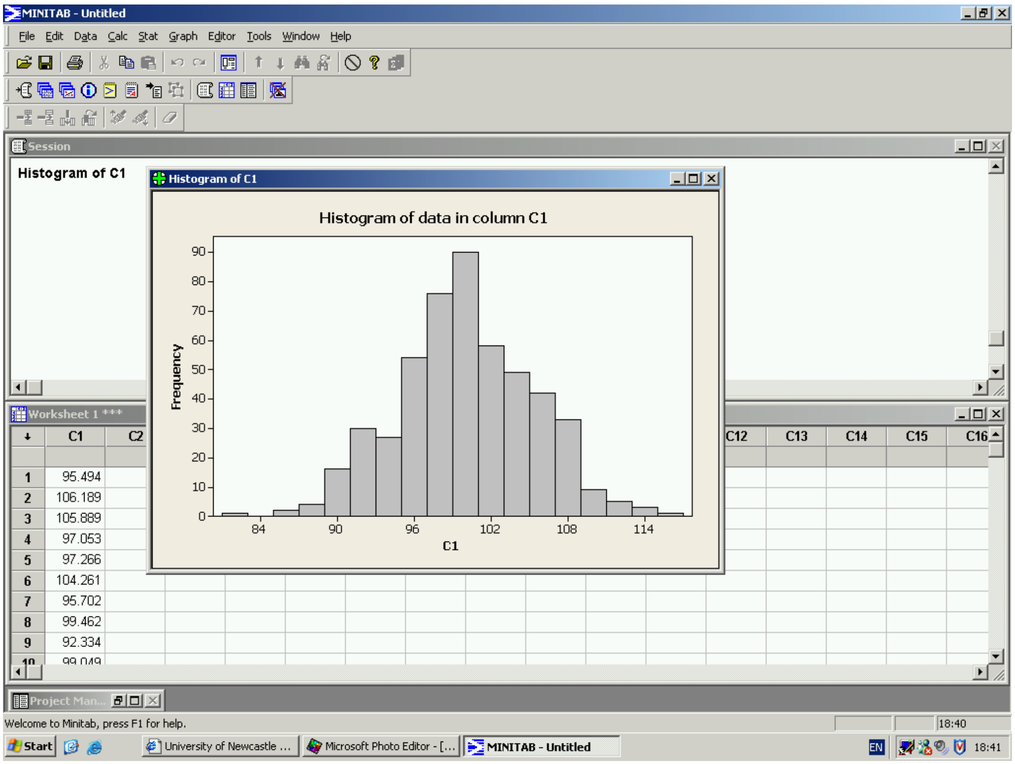
4. Select C1 under `Graph variables`.



Note: various advanced options are available e.g. a title can be added by clicking `Labels`

5. When happy with your choices click `OK`.

These instructions produce the following histogram:

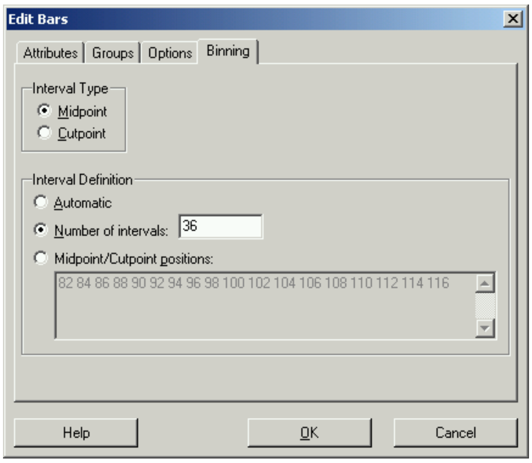

MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Histogram of C1
Histogram of C1
Histogram of data in column C1
Frequency
0
10
20
30
40
50
60
70
80
90
84
90
96
102
108
114
C1
Worksheet 1 ***
C12
C13
C14
C15
C16
Project Man...
Welcome to Minitab, press F1 for help.
18:40
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - Untitled
EN
18:41

↓	C1	C2
1	95.494	
2	106.189	
3	105.889	
4	97.053	
5	97.266	
6	104.261	
7	95.702	
8	99.462	
9	92.334	
10	99.049	

The histogram produced can be amended by `right-clicking` on the graph. For example, the intervals used in the histogram can be changed or, more simply, the number of intervals using

```
Edit bars > Binning
```

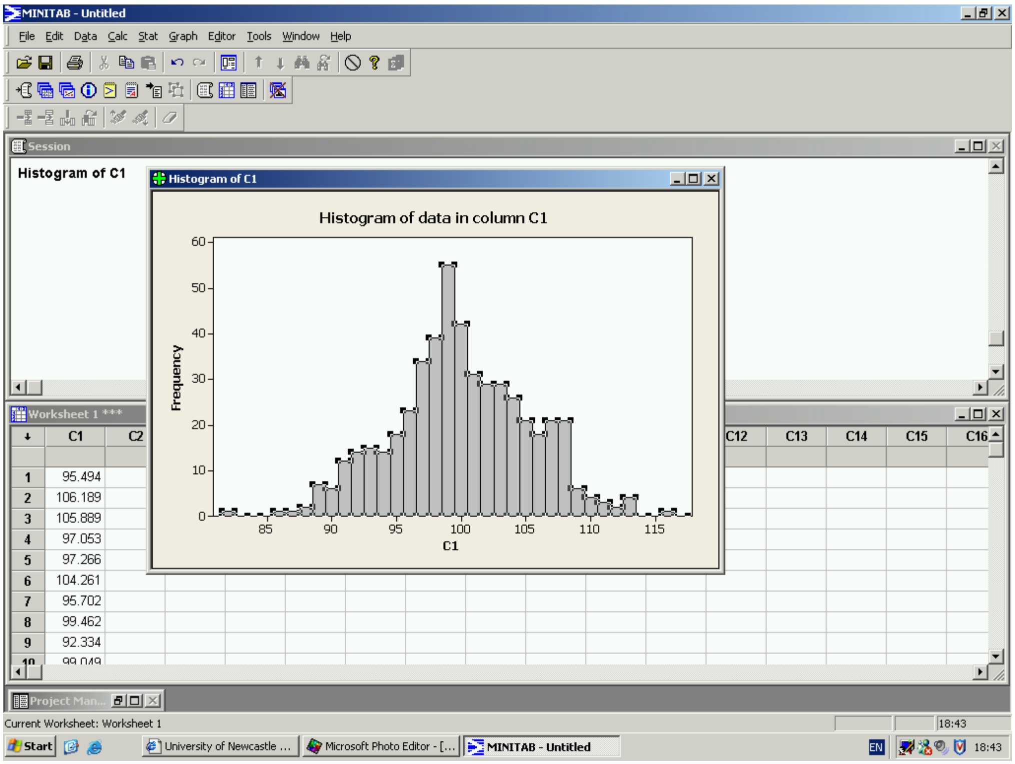
We can double the number of intervals (from 18 to 36 intervals) using the `Binning` dialogue box



Interval Type
● Midpoint
○ Cutpoint

Interval Definition
○ Automatic
● Number of intervals: 36
○ Midpoint/Cutpoint positions:
82 84 86 88 90 92 94 96 98 100 102 104 106 108 110 112 114 116

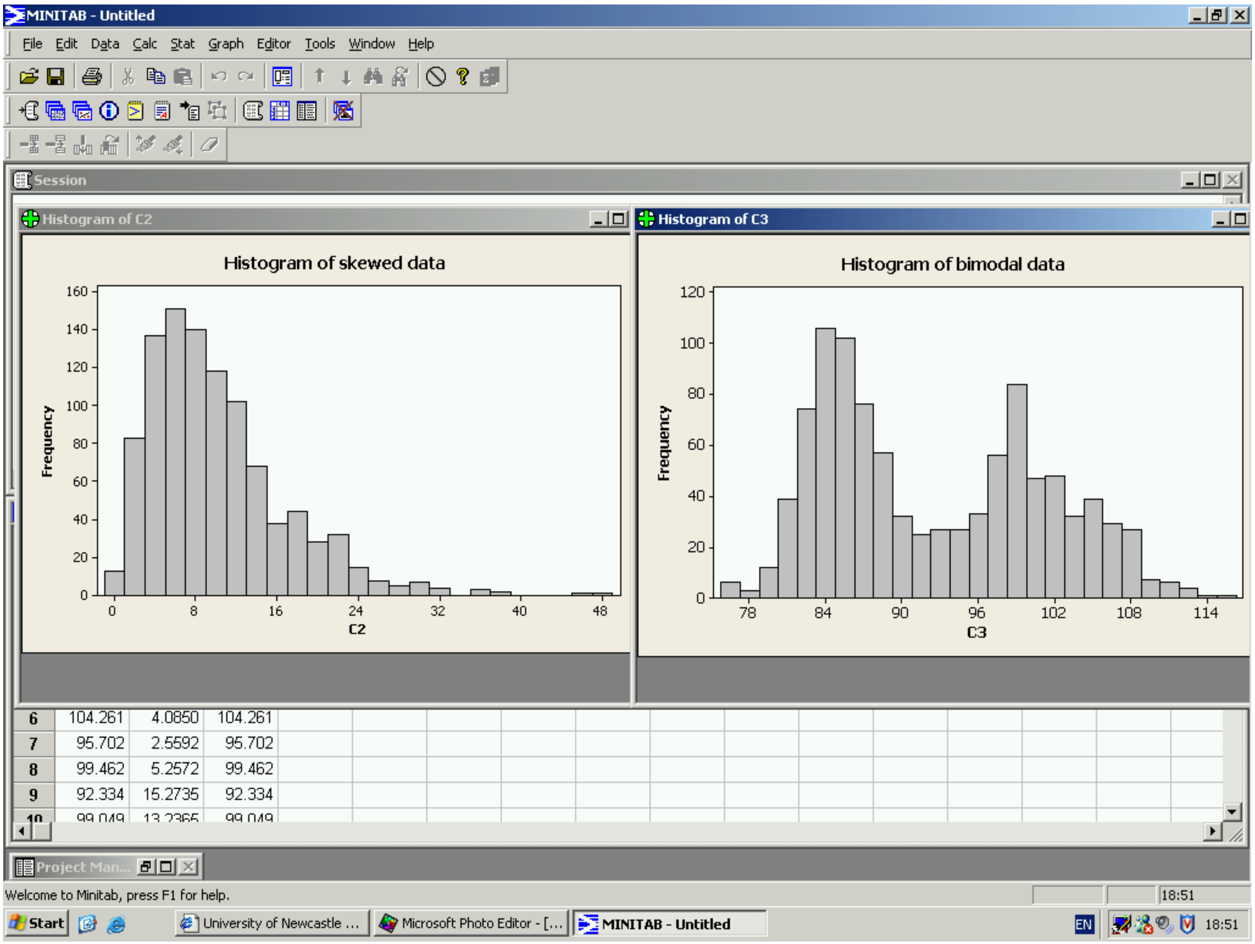
This changes the histogram to



↓	C1	C2
1	95.494	
2	106.189	
3	105.889	
4	97.053	
5	97.266	
6	104.261	
7	95.702	
8	99.462	
9	92.334	
10	99.049	

Histograms are useful tools in data analysis. They are easy to produce in **MINITAB** for large data sets and provide a clear visual representation of the data. Using histograms, it is easy to spot the *modal* or most popular class in the data, the one with the highest peak. It is also easy to spot simple patterns in the data. Is the frequency distribution symmetric, as the histograms produced above, or is it skewed to one side like the left-hand histogram in the following graphic.

Histograms also allow us to make early judgements as to whether all our data come from the same population. Consider the right-hand histogram in the graphic below. It clearly contains two separate modes (peaks), each of which has its own symmetric pattern of data. This clearly suggests that the data come from two separate populations, one centred around 85 with a narrow spread and one centred around 100 with a wider spread. In real situations it is unlikely that the difference would be as dramatic, unless you had a poor sampling method. However the drawing of histograms is often the first stage of more complex analysis.



6	104.261	4.0850	104.261
7	95.702	2.5592	95.702
8	99.462	5.2572	99.462
9	92.334	15.2735	92.334
10	99.049	13.2365	99.049

Welcome to Minitab, press F1 for help.

Finally, be careful when drawing histograms of observations on variables which have boundaries on their ranges. For example heights, weights, times to complete tasks etc. can not take negative values so there is a lower limit at zero. Computer programs do not automatically know this. You should make sure that the lower limit of the first class interval is not negative in such cases.

2.6 Exercises 2

1. The following table shows the weight in kilograms of 50 sacks of potatoes leaving a farm shop.

10.41	10.06	9.38	11.36	9.65
11.24	10.58	8.55	10.47	8.22
9.36	9.63	10.33	10.05	11.57
11.36	10.82	8.93	10.08	9.53
10.05	11.30	11.01	9.72	10.67
9.91	10.26	10.67	10.21	8.18
8.70	9.49	10.98	10.01	9.92
9.27	11.69	9.66	9.52	10.40
10.61	8.83	10.11	10.37	9.73
10.72	10.63	12.86	10.62	10.26

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

2. 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

3. Produce a histogram for the data on length of mobile phone calls in Exercises 1 (listed again below) and comment on it.

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

Chapter 3

More graphical methods for presenting data

3.1 Introduction

We have seen some basic ways in which we might present data graphically. These methods will often provide the mainstay of business presentations. There are, however, other techniques which are useful and offer advantages in some applications over histograms and bar charts.

3.2 Percentage Relative Frequency Histograms

When we produced frequency tables in Chapter 2, 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. Recall the data on service time (in seconds) for calls to a credit card service 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

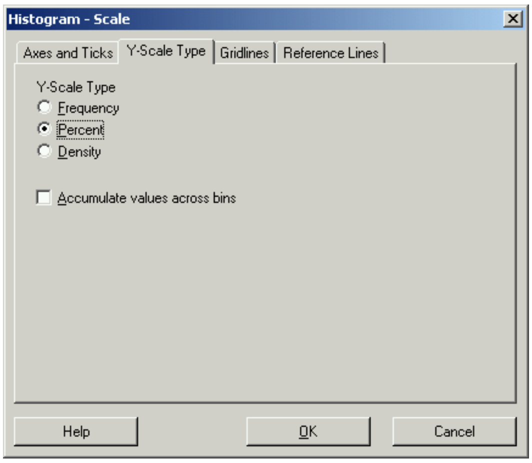
A percentage relative frequency table for these data is

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

You can easily plot these data like an ordinary histogram, except, instead of using frequency on the vertical axis (y-axis), you use the percentage relative frequency.

This can be done in MINITAB as follows.

1. Place the data to be graphed in a column of the worksheet. For illustrative purposes 500 observations have been generated in column C1.
2. `Graph > Histogram`
3. As with ordinary histograms, select the `Simple` graph format, click on **OK**, select column C1 under `Graph variables`.
4. Select `Scale...` then `Y-Scale Type` and check the `Percent` button

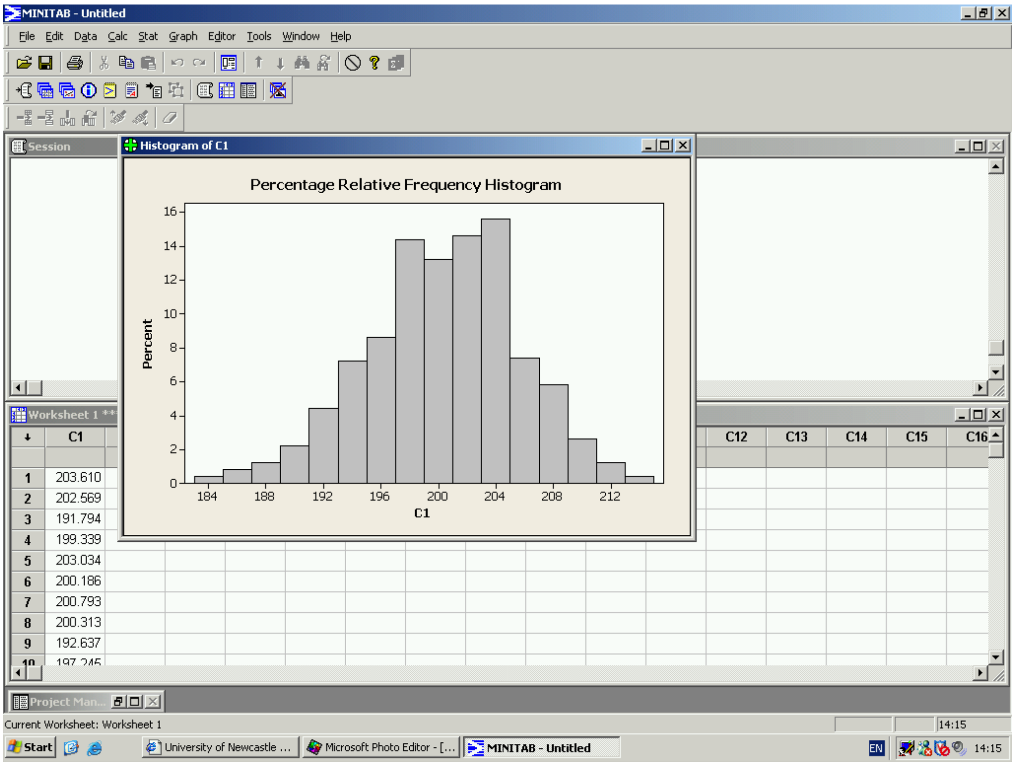


Y-Scale Type
○ Frequency
● Percent
○ Density

□ Accumulate values across bins

5. Click on **OK** and again on **OK**.

This produces the following histogram:

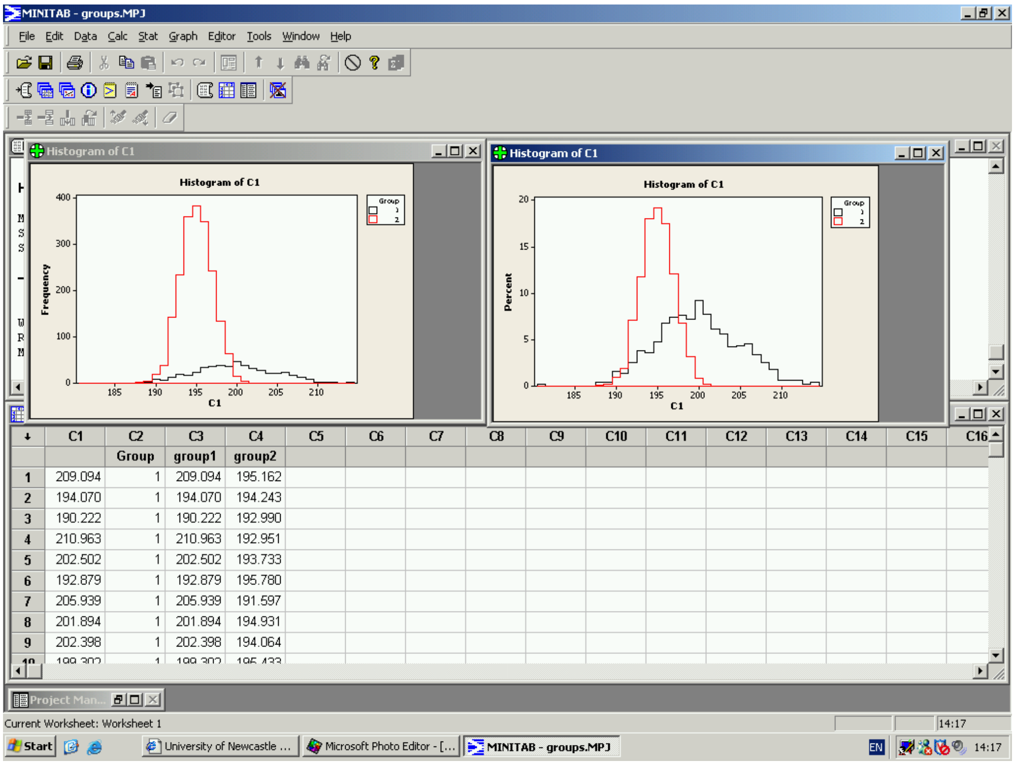

MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Histogram of C1
Percentage Relative Frequency Histogram
Percent
16
14
12
10
8
6
4
2
0
184
188
192
196
200
204
208
212
C1
Worksheet 1 ***
C12
C13
C14
C15
C16
Project Man...
Current Worksheet: Worksheet 1
14:15
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - Untitled
EN
14:15

	C1
1	203.610
2	202.569
3	191.794
4	199.339
5	203.034
6	200.186
7	200.793
8	200.313
9	192.637
10	197.245

Note that the y-axis now contains the relative percentages rather than the frequencies.

You might well ask why we would 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. It is really just a matter of making the vertical scales comparable.

In the following graph 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	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
		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												
10	199.302	1	199.302	195.433												

Current Worksheet: Worksheet 1

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.

3.3 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 the points together with straight lines.

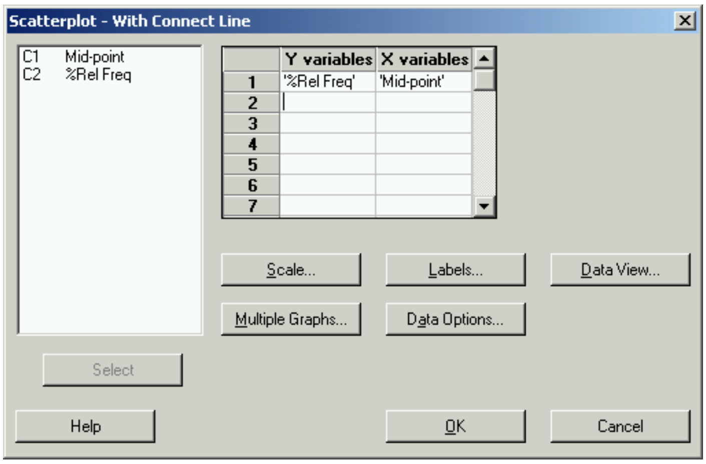
Consider the following simple example.

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.

Alternatively you can use **MINITAB**.

1. Place the data in the worksheet using column C1 for the mid-points and column C2 for the percentage relative frequencies.
2. `Graph > Scatterplot...`
3. Select the `With Connect Line` option and click on **OK**.
4. Enter the column with the percentage frequencies (C2) under `Y variables` and the column with the midpoints (C1) under `X variables`


Scatterplot - With Connect Line

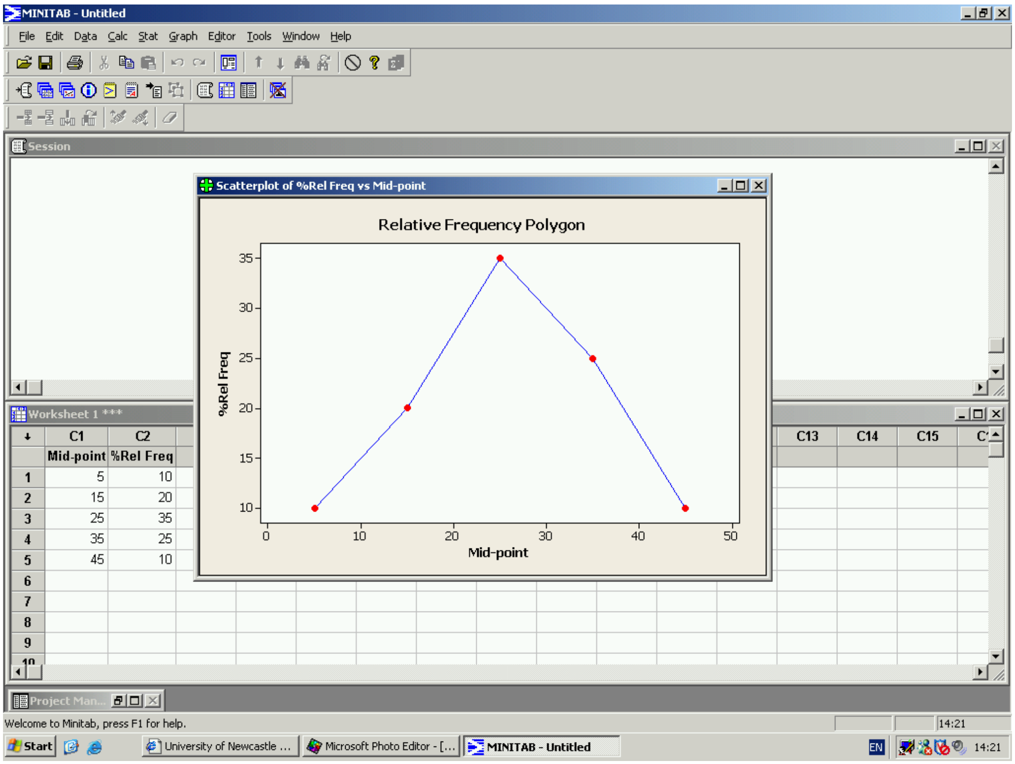
C1 Mid-point
C2 %Rel Freq

	Y variables	X variables
1	'%Rel Freq'	'Mid-point'
2		
3		
4		
5		
6		
7		

5. Add a title by clicking on **Labels...** etc.

6. Click on **OK**.

These instructions produce the graph



MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Scatterplot of %Rel Freq vs Mid-point
Relative Frequency Polygon
35
30
25
20
15
10
%Rel Freq
0
10
20
30
40
50
Mid-point
Worksheet 1 ***
C13
C14
C15
Project Man...
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - Untitled
EN
14:21

↓	C1	C2
	Mid-point	%Rel Freq
1	5	10
2	15	20
3	25	35
4	35	25
5	45	10
6		
7		
8		
9		
10		

Welcome to Minitab, press F1 for help.

These percentage relative frequency polygons are of most use however for comparison between two samples. Consider the following data on gross weekly income 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

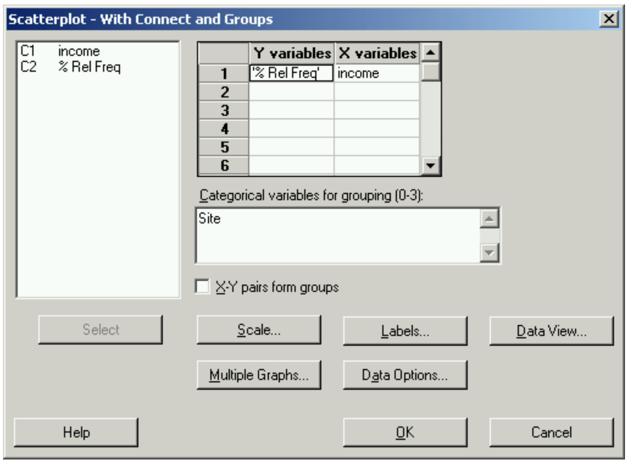
We can produce a graph containing polygons for both locations using MINITAB instructions very similar to those above:

1. Place the data in the worksheet using column C1 for the mid-points, column C2 for the percentage relative frequencies and column C3 for the site where the data were taken.

2. `Graph > Scatterplot...`

3. Select the `With Connect and Groups` option and click on **OK**

4. Enter the column with the percentage frequencies (C2) under `Y variables` and the column with the midpoints (C1) under `X variables`. Also enter the `Site` column (C3) in the box for `Categorical variables for grouping`



Scatterplot - With Connect and Groups

C1 income
C2 % Rel Freq

	Y variables	X variables
1	'% Rel Freq'	income
2		
3		
4		
5		
6		

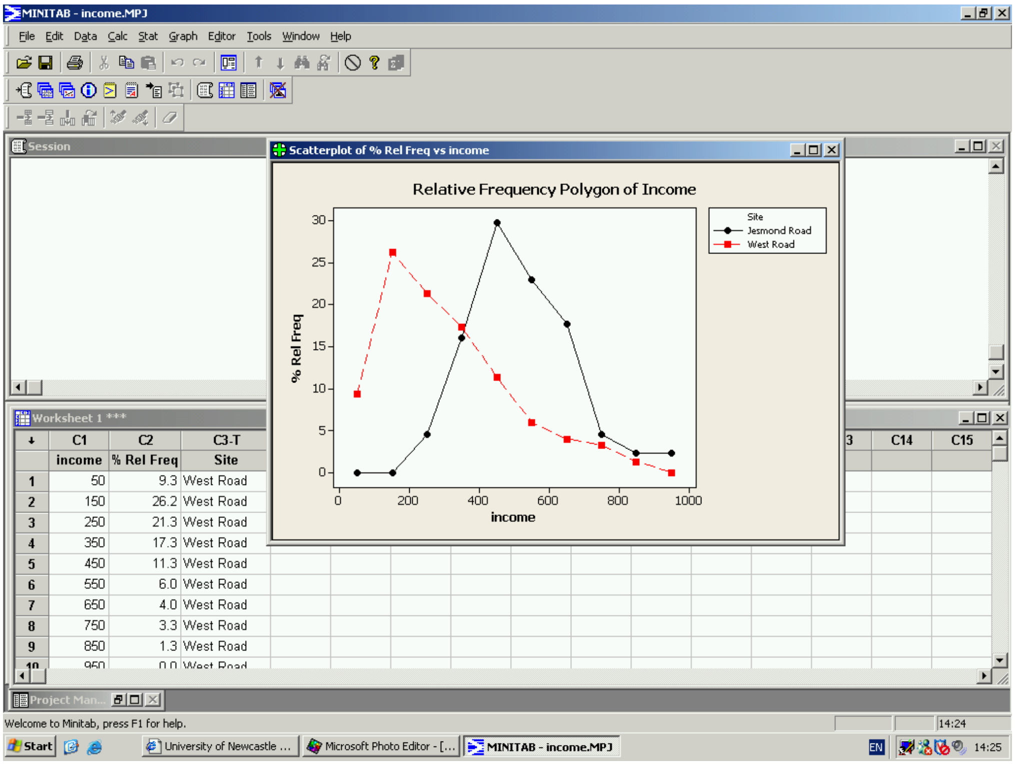
Categorical variables for grouping (0-3): Site

☐ X-Y pairs form groups

5. Add a title by clicking on **Labels...** etc.

6. Click on **OK**.

The polygon produced looks like



MINITAB - income.MPJ

Scatterplot of % Rel Freq vs income

Worksheet 1 ***

	C1	C2	C3-T
	income	% Rel Freq	Site
1	50	9.3	West Road
2	150	26.2	West Road
3	250	21.3	West Road
4	350	17.3	West Road
5	450	11.3	West Road
6	550	6.0	West Road
7	650	4.0	West Road
8	750	3.3	West Road
9	850	1.3	West Road
10	950	0.0	West Road

Welcome to Minitab, press F1 for help.

We can clearly see the differences between the two samples. The line connecting the circles represents the data from the West Road and the line connecting the boxes represents those for Jesmond Road. The distribution of incomes on the 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 on the West Road.

3.4 Cumulative Frequency Polygons (Ogive)

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

This graph can be produced using the following **MINITAB** instructions:

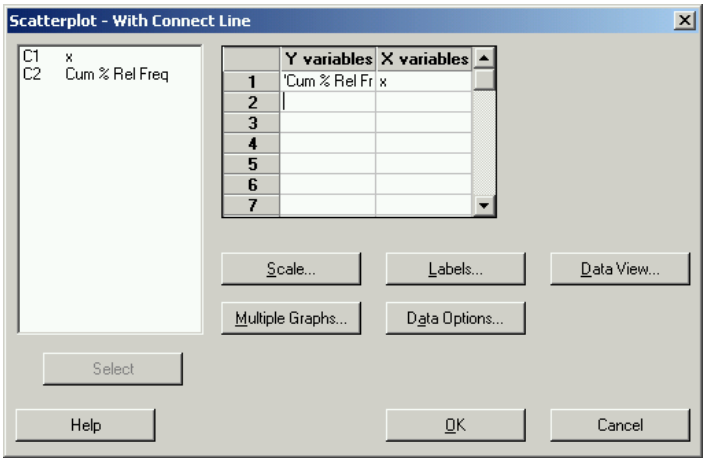
1. In column **C1**, enter the end points of the class intervals, as well the starting point of the smallest class.

2. In column **C2**, enter 0 against the starting point and the cumulative percentage relative frequencies against the relevant end point.

3. `Graph > Scatterplot...`

4. Select the `With Connect Line` option and click on **OK**

5. Enter the column with the percentage frequencies (C2) under `Y variables` and the column with the midpoints (C1) under `X variables`



Scatterplot - With Connect Line

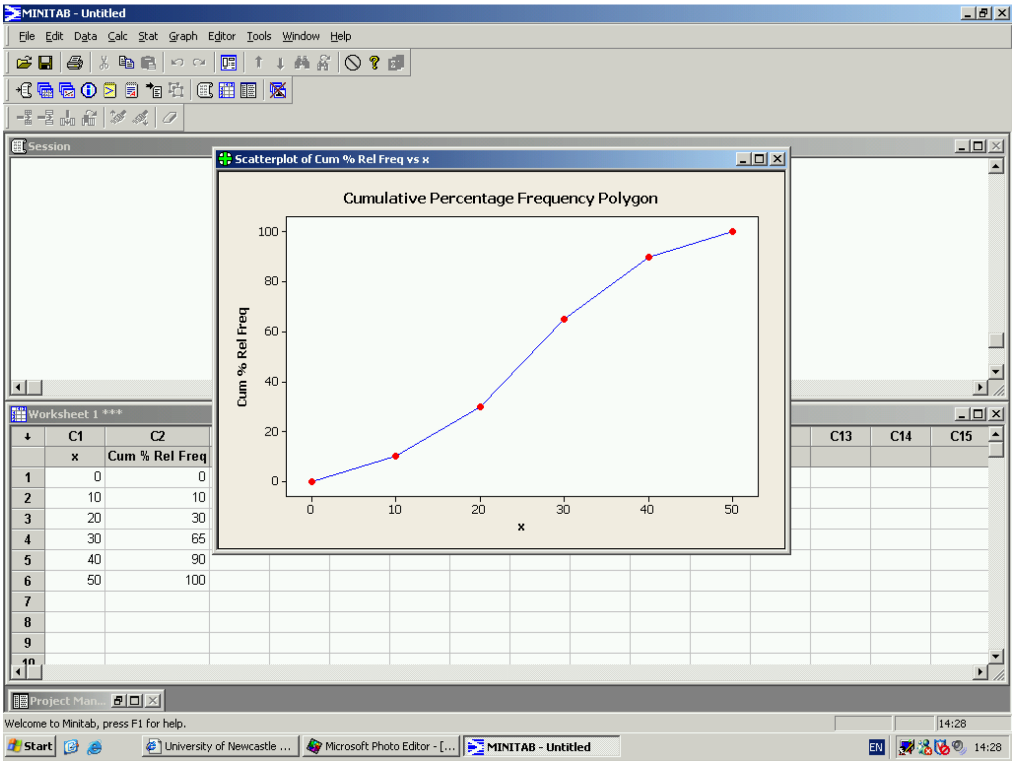
C1 x
C2 Cum % Rel Freq

	Y variables	X variables
1	'Cum % Rel Fr	x
2		
3		
4		
5		
6		
7		

6. Add a title by clicking on **Labels...** etc.

7. Click on **OK**.

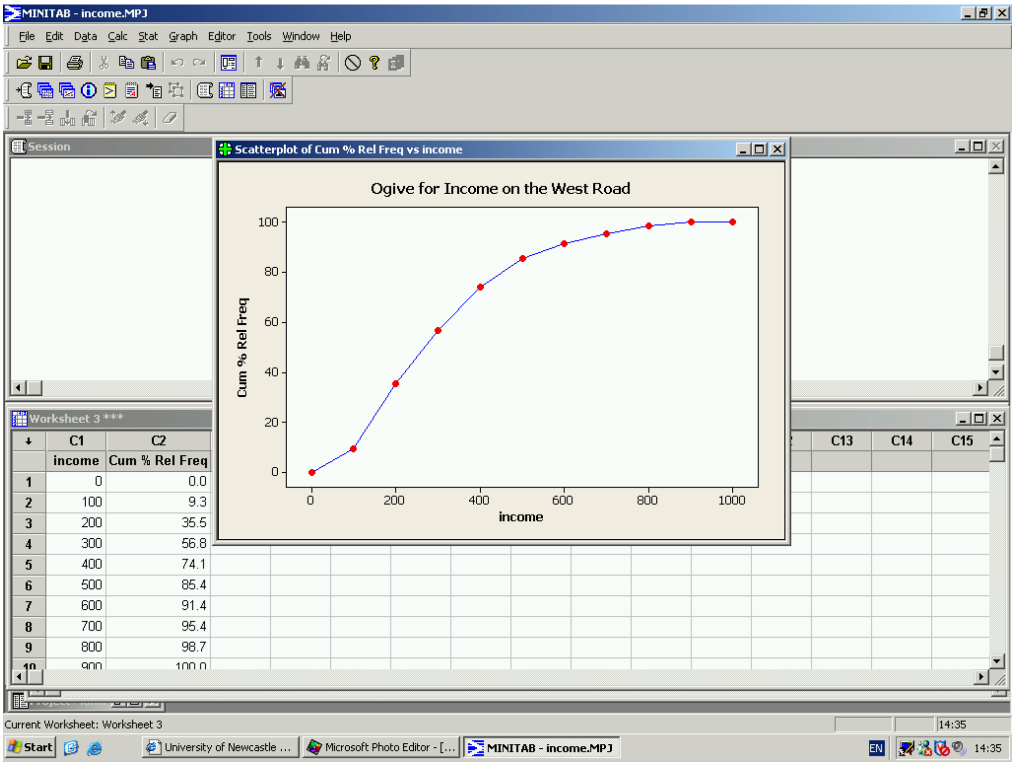
This produces the following graph:


MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Scatterplot of Cum % Rel Freq vs x
Cumulative Percentage Frequency Polygon
Cum % Rel Freq
100
80
60
40
20
0
0
10
20
30
40
50
x
Worksheet 1 ***
C13
C14
C15
Project Man...
14:28
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - Untitled
EN
14:28

↓	C1	C2
	x	Cum % Rel Freq
1	0	0
2	10	10
3	20	30
4	30	65
5	40	90
6	50	100
7		
8		
9		

Welcome to Minitab, press F1 for help.

Applying this procedure to the income data from the West Road survey gives the ogive:


MINITAB - income.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Scatterplot of Cum % Rel Freq vs income
Ogive for Income on the West Road
Cum % Rel Freq
100
80
60
40
20
0
0
200
400
600
800
1000
income
Worksheet 3 ***
C13
C14
C15
14:35
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - income.MPJ
EN
14:35

↓	C1	C2
	income	Cum % Rel Freq
1	0	0.0
2	100	9.3
3	200	35.5
4	300	56.8
5	400	74.1
6	500	85.4
7	600	91.4
8	700	95.4
9	800	98.7
10	900	100.0

Current Worksheet: Worksheet 3

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

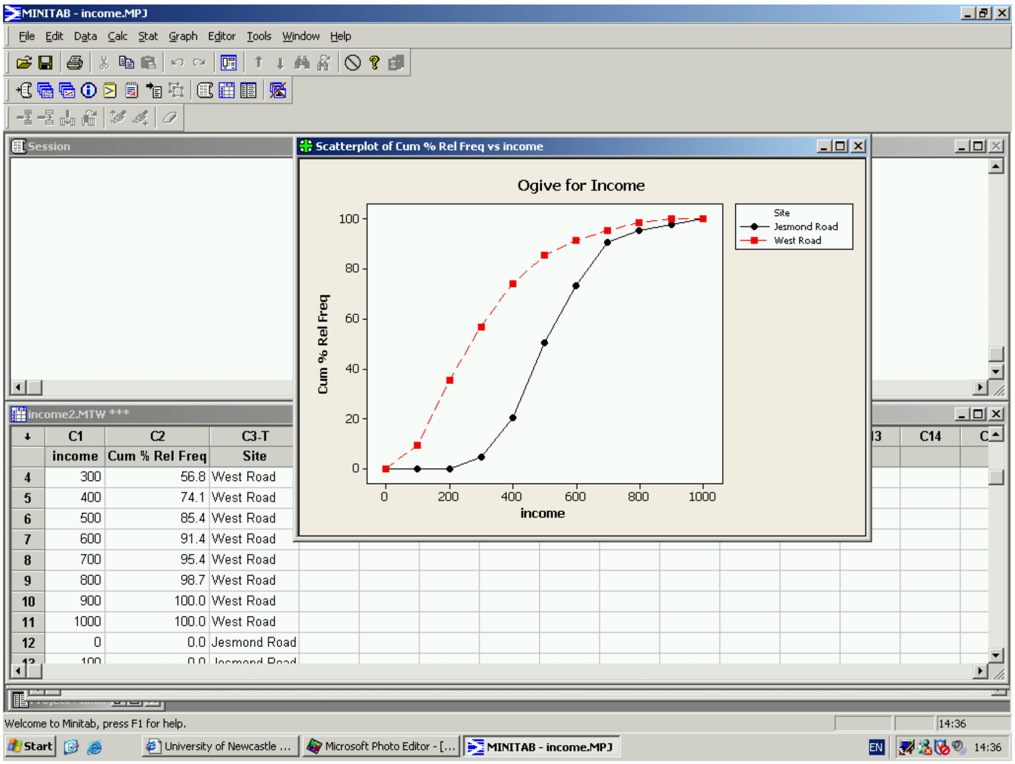
1. Find x on the x-axis and draw a line up from this value until you reach the ogive.

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 on the 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.


MINITAB - income.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Scatterplot of Cum % Rel Freq vs income
Ogive for Income
Cum % Rel Freq
0
20
40
60
80
100
0
200
400
600
800
1000
income
Site
Jesmond Road
West Road
income2.MTW ***
C14
Welcome to Minitab, press F1 for help.
14:36
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - income.MPJ
EN
14:36

↓	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
13	100	0.0	Jesmond Road

It clearly shows the ogive for Jesmond shifted to the left of that for the 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 between 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.

3.5 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 hte proportions of a whole. They are best used where there are only a handful of categories to display.

A pie chart consists of a circle divided into segments, one segment for each category. The size of each segment is determined by the frequency of the category and measured by the angle of the segment. As the total number of degrees in a circle is 360, the angle given to a segment is $360°$ times the fraction of the data in the category, that is

$$\text{angle} = \frac{\text{Number in category}}{\text{Total number in sample}(n)} \times 360.$$

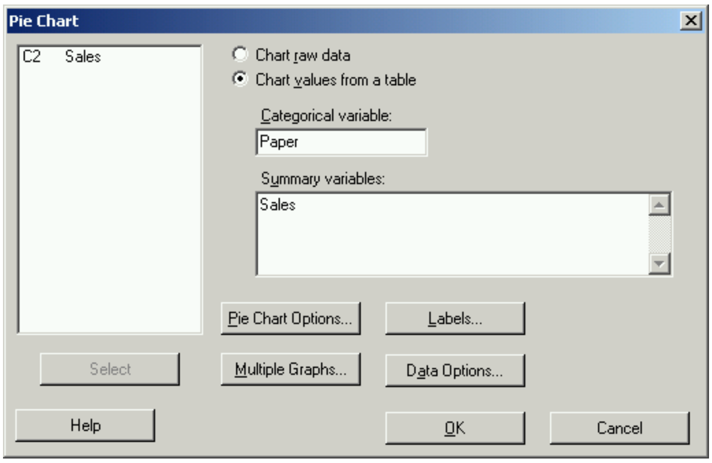
Consider again the data on newspaper sales to 650 students.

Paper	**Frequency**	**Degrees**
The Times	140	77.5
The Sun	200	110.8
The Sport	50	27.7
The Guardian	120	66.5
The Financial Times	20	11.1
The Mirror	80	44.3
The Daily Mail	10	5.5
The Independent	30	16.6
Totals	650	360.0

The pie chart is constructed by first drawing a circle and then dividing it up with segments with angles calculated using this formula.

In **MINITAB**, a pie chart for these data would be obtained as follows:

1. Enter data into worksheet, with category name in column C1 and frequencies in column C2.
2. `Graph > Pie Chart...`
3. Check the button for `Chart values from a table`
4. Enter the `Category` column under `Categorical variable:` and the `Frequency` column under `Summary variables:`



Pie Chart

C2 Sales

○ Chart raw data

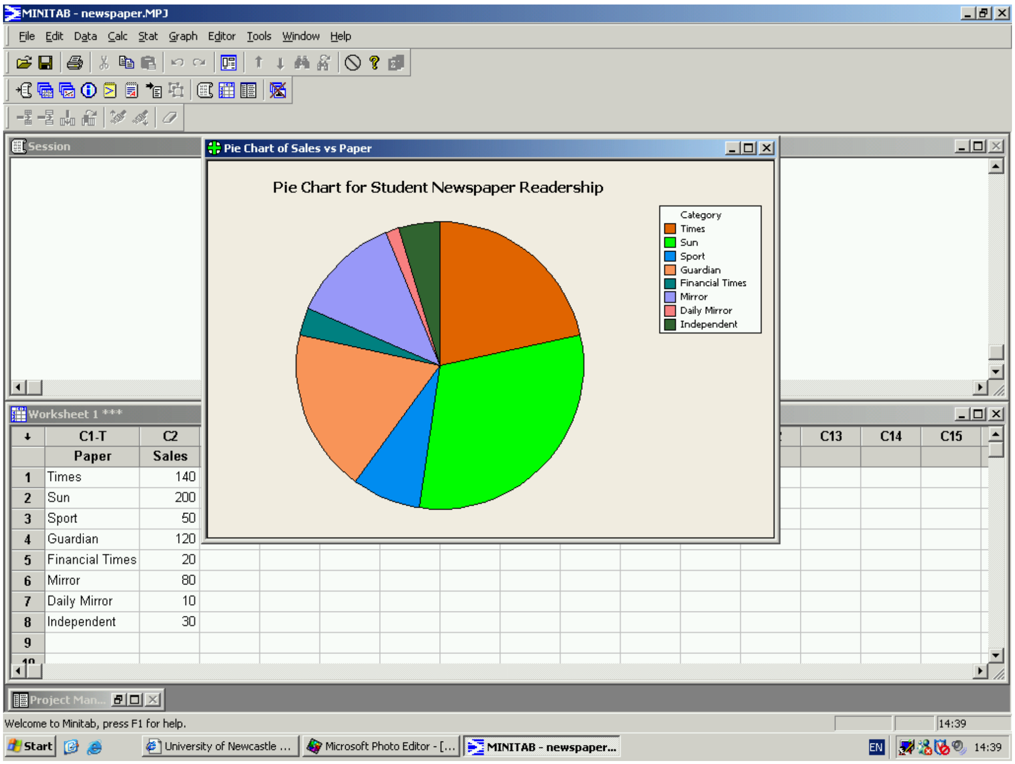
● Chart values from a table

Categorical variable: Paper

Summary variables: Sales

5. Add a title and click **OK**

This produces the following pie chart



↓	C1-T	C2
	Paper	Sales
1	Times	140
2	Sun	200
3	Sport	50
4	Guardian	120
5	Financial Times	20
6	Mirror	80
7	Daily Mirror	10
8	Independent	30
9		

Welcome to Minitab, press F1 for help.

It shows that The Sun, The Times and The Guardian are the most popular papers.

Note that the pie chart is a simple circle. Some computer software will draw “perspective” pie charts, pie charts with slices detached etc. It is best to avoid such gimmicks which merely obscure the information contained in the chart.

3.6 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 simply using time as the x-axis.

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.

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

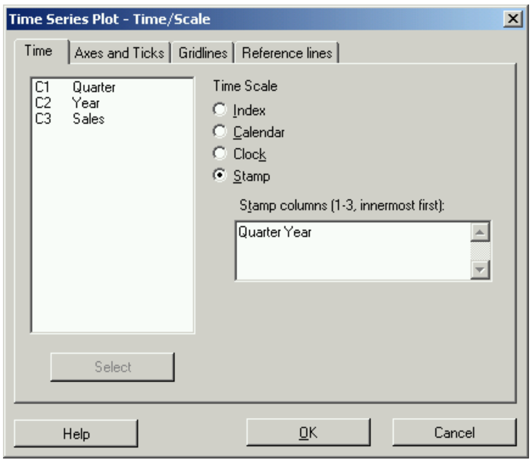
By hand, a time series plot is constructed as follows:

1. Draw the x-axis and label over the time scale.
2. Draw the y-axis and label with an appropriate scale.
3. Plot each point according to time and value.
4. Draw lines connecting all points.

In **MINITAB** the plot can be obtained using

1. Enter the data into a worksheet, with the `Quarter, Year` and `Sales` in columns C1, C2 and C3.

2. Click on `Graph` and select `Time Series Plot...`
3. Select the `Simple` graph format and click on `OK`.
4. Enter the `Sales` column in the `Series:` box.
5. Now click on `Time/Scale...`, check the `Stamp` button and enter the `Quarter` and `Year` columns under `Stamp columns`



Time Series Plot - Time/Scale

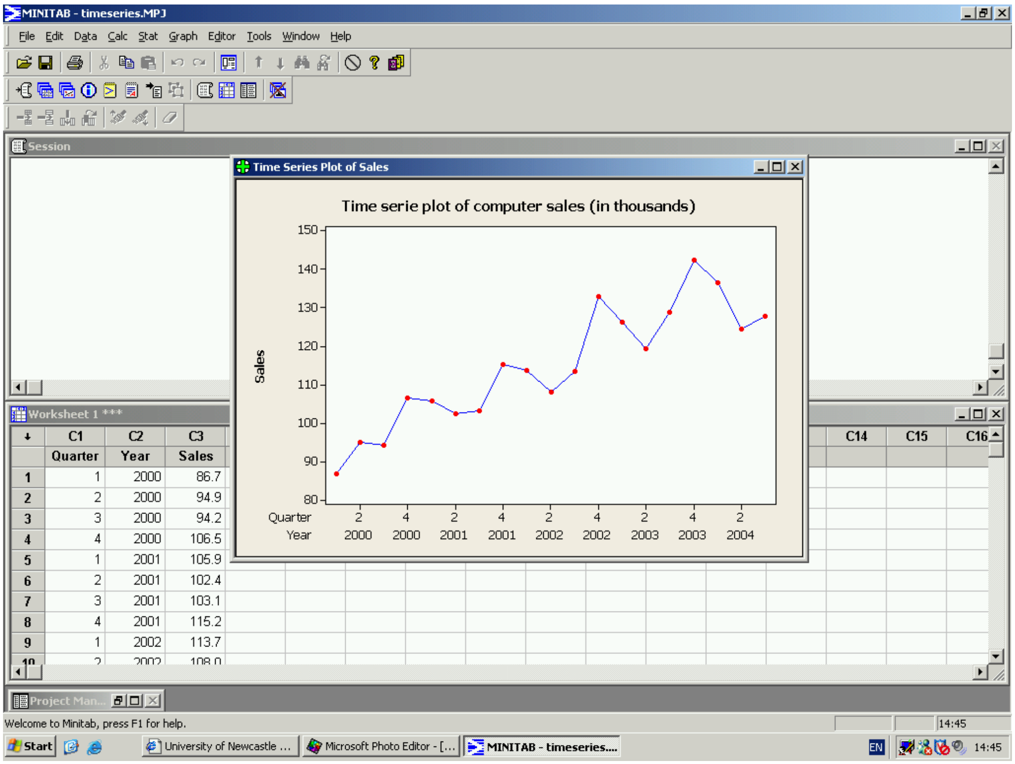
C1 Quarter
C2 Year
C3 Sales

Time Scale
○ Index
○ Calendar
○ Clock
● Stamp

Stamp columns (1-3, innermost first): Quarter Year

6. Click **OK**.
7. Add a title etc.
8. Click **OK**.

The time series plot is



MINITAB - timeseries.MPJ

File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Time Series Plot of Sales

Worksheet 1 ***

↓	C1	C2	C3
	Quarter	Year	Sales
1	1	2000	86.7
2	2	2000	94.9
3	3	2000	94.2
4	4	2000	106.5
5	1	2001	105.9
6	2	2001	102.4
7	3	2001	103.1
8	4	2001	115.2
9	1	2002	113.7
10	2	2002	108.0

Time serie plot of computer sales (in thousands)
Sales
150
140
130
120
110
100
90
80
Quarter 2 4 2 4 2 4 2 4 2
Year 2000 2000 2001 2001 2002 2002 2003 2003 2004
C14
C15
C16
Project Man...

Welcome to Minitab, press F1 for help.

14:45
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - timeseries....
EN
14:45

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.

3.7 Scatter Plots

The final type of graph we are going to look at is *scatter plots*. These 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.

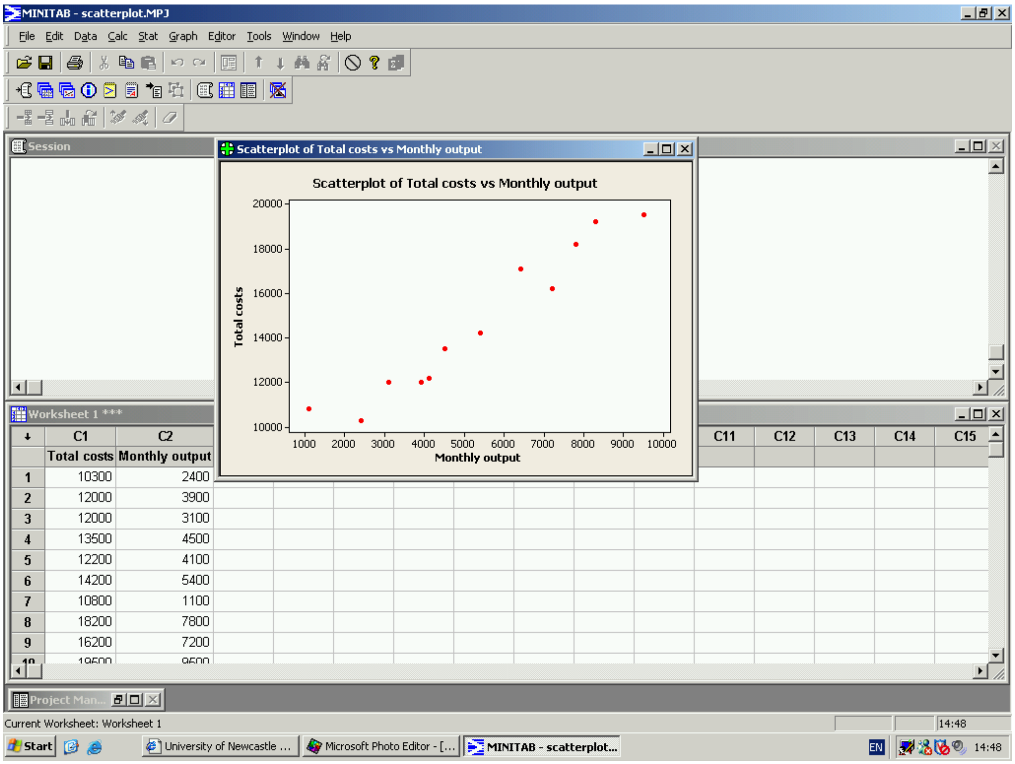
Consider the following data for monthly output and total costs at a factory.

Total costs (£)	**Monthly Output**
10300	2400
12000	3900
12000	3100
13500	4500
12200	4100
14200	5400
10800	1100
18200	7800
16200	7200
19500	9500
17100	6400
19200	8300

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.

1. The "response" variable is placed on the y-axis. Here we are trying to understand how total costs relate to monthly output and so the response variable is "total costs".
2. The variable that is used to try to explain the response variable (here, monthly output) is placed on the x-axis.
3. Plot the pairs of points on the graph.

This graph can be produced using **MINITAB**. (Select *Graph* then *Scatter Plot* then *Simple* and insert the required variables).


MINITAB - scatterplot.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Scatterplot of Total costs vs Monthly output
Scatterplot of Total costs vs Monthly output
Total costs
20000
18000
16000
14000
12000
10000
1000 2000 3000 4000 5000 6000 7000 8000 9000 10000
Monthly output
Worksheet 1 ***
C11 C12 C13 C14 C15
Project Man...
Current Worksheet: Worksheet 1
14:48
Start
University of Newcastle ...
Microsoft Photo Editor - [...
MINITAB - scatterplot...
EN
14:48

↓	C1	C2
	Total costs	Monthly output
1	10300	2400
2	12000	3900
3	12000	3100
4	13500	4500
5	12200	4100
6	14200	5400
7	10800	1100
8	18200	7800
9	16200	7200
10	19500	9500

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 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 more sophisticated techniques which we will develop later in the course.

3.8 Exercises 3

1. 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) Calculate the cumulative percentage relative frequency for before and after.

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

   (d) Plot the ogives for both on one graph.

   (e) Find the level of sales before and after that are reached on 25%, 50% and 75% of days.

2. The following table shows data for the monthly sales of a small department store and their monthly advertising expenditure.

Advertising Expenditure	**Monthly Sales**
52000	1200000
20500	650000
35000	870000
76000	1600000
65000	1200000
27000	850000
55000	1100000
39000	1000000
45000	1110000
27000	700000
38000	900000
52000	1150000

   Plot these data on an appropriate graph and comment on the relationship between advertising expenditure and monthly sales.

3. The data in table 3.1 give the amounts, in £, spent by 200 customers at a “Farmer’s Market” stall who bought at least one item. Use a histogram to display the data. The data are also available from the module Web page.

19.32	5.74	12.52	8.57	9.97	5.43	5.76	12.63	0.67	10.92
9.59	0.39	19.92	11.25	7.71	10.91	4.77	23.88	18.13	8.25
3.84	5.17	21.78	11.06	8.29	12.43	16.68	12.03	4.29	11.06
5.73	6.95	10.92	5.67	19.66	12.69	19.84	5.78	7.33	3.42
9.13	2.80	5.11	4.35	12.58	15.71	24.78	13.88	5.38	14.59
11.98	14.48	15.18	13.37	7.64	5.10	1.54	6.46	4.85	7.54
14.45	11.26	6.48	3.50	6.59	3.30	8.35	2.53	8.19	6.39
13.41	4.96	13.18	46.59	26.42	14.81	4.21	12.89	14.92	18.02
7.82	6.45	3.92	2.28	3.97	14.35	6.72	8.84	4.88	1.88
7.34	2.75	9.71	4.29	11.37	10.00	5.04	5.76	8.74	2.14
15.11	1.37	3.68	10.99	2.75	20.77	7.39	5.92	12.57	6.57
11.56	10.86	7.37	4.44	9.24	18.48	3.71	9.19	10.61	7.85
12.57	6.65	10.54	14.54	14.00	9.73	14.37	2.56	2.01	0.85
8.39	11.66	4.65	17.29	6.12	18.36	4.89	5.89	10.44	5.35
12.10	8.43	26.18	8.92	9.79	10.93	5.92	18.00	6.01	2.68
10.40	0.91	11.46	16.73	19.16	12.06	15.22	10.53	6.78	6.33
7.67	4.76	7.38	21.10	10.86	14.88	6.35	8.02	5.29	1.16
19.93	3.38	4.08	5.88	5.32	9.41	29.92	17.19	11.72	10.10
8.01	3.98	4.95	2.13	1.57	10.08	17.81	5.78	4.77	17.80
4.31	20.42	2.28	2.40	26.99	9.17	2.86	14.58	36.25	20.96

Table 3.1: Amounts spent at a farmer’s market stall

Industry	UK	Ireland
Agriculture	2.7	23.2
Mining	1.4	1.0
Manufacturing	30.2	20.7
Power supplies	1.4	1.3
Construction	6.9	7.5
Service Industries	16.9	16.8
Finance	5.7	2.8
Social and personal services	28.3	20.8
Transport and communications	6.4	6.1

Table 3.2: Percentages employed in different industries

Month	1980	1981	1982	1983	1984	1985
January	1998	1924	1969	2149	2319	2137
February	1968	1959	2044	2200	2352	2130
March	1937	1889	2100	2294	2476	2154
April	1827	1819	2103	2146	2296	1831
May	2027	1824	2110	2241	2400	1899
June	2286	1979	2375	2369	3126	2117
July	2484	1919	2030	2251	2304	2266
August	2266	1845	1744	2126	2190	2176
September	2107	1801	1699	2000	2121	2089
October	1690	1799	1591	1759	2032	1817
November	1808	1952	1770	1947	2161	2162
December	1927	1956	1950	2135	2289	2267

Table 3.3: Jeans sales in the UK

4. Table 3.2 shows the percentages employed in different industries in the UK and Ireland in the late 1970s. Use pie charts to compare the two countries' proportions.

5. Table 3.3 shows the estimated monthly sales of pairs of jeans, in 1000s, in the UK over six years. Use these data to make a time series plot. The data are also available from the module Web page.

Chapter 4

Numerical summaries for data

4.1 Introduction

So far we have only considered graphical methods for presenting data. These are always useful starting points. For many purposes, though, we might also require numerical methods for summarising data.

4.2 Mathematical notation

Before we can talk more about numerical techniques we first need to define some basic notation. This is to allow us the 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 letter to represent sample data and use 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 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 data 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 $sigma$. 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.

Two other mathematical basics need to be reintroduced. First, the use of powers is important in many statistical formulae. We all know that, for example, the square of three means raising 3 to the power 2, i.e. $3^2 = 3 \times 3 = 9$. This can be generalised to x^k, which means multiplying x by itself k times.

The other important idea is the use of brackets. Brackets are used to impose an ordering on the way operations are carried out. The operation inside the bracket is carried out before the one outside. Consider the following three cases:

$$\begin{aligned} 3 + 4^2 &= 19 \\ 3^2 + 4^2 &= 25 \\ (3 + 4)^2 &= 49. \end{aligned}$$

In the first case, we simply square 4 and then add this to 3. In the second case, we square both numbers and then add them together, while in the third case, because of the brackets, we add the numbers together and then square the result. Each one of these seemingly similar formulae gives a very different result. If we consider the last two formulae in general terms we could represent the second as $\sum x^2$, that is, we raise all the xs to the power 2 and then add then together. The third equation can be represented as $\left(\sum x\right)^2$, that is, all the xs are summed together and then this sum raised to the power 2. This is an important distinction which we will use later.

4.3 Measures of Location

These are also referred to as measures of centrality or 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.

4.3.1 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 $\bar{x}$. 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 (**SD**) mode on a university approved calculator.

Sometimes we might not have the raw data but data are 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**
01/07/04	9	08/07/04	10
02/07/04	8	09/07/04	5
03/07/04	6	10/07/04	8
04/07/04	7	11/07/04	4
05/07/04	7	12/07/04	6
06/07/04	10	13/07/04	8
07/07/04	11	14/07/04	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

$$\bar{x} = \frac{4 \times 1 + 5 \times 1 + 6 \times 2 + \ldots + 11 \times 1}{14} = 7.71.$$

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

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

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).

4.3.2 The Median

The median is occasionally used instead of the mean, particularly when the data have an asymmetric profile (as indicated by a histogram). 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**.

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

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

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 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 ($n = 8$) and 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.

4.3.3 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. Recalling the newspaper example, the mode was The Sun, as it was the most popular paper.

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.

4.4 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*.

4.4.1 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.

4.4.2 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:

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

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 start to make comparisons between the ranges of two data sets, without the problems caused by outliers or uneven sample sizes.

4.4.3 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. Strickly 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 as simplified 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 **SD** 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

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

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

We have already calculated the sample mean as $\bar{x} = 9.73$. Now

$$\begin{aligned}\sum x^2 &= 8.4^2 + 8.7^2 + \cdots + 11.4^2 = 1904.38\\ n(\bar{x})^2 &= 1893.458\end{aligned}$$

and so the sample variance is

$$s^2 = \frac{1}{19}(1904.38 - 1893.458) = 0.57484$$

and the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{0.57484} = 0.75818.$$

To ensure you understand the formulae and notation it would be a good idea for you to work through the following example:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	x_i^2
1			
1			
2			
3			
5			
6			
6			
6			
7			
8			
Totals			

$$n =$$

$$\sum x =$$

$$\overline{x} = \frac{\sum x}{n} =$$

$$\sum (x - \bar{x})^2 =$$

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 =$$

$$\sum x^2 =$$

$$s^2 = \frac{1}{n-1} \left\{ \sum x^2 - n \left(\bar{x} \right)^2 \right\} =$$

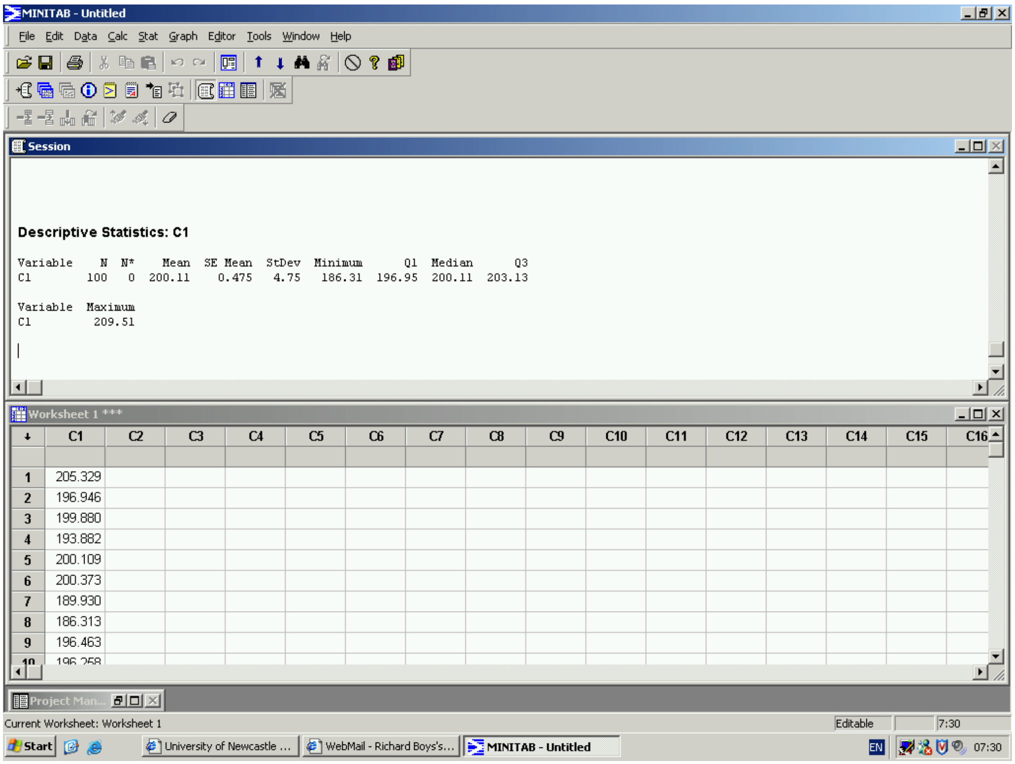
$$s = \sqrt{s^2} =$$

4.5 Summary statistics in MINITAB

MINITAB can be used to calculate many of the basic numerical summary statistics described so far. These summaries for data in a selected column can be obtained using the commands

```
Stats > Basic Statistics > Display Descriptive Statistics
```

The results are output in the session window as follows:



MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Worksheet 1 ***
Project Man...
Current Worksheet: Worksheet 1
Editable
7:30
Start
University of Newcastle ...
WebMail - Richard Boys's...
MINITAB - Untitled
EN
07:30

Descriptive Statistics: C1

```
Variable    N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
C1        100   0  200.11    0.475   4.75   186.31  196.95  200.11  203.13

Variable  Maximum
C1         209.51
```

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
1	205.329															
2	196.946															
3	199.880															
4	193.882															
5	200.109															
6	200.373															
7	189.930															
8	186.313															
9	196.463															
10	196.258															

4.6 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". A box and whisker plot for data with summary statistics

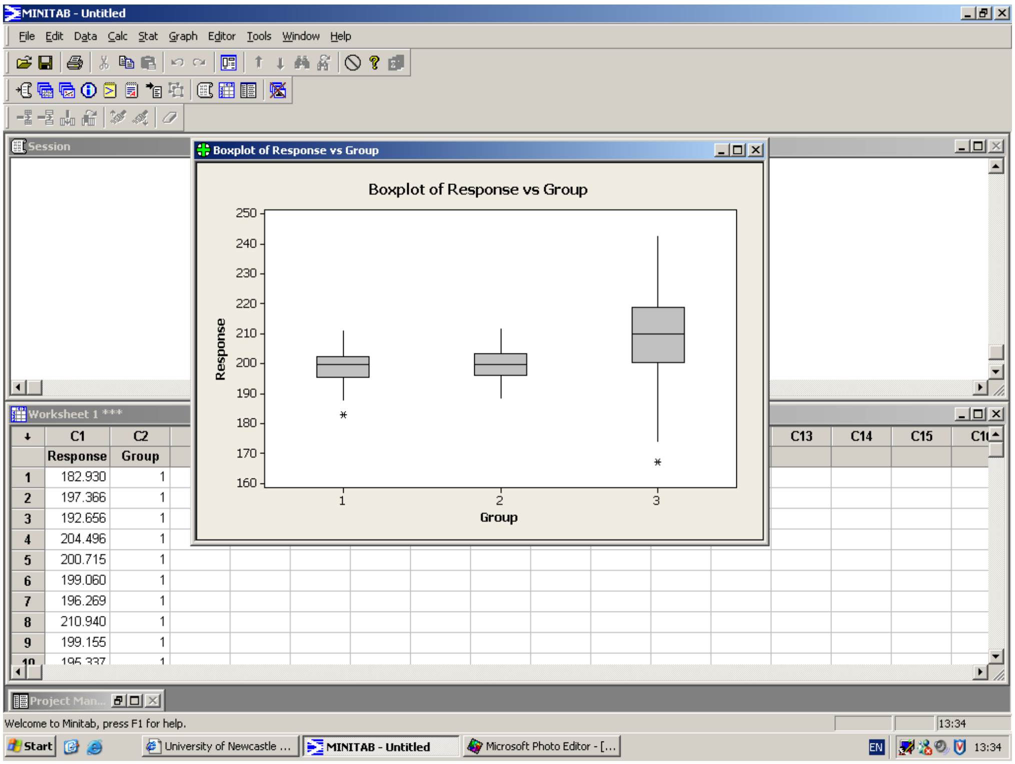
Minimum	$min = 10$
Lower quartile	$Q1 = 40$
Median	$Q2 = 43$
Upper quartile	$Q3 = 45$
Maximum	$max = 50$

would look like

MINITAB will produce box and whisker plots using the following commands.

1. Enter the data into the worksheet, say column C1
2. `Graph > Boxplot...` and select the `Simple` graph format
3. Next enter the column containing the data under `Graph variables:`
4. Add a title using `Labels...`
5. Click on **OK**.

If the data have subgroups, such as results from three different surveys, then box and whisker plots of the sample data can be plotted by group by first entering the group variable into the worksheet, say as column C2, and then selecting the `With Groups` graph format. The group variable is then entered into the subsequent dialogue box under `Categorical variables for grouping`. Displaying group structure is one of the main uses of box and whisker plots. For example



MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Session
Boxplot of Response vs Group
Boxplot of Response vs Group
250
240
230
220
210
200
190
180
170
160
Response
1
2
3
Group
C13
C14
C15
C1
Project Man...
13:34
Start
University of Newcastle ...
MINITAB - Untitled
Microsoft Photo Editor - [...
EN
13:34

Worksheet 1 ***

↓	C1	C2
	Response	Group
1	182.930	1
2	197.366	1
3	192.656	1
4	204.496	1
5	200.715	1
6	199.060	1
7	196.269	1
8	210.940	1
9	199.155	1
10	195.337	1

Welcome to Minitab, press F1 for help.

clearly shows that although there is overlap between the three sets of data, the first and second datasets contain roughly similar responses and that these are quite different from those in the third set. Note that the asterisks (*) at the ends of the whiskers is the way MINITAB highlights outlying values.

4.7 Exercises 4

Recall the data from Exercise 2 on the weight (in kg) of 50 sacks of potatoes leaving a farm shop.

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

1. Calculate the mean of the data.
2. Put the data in a grouped frequency table.
3. Estimate the sample mean from the grouped frequency table.
4. Calculate the median of the data.
5. Find the modal class.
6. Calculate the range of the data.
7. Calculate the inter-quartile range.
8. Calculate the sample standard deviation.
9. Draw a box and whisker plot for these data and comment on it.

Chapter 5

Introduction to Probability

5.1 Introduction

Probability is the language we use to model uncertainty. We all intuitively understand that few things in life are certain. There is usually an element of uncertainty or randomness around outcomes of our choices. In business this uncertainty can make all the difference between a good investment and a poor one. Hence an understanding of probability and how we might incorporate this into our decision making processes is important. In this chapter, we look at the logical basis for how we might express a probability and some basic rules that probabilities should follow. In the next chapter, we look at how we can use probabilities to aid decision making.

5.1.1 Definitions

We often use the letter P to represent a probability. For example, $P(Rain)$ would be the probability of the event of it raining.

Experiment An experiment is an activity where we do not know for certain what will happen but we will observe what happens. For example:

- We will ask someone whether or not they have used our product.
- We will observe the temperature at mid day tomorrow.
- We will toss a coin and observe whether it shows "heads" or "tails".

Outcome An outcome, or *elementary event*, is one of the possible things that can happen. For example, suppose that we are interested in the (UK) shoe size of the next customer to come into a shoe shop. Possible outcomes include "eight", "twelve", "nine and a half" and so on. In any experiment, one and only one outcome occurs.

Sample space The sample space is the set of all possible outcomes. For example it could be the set of all shoe sizes.

Event An even is a set of outcomes. For example "the shoe size of the next customer is less than 9" is an event. It is made of of all of the outcomes where the shoe size is less than 9. Of course an event might contain just one outcome.

Probabilities are usually expressed in terms of fractions or decimal numbers or percentages. Therefore we could express the probability of it raining today as

$$P(Rain) = \frac{1}{20} = 0.05 = 5\%.$$

All probabilities are measured on a scale ranging from zero to one. The probabilities of most events lie strictly between zero and one as an event with probability zero is an impossible event and one with probability one is a certain event.

The collection of all possible outcomes, that is the sample sapce, has a probability of 1. For example, if an event consists of only two outcomes *success* or *failure* then the probability of either a *success* or a *failure* is 1. That is $P(success$ or $failure) = 1$.

Two events are said to be *mutually exclusive* if both can not occur simultaneously. In the example above, the outcomes *success* and a *failure* are mutually exclusive.

Two events are said to be *independent* if the occurence of one does not affect the probability of the second occurring. For example, if you toss a coin and look out of the window, it would be reasonable to suppose that the events "get heads" and "it is raining" would be independent. However, not all events are independent. For example, if you go into the Students' Union Building and pick a student at random, then the events "the student is female" and "the student is studying engineering" are not independent since there is a greater proportion of male students on engineering courses than on other courses at the University (and this probably applies to those students found in the Union).

5.2 How do we measure Probability?

There are three main ways in which we can measure probability. All three obey the basic rules described above. Different people argue in favour of the different views of probability and some will argue that each kind has its uses depending on the circumstances.

5.2.1 Classical

If all possible outcomes are "equally likely" then we can adopt the *classical* approach to measuring probability. For example if we tossed a fair coin, there are only two possible outcomes, a head or a tail both of which are equally likely and hence

$$P(Head) = \frac{1}{2} \quad \text{and} \quad P(Tail) = \frac{1}{2}.$$

The underlying idea behind this view of probability is *symmetry*. In this example, there is no reason to think that the outcome *Head* and the outcome *Tail* have different probabilities and so they should have the same probability. Since there are two outcomes and one of them must occur, both outcomes must have probability 1/2.

Another commonly used example is rolling dice. There are six possible outcomes (1,2,3,4,5,6) when a die is rolled and each of them should have an equal chance of occuring. Hence the $P(1) = \frac{1}{6}$, $P(2) = \frac{1}{6}, \ldots$.

Other calculations can be made such as $P(\text{Even Number}) = \frac{3}{6} = \frac{1}{2}$. This follows from the formula

$$P(\text{Event}) = \frac{\text{Total number of outcomes in which event occurs}}{\text{Total number of possible outcomes}}.$$

Note that this formula only works when all possible events are equally likely – not a practical assumption for most real life situations.

5.2.2 Frequentist

When the outcomes of an experiment are not equally likely, we can conduct experiments to give us some idea of how likely the different outcomes are. For example, suppose we were interested in measuring the probability of producing a defective item in a manufacturing process. This probability could be measured by monitoring the process over a reasonably long period of time and calculating the proportion of defective items. What constitutes a reasonably long period of time is, of course, a difficult question to answer. In a more simple case, if we did not believe that a coin was fair, we could toss the coin a large number of times and see how often we obtained a head. In both cases we perform the same experiment a large number of times and observe the outcome. This is the basis of the frequentist view. By conducting experiments the probability of an event can easily be estimated using the following formula:

$$P(\text{Event}) = \frac{\text{Number of times an event occurs}}{\text{Total number of times experiment done}}.$$

The larger the experiment, the closer this probability is to the “true” probability. The frequentist view of probability regards probability as the long run relative frequency (or proportion). So, in the defects example, the “true” probability of getting a defective item is the proportion obtained in a very large experiment (strictly an *infinitely* long sequence of trials).

In the frequentist view, probability is a property of nature and, since, in practice, we can not conduct infinite sequences of trials, in many cases we never really know the “true” values of probabilities. We also have to be able to imagine a long sequence of “identical” trials. This does not seem to be appropriate for “one-off” experiments like the launch of a new product. For these reasons (and others) some people prefer the *subjective* or *Bayesian* view of probability.

5.2.3 Subjective/Bayesian

We are probably all intuitively familiar with this method of assigning probabilities. When we board an aeroplane, we judge the probability of it crashing to be sufficiently small that we are happy to undertake the journey. Similarly, the odds given by bookmakers on a horse race reflect people's beliefs about which horse will win. This probability does not fit within the frequentist definition as the race cannot be run a large number of times.

One potential difficulty with using subjective probabilities is that it *is* subjective. So the probabilities which two people assign to the same event can be different. This becomes important if these probabilities are to be used in decision making. For example, if you were deciding whether to launch a new product and two people had very different ideas about how likely success or failure of this product was, then the decision to go ahead could be controversial. If both individuals assessed the probability of success to be 0.8 then the decision to go ahead could easily be based on this belief. However, if one said 0.8 and the other 0.3, then the decision is not straightforward. We would need a way to reconcile these different positions.

Subjective probability is still subject to the same rules as the other forms of probability, namely that all probabilities should be positive and that the probability of all outcomes should sum to one. Therefore, if you assess $P(Success) = 0.8$ then you should also assess $P(Failure) = 0.2$.

5.3 Laws of Probability

5.3.1 Multiplication Law

The probability of two *independent* events E_1 and E_2 both occurring can be written as

$$P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2).$$

For example, if the probability of throwing a six followed by another six on two rolls of a die is calculated as follows. The outcomes of the two rolls of the die are independent. Let E_1 denote a six on the first roll and E_2 a six on the second roll. Then

$$P(\text{two sixes}) = P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

This method of calculating probabilities extends to when there are many *independent* events

$$P(E_1 \text{ and } E_2 \text{ and } \cdots \text{ and } E_n) = P(E_1) \times P(E_2) \times \cdots \times P(E_n).$$

(There is a more complicated rule for multiplying probabilities when the events are not independent).

5.3.2 Addition Law

The multiplication law is concerned with the probability of two or more independent events occurring. The *addition law* describes the probability of any of two or more events occurring. The addition law for two events E_1 and E_2 is

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2).$$

This describes the probability of *either* event E_1 *or* event E_2 happening.

Consider the following information: 50 percent of families in a certain city subscribe to the morning newspaper, 65 percent subscribe to the afternoon newspaper, and 30 percent of the families subscribe to both newspapers. What proportion of families subscribe to at least one newspaper?

We are told $P(\text{Morning}) = 0.5$, $P(\text{Afternoon}) = 0.65$ and $P(\text{Morning and Afternoon}) = 0.3$. Therefore

$$\begin{aligned} P(\text{at least one paper}) &= P(\text{Morning or Afternoon}) \\ &= P(\text{Morning}) + P(\text{Afternoon}) - P(\text{Morning and Afternoon}) \\ &= 0.5 + 0.65 - 0.3 \\ &= 0.85. \end{aligned}$$

So 85% of of the city subscribe to at least one of the newspapers.

A more basic version of the rule works where events are mutually exclusive: if events E_1 and E_2 are mutually exclusive then

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2).$$

This simplification occurs because when two events are mutually exclusive they cannot happen together and so $P(E_1 \text{ and } E_2) = 0$.

These two laws are the basis of more complicated problem solving we will see later.

5.3.3 Example

A building has three rooms. Each room has two separate electric lights. There are thus six electric lights altogether. After a certain time there is a probability of 0.1 that a given light will have failed and all light are independent of all other lights. Find the probability that, after this time, there is at least one room in which both lights have failed.

Solution

For a given light, the probability that it has failed is 0.1.

For a given room, the probability that *both* lights have failed is

$$0.1 \times 0.1 = 0.01.$$

For a given room, the probability that it is not true that both lights have failed, that is the probability that at least one of the two lights is working, is

$$1 - 0.01 = 0.99.$$

The probability that at least one light is working in every one of the three rooms (that is, in Room A *and* in Room B *and* in Room C) is

$$0.99 \times 0.99 \times 0.99 = 0.99^3 = 0.970299.$$

The probability that there is at least one room in which both lights have failed (that is the probability that it is not true that there is at least one light working in every room) is

$$1 - 0.970299 = 0.029701$$

or just under 3%.

N.B. We also can obtain this answer by extending the addition law to cover three events. Let $A,\ B,\ C$ be the events "both lights have failed in Room A," " both lights have failed in Room B," "both lights have failed in Room C." We can show that

$$\begin{aligned} P(A \text{ or } B \text{ or } C) &= P(A) + P(B) + P(C) - P(A \text{ and } B) - P(A \text{ and } B) - P(B \text{ and } C) \\ &\quad + P(A \text{ and } B \text{ and } C) \end{aligned}$$

where "A or Bor C" means "at least one of $A,\ B,\ C$" and "A and $B$and C" means "all three of $A,\ B,\ C$". So, the required probability is

$$\begin{aligned} P(A \text{ or } B\text{or } C) &= 0.01 + 0.01 + 0.01 - (0.01 \times 0.01) - (0.01 \times 0.01) - (0.01 \times 0.01) \\ &\quad + (0.01 \times 0.01 \times 0.01) \\ &= 3 \times 0.01 - 3 \times 0.0001 + 0.000001 \\ &= 0.03 - 0.0003 + 0.000001 = 0.029701. \end{aligned}$$

5.4 Exercises 5

1. A company manufactures a device which contains three components A, B and C. The device fails if any of these components fail and the company offers to its customers a full money-back warranty if the product fails within one year. The company has assessed the probabilities of each of the components lasting at least a year as 0.98, 0.99 and 0.95 for A, B and C respectively. The three components within a single device are considered to be independent. Consider a single device chosen at random. Calculate the probability that

   (a) all three components will last for at least a year;

   (b) the device will be returned for a refund.

2. The following data refer to a class of 18 students. Suppose that we will choose one student at random from this class.

Student Number	Sex	Height (m)	Weight (kg)	Shoe Size	Student Number	Sex	Height (m)	Weight (kg)	Shoe Size
1	M	1.91	70	11.0	10	M	1.78	76	8.5
2	F	1.73	89	6.5	11	M	1.88	64	9.0
3	M	1.73	73	7.0	12	M	1.88	83	9.0
4	M	1.63	54	8.0	13	M	1.70	55	8.0
5	F	1.73	58	6.5	14	M	1.76	57	8.0
6	M	1.70	60	8.0	15	M	1.78	60	8.0
7	M	1.82	76	10.0	16	F	1.52	45	3.5
8	M	1.67	54	7.5	17	M	1.80	67	7.5
9	F	1.55	47	4.0	18	M	1.92	83	12.0

Find the probabilities for the following events.

(a) The student is female.

(b) The student’s weight is greater than 70kg.,

(c) The student’s weight is greater than 70kg. and the student’s shoe-size is greater than 8,

(d) The student’s weight is greater than 70kg. or the student’s shoe-size is greater than 8.

Chapter 6

Decision Making using Probability

In this chapter, we look at more complicated notions of probability and how we can use probability in order to aid in management decision making.

6.1 Conditional Probability

So far we have only considered probabilities of single events or of several independent events, like two rolls of a die. However in reality many events are related. For example the probability of it raining in 5 minutes time is dependent on whether or not it is raining now.

We need a mathematical notation to capture how the probability of one event depends on other events taking place. We do this as follows. Consider two events A and B. We write

$$P(A|B)$$

for the probability of A given B has already happened. We describe $P(A|B)$ as the conditional probability of A given B. For example, the probability of it raining in 5 minutes time given that it is raining now would be

$$P(\text{Rain in 5 minutes}|\text{Raining now}).$$

Utility companies need to be able to forecast periods of high demand. They describe their forecasts in terms of probabilities. Gas and electricity suppliers might relate them to air temperature. For example,

$$\begin{aligned}
&P(\text{High demand}|\text{air temperature is below normal}) = 0.6\\
&P(\text{High demand}|\text{air temperature is normal}) = 0.2\\
&P(\text{High demand}|\text{air temperature is above normal}) = 0.05.
\end{aligned}$$

We can calculate these conditional probabilities using the formula

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)},$$

that is, in terms of the probability of both events occurring, $P(A$ and $B)$, and the probability of the event that has already taken place, $P(B)$.

To see how this formula works, let's consider a simple example based on the class of students in Exercises 5.

Student Number	Sex	Height (m)	Weight (kg)	Shoe Size	Student Number	Sex	Height (m)	Weight (kg)	Shoe Size
1	M	1.91	70	11.0	10	M	1.78	76	8.5
2	F	1.73	89	6.5	11	M	1.88	64	9.0
3	M	1.73	73	7.0	12	M	1.88	83	9.0
4	M	1.63	54	8.0	13	M	1.70	55	8.0
5	F	1.73	58	6.5	14	M	1.76	57	8.0
6	M	1.70	60	8.0	15	M	1.78	60	8.0
7	M	1.82	76	10.0	16	F	1.52	45	3.5
8	M	1.67	54	7.5	17	M	1.80	67	7.5
9	F	1.55	47	4.0	18	M	1.92	83	12.0

Suppose we want the probability that a student chosen at random from this class will be female given that the student's shoe size is less than 8. We could simply find the proportion of students with shoe sizes less than 8 who are female. There are 7 students with shoe sizes less than 8 and 4 of these are female. So

$$P(\text{Female}|\text{Shoe size less than 8}) = \frac{4}{7}.$$

This probability can also be calculated using the above formula as follows:

$$\begin{aligned} P(\text{Shoe size less than 8}) &= \frac{7}{18} \\ P(\text{Shoe size less than 8 and female}) &= \frac{4}{18} \end{aligned}$$

and so

$$P(\text{Female}|\text{Shoe size less than 8}) = \frac{P(\text{Shoe size less than 8 and female})}{P(\text{Shoe size less than 8})} = \frac{4/18}{7/18} = \frac{4}{7}.$$

6.2 Multiplication of probabilities

We saw in Chapter 5 that, if two events A and B are independent, then $P(A \text{ and} B) = P(A)P(B)$. Now we know that

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)},$$

we can easily see that

$$P(A \text{ and } B) = P(B)P(A|B).$$

Of course it is also true that $P(A \text{ and } B) = P(A)P(B|A)$.

For example, consider a student chosen at random from the example class. Let F be the event “the student is female” and S be the event “the student’s weight is less than 60kg.” Then the probability that the student is female and has a weight less than 60kg is

$$\begin{aligned} P(F \text{and} S) &= P(S)P(F|S) = \frac{7}{18} \times \frac{3}{7} = \frac{3}{18} \\ &= P(F)P(S|F) = \frac{4}{18} \times \frac{3}{4} = \frac{3}{18} \end{aligned}$$

Notice that, if M is the event “the student is male,” then $P(S|M) = 4/14 = 0.286$ and this is not equal to $P(S|F) = 3/4 = 0.75$. So the probability of the student having a weight less than 60kg depends on the student’s sex, that is whether the student is female or male. The events S and F are not independent. Similarly $P(F|S) = 3/7 = 0.429$ while $P(F|L) = 1/11 = 0.091,$ where L is the event “the students’s weight is not less than 60kg.” So, knowing whether or not a student’s weight is less than 60kg gives us information about whether the student is likely to be male or female.

Let $\bar{B}$ be the event “not B.” So, for example $\bar{F} = M$. Then we say that tow events A and B are independent if $P(A|B) = P(A|\bar{B}) = P(A)$. It is easy to show that this is equivalent to $P(B|A) = P(B|\bar{A}) = P(B)$. If A and B are independent then $P(A \text{and} B) = P(A)P(B).$

For example, consider the following probabilities for customers at a cafe who can choose eiether icecream or treacle sponge and custard.

	Icecream	Treacle sponge
Male	0.250	0.150
Female	0.375	0.225

We see that $P(\text{male}) = 0.250+0.150 = 0.4$ and $P(\text{female}) = 0.375+0.225 = 0.6 = 1-P(\text{male}).$ Now

$$P(\text{Icecream}|\text{Male}) = \frac{0.250}{0.4} = 0.625$$

and

$$P(\text{Icecream}|\text{Female}) = \frac{0.375}{0.6} = 0.625$$

so Icecream and Male are independent events. In fact the variables Sex and Dessert-choice are independent in this example. So the probability that a customer is male and chooses icecream is just $P(\text{Male})P(\text{Icecream}) = 0.4 \times 0.625 = 0.25$. (The probability of icecream is just $0.250 + 0.375 = 0.625$).

Another example relates to the age and sex distribution of purchasers of CD singles at an outlet:

	< 30	$30 - 50$	$50+$
Male	0.275	0.125	0.025
Female	0.325	0.175	0.075

From this table, we can calculate

$$\begin{aligned} P(\text{Male}) &= P(\text{Male and } < 30) + P(\text{Male and } 30 - 50) + P(\text{Male and } 50+) \\ &= 0.275 + 0.125 + 0.025 = 0.425 \end{aligned}$$

and

$$\begin{aligned} P(\text{Female}) &= P(\text{Female and } < 30) + P(\text{Female and } 30-50) + P(\text{Female and } 50+) \\ &= 0.325 + 0.175 + 0.075 = 0.575. \end{aligned}$$

Also, the age distribution of the customers is

$$\begin{aligned} P(<30) &= P(\text{Male and } < 30) + P(\text{Female and } < 30) = 0.275 + 0.325 = 0.6 \\ P(30-50) &= P(\text{Male and } 30-50) + P(\text{Female and } 30-50) = 0.125 + 0.175 = 0.3 \\ P(50+) &= P(\text{Male and } 50+) + P(\text{Female and } 50+) = 0.025 + 0.075 = 0.1. \end{aligned}$$

Using this information we can calculate various probabilities such as:

$$\begin{aligned} P(\text{Male}|30-50) &= \frac{P(\text{Male and } 30-50)}{P(30-50)} = \frac{0.125}{0.3} = 0.4167 \\ P(\text{Female}|30-50) &= 1 - P(\text{Male}|30-50) = 1 - 0.4167 = 0.5833 \end{aligned}$$

and

$$\begin{aligned} P(<30|\text{Male}) &= \frac{P(\text{Male and } < 30)}{P(\mathrm{Male})} = \frac{0.275}{0.425} = 0.6471 \\ P(30-50|\text{Male}) &= \frac{P(\text{Male and } 30-50)}{P(\mathrm{Male})} = \frac{0.125}{0.425} = 0.2941 \\ P(50+|\text{Male}) &= 1 - P(<30|\text{Male}) - P(30-50|\text{Male}) = 1 - 0.6471 - 0.2941 = 0.0588. \end{aligned}$$

6.3 Tree Diagrams

Tree diagrams or probability trees are simple clear ways of presenting probabilistic information. Let us first consider a simple example in which a die is rolled twice. Suppose we are interested in the probability that we score a six on both rolls. This probability can be calculated as

$$\begin{aligned} P(\text{Six and Six}) &= P(\text{Six on 1st throw}) \times P(\text{Six on 2nd throw}|\text{Six on 1st throw}) \\ &= \frac{1}{6} \times \frac{1}{6} \\ &= \frac{1}{36}. \end{aligned}$$

This example can be represented as a tree diagram in which experiments are represented by circles (called *nodes*) and the outcomes of the experiments as *branches*. The branches are annotated by the probability of the particular outcome.

Here the probability of a six followed by a six is found by tracing the branch corresponding to this outcome through the tree. Note that the ends of the branches of the tree are usually known as *terminal nodes*.

Consider a more complicated example. A machine is used to produce components. Each time it produces a component there is a chance that the component will be defective. When the machine is working correctly the probability that a component is defective is 0.05. Sometimes, though, the machine requires adjustment and, when this is the case, the probability that a component is defective is 0.2. At the time in question there is a probability of 0.1 that the machine requires adjustment. Components produced by the machine are tested and either accepted or rejected. A component which is not defective is accepted with probability 0.97 and (falsely) rejected with probability 0.03. A defective component is (falsely) accepted with probability 0.15 and rejected with probability 0.85.

We can calculate various probabilities. For example:

$$\begin{aligned}
P(\text{accepted}) &= 0.82935 + 0.00675 + 0.07760 + 0.00300 = 0.9167 \\
P(\text{defective}) &= (0.9 \times 0.05) + (0.1 \times 0.2) = 0.045 + 0.02 = 0.065 \\
P(\text{defective and accepted}) &= 0.00675 + 0.00300 = 0.00975 \\
P(\text{accepted} \mid \text{defective}) &= \frac{0.00975}{0.065} = 0.15 \\
P(\text{defective} \mid \text{accepted}) &= \frac{0.00975}{0.9167} = 0.010636 \\
P(\text{machine OK and accepted}) &= 0.82935 + 0.00675 = 0.8361 \\
P(\text{machine OK} \mid \text{accepted}) &= \frac{0.8361}{0.9167} = 0.9121 \\
P(\text{machine OK and rejected}) &= 0.02565 + 0.03825 = 0.0639 \\
P(\text{rejected}) &= 1 - P(\text{accepted}) = 0.0833 \\
P(\text{machine OK} \mid \text{rejected}) &= \frac{0.0639}{0.0833} = 0.7671
\end{aligned}$$

6.4 Expected Monetary Value and Probability Trees

Probability trees can be used to see the effect of making particular decisions in the face of uncertainty. This is achieved by weighting the probability of different outcomes by their value. Often this value is financial. The *Expected Monetary Value (EMV)* of a single event is simply the probability of that event multiplied by the monetary value of that outcome. For example, if you would win £5 if you pulled an ace from a pack of cards, the EMV would be

$$EMV(Ace) = \frac{1}{13} \times 5 = 0.38.$$

In other words, if you repeated this bet a large number of times, overall you would come out an average of 38 pence better off per bet. Therefore you would want to pay no more than $38p$ for such a bet.

Consider another bet. When rolling a die, if it's a six you have to pay £5 but if it's any other number you receive £2.50. Would you take on this bet?

Probability	**Financial outcome**
$P(6) = 1/6$	-£5
$P(\text{Not a } 6) = 5/6$	£2.50

Therefore

$$\begin{aligned} EMV(\text{Six}) &= \frac{1}{6} \times -5.00 = -0.833 \\ EMV(\text{Not a Six}) &= \frac{5}{6} \times 2.50 = 2.0833 \end{aligned}$$

and hence the expected monetary value of the bet is

$$EMV(\text{Bet}) = -0.833 + 2.083 = 1.25.$$

Therefore, in the long run, this would be a bet to take on as it has a positive expected monetary value.

In general, the expected monetary value of a project or bet is given by the formula

$$EMV = \sum P(\text{Event}) \times \text{ Monetary value of Event}$$

where the sum is over all possible events. The EMV of a project can be used as a decision criterion for choosing between different projects and has applications in a large number of situations. This is illustrated by the following example.

A small company is trying to decide how to launch a new and innovative product. It could go for a direct approach, launching onto the whole of the domestic market through traditional distribution channels, or it could launch only on the internet. A third option exists where the product is licensed to a larger company through the payment of a licence fee irrespective of the success of the product. How should the company launch the product? The company has done some initial market research and the managing director believes the probability of the product being successful can be classed into three categories: high, medium or low. She thinks that these categories will occur with probabilities 0.2, 0.35 and 0.45 respectively and her thoughts on the likely profits (in £K) to be earned in each plan are

	High	Medium	Low
Direct	100	55	-25
Internet	46	25	15
Licence	20	20	20

The EMV of each plan can be calculated as follows:

$$\begin{aligned} EMV(\text{Direct}) &= 0.2 \times 100 + 0.35 \times 55 + 0.45 \times (-25) = \pounds 28\text{K} \\ EMV(\text{Internet}) &= 0.2 \times 46 + 0.35 \times 25 + 0.45 \times 15 = \pounds 24.7\text{K} \\ EMV(\text{Licence}) &= 0.2 \times 20 + 0.35 \times 20 + 0.45 \times 20 = \pounds 20\text{K}. \end{aligned}$$

On the basis of expected monetary value, the best choice is the Direct approach.

In this example we have to make a decision. When we include a decision in a probability tree we use a rectangular node, called a *decision node* to represent the decision. The diagram is then called a *decision tree*. There are no probabilities at a decision node but we evaluate the expected monetary values of the options. In a decision tree the first node is always a decision node. There may also be other decision nodes. If there is another decision node then we evaluate the options there and choose the best and the expected value of this option becomes the expected value of the branch leading to the decision node.

6.5 Exercises 6

1. A company has installed a new computer system and some employees are having difficulty logging on to the system. They have been given training and the problems which arose during training were recorded and their probabilities calculated as follows:

   - An employee has a probability of 0.9 of logging on successfully on the first attempt.
   - If the employee logs in successfully then the employee will also be successful on each later attempt with probability 0.9.
   - If the employee tries to log in and is not successful then the employee loses confidence and the probability of a successful log-in on later occasions drops to 0.5.

   Use a tree diagram to find the following probabilities:

   (a) An employee successfully logs on in each of the first three attempts.

   (b) An employee fails in the first attempt but is successful in the next two attempts.

   (c) An employee logs on successfully only once in three attempts.

   (d) An employee does not manage to log on successfully in three attempts.

2. The owner of a small business has the right to have a retail stall at a large festival to be held during the summer. She judges that this would either be a success or a failure and that the probability that it is a success is 0.4. If the stall was a success, the net income from it would be £90,000. If the stall was a failure, there would be a net loss of £30,000 from it. To help to make the decision, the owner could pay for market research. This would cost £5,000. The market research will either give a positive indication or a negative indication. The conditional probability that it gives a positive indication, given that the stall will actually be a success, is 0.75. The conditional probability that it gives a positive indication, given that the stall will actually be a failure, is $1/3$.

   The owner has various options:

   - Do nothing.
   - Go ahead without market research.
   - Pay for the market research.
   - Sell her right to a stall for £10,000.

   If she pays for the market reserach then, depending on the outcome, she can:

   - Do nothing more.
   - Go ahead.
   - Sell her right to a stall. If the market research gave a positive indication the price would be £35,000. If the market research gave a negative indication the price would be only £3,000.

   What should she do?

   This is a fairly complicated question so it is best to tackle it in stages.

(a) using a probability tree for the success or failure of the stall and the market research outcome, find the following probabilities.

i. The probability of a positive market research outcome.

ii. The probability of a negative market research outcome.

iii. The conditional probability of a successful stall given a positive market research outcome.

iv. The conditional probability of a failure given a positive market research outcome.

v. The conditional probability of a successful stall given a negative market research outcome.

vi. The conditional probability of a failure given a negative market research outcome.

(b) Represent the owner's decision problem using a tree diagram.

(c) Suppose that the owner has the market research done and that the outcome is positive. Evaluate the expected monetary values, under these circumstances, of the three options:

- Sell.
- Go ahead.
- Do nothing more.

and hence find what she should do if these circumstances arise.

(d) Suppose that the owner has the market research done and that the outcome is negative. Evaluate the expected monetary values, under these circumstances, of the three options:

- Sell.
- Go ahead.
- Do nothing more.

and hence find what she should do if these circumstances arise.

(e) Hence find the expected monetary value of the initial option of "Pay for the market research."

(f) Find the expected monetary values of the three other inital options:

- Do nothing.
- Go ahead without market research.
- Sell her right to a stall for £10,000.

(g) Determine the owner's best strategy.

Chapter 7

Discrete Probability Models

7.1 Introduction

The link between probability and statistics arises because, in order to see, for example, how strong the evidence is in some data, we need to consider the probabilities concerned with how we came to observe the data. In this chapter, we describe some standard probability models which are often used used with data from various sources such as market research. However, before we describe these in detail, we need to establish some ground rules for counting.

7.2 Permutations and Combinations

7.2.1 Numbers of sequences

Imagine that your cash point card has just been stolen. What is the probability of the thief guessing your 4 digit PIN in one go? To answer this question, we need to know how many different 4 digit PINs there are. We are also assuming that the thief chooses in such a way that all possibilities are equally likely. With this assumption the probability of a correct guess (in one go) is

$$\begin{aligned} P(\text{Guess correctly}) &= \frac{\text{number of correct PINs}}{\text{number of possible 4 digit PINs}} \\ &= \frac{1}{\text{number of possible 4 digit PINs}}. \end{aligned}$$

There is, of course, only one correct PIN. The number of possible 4 digit PINs is calculated as follows. There are 10 choices for the first digit, another 10 choices for the second digit, and so on. Therefore the number of possible choices is

$$10 \times 10 \times 10 \times 10 = 10,000.$$

So the probability of a correct guess is

$$\begin{aligned} P(\text{Guess correctly}) &= \tfrac{1}{10\times10\times10\times10} \\ &= \tfrac{1}{10{,}000} \\ &= 0.00001. \end{aligned}$$

7.2.2 Permutations

The calculation of the card-thief's correct guess of a PIN changes if the thief knows that your PIN uses 4 different digits. Now the number of possible PINs is smaller. To find this number we need to work out how many ways there are to arrange 4 digits out of a lsit of 10.

In more general terms, we need to know how many different ways there are of arranging r objects from a list of n objects. The best way of thinking about this to consider the choice of each item as a different experiment. The first experiment has n possible outcomes. The second experiment only has $n-1$ possible outcomes, as one object has already been selected. The third experiment has $n-2$ outcomes and so on until the rth experiment, which has $n-r+1$ possible outcomes. Therefore the number of possible selections is

$$n\times(n-1)\times(n-2)\times\cdots\times(n-r+1) = \frac{n\times(n-1)\times(n-2)\times\cdots\times3\times2\times1}{(n-r)\times(n-r-1)\times\cdots\times3\times2\times1} = \frac{n!}{(n-r)!}.$$

Here

$$n! = n(n-1)(n-2)(n-3)\times\cdots\times3\times2\times1$$

and is called n *factorial* - it can be found on many calculators. The formula

$$\frac{n!}{(n-r)!}$$

is a commonly encountered expression in counting calculations (combinatorics) and has its own notation. The number of ordered ways of selecting r objects from n is denoted ${}^n\mathrm{P}_r$, where

$${}^n\mathrm{P}_r = \frac{n!}{(n-r)!}.$$

We refer to ${}^n\mathrm{P}_r$ as the number of *permutations* of r out of n objects.

If we are interested solely in the number of ways of arranging n objects, then this is clearly just

$${}^n\mathrm{P}_n = n!$$

Returning to the example in which the thief is trying to guess your 4-digit PIN, if the thief knows that the PIN contains no repreated digits then the number of possible PINS is

$${}^{10}\mathrm{P}_4 = 5040$$

so, assuming that each is equally likely to be guessed, the probability of a correct guess is

$$P(\text{Guess correctly}) = \frac{1}{5040} = 0.0001984.$$

This illustrates how important it is to keep secret all information about your PIN. These probability calculations show that even knowing whether or not you repeat digits in your PIN is informative for a thief – it reduces the number of possible PINs by a factor of around 2.

We now consider a more complicated problem. In a tutorial group there are 40 students. What is the probability that at least two students share a birthday?

First, let's make some simplifying assumptions. We will assume that there are 365 days in a year and that each day is equally likely to be a birthday.

Call the event we are interested in $Match$. We will first calculate the probability of $No\,Match$, the probability that *no two* people have the same birthday, and calculate the probability we want using $P(Match) = 1 - P(No\,Match)$. The number of ways 40 birthdays could occur is 365^{40} since there are 365 choices for the first person and 365 choices for the second person and so on. The number of ways we can have 40 *distinct* birthdays is the same as the number of permutations of 40 objects from 365 objects, that is, ${}^{365}\mathrm{P}_{40}$. So, the probability of all birthdays being distinct is

$$P(No\,Match) = \frac{{}^{365}\mathrm{P}_{40}}{365^{40}} = \frac{365!}{325!365^{40}} \simeq 0.1$$

and so

$$P(Match) = 1 - P(No\,Match) \simeq 0.9.$$

That is, there is a probability of 0.9 that we have a match. In fact, the fact that birthdays are *not* distributed uniformly over the year makes the probability of a match even higher! This is a somewhat counter-intuitive result, and the reason is that people think more intuitively about the probability that someone has the same birthday as *themselves*. This is an entirely different problem.

An alternative way of thinking about the probability of $No\,Match$ is to consider the sequences of choices as individuals are picked from the tutorial group:

$$\begin{aligned} P(No\,Match) &= P(\text{Pick person } 1) \times P(\text{Pick person } 2 \text{ with different birthday to person } 1) \\ &\qquad \times P(\text{Pick person } 3 \text{ with different birthday to persons } 1 \text{ and } 2) \times \ldots \\ &= \frac{365}{365} \times \frac{364}{365} \times \ldots \times \frac{326}{365} \\ &\simeq 0.1. \end{aligned}$$

7.2.3 Combinations

We now have a way of counting permutations, but often when selecting objects, all that matters is *which* objects were selected, not the order in which they were selected. Suppose that we have a collection of n objects and that we wish to make r selections from this list of objects, where the order does not matter. An unordered selection such as this is referred to as a *combination*. How many ways can this be done? Notice that this is equivalent to asking how many different ways are there of choosing r objects from n objects.

For example, a company has 20 retail outlets. It is decided to try a sales promotion at 4 of these outlets. How many selections of 4 can be chosen? It may be important to know this when we come to look at the results of the trial.

This calculation is very similar to that of permutations except that the ordering of objects no longer matters. For example, if we select two objects from three objects A, B and C, there are ${}^3\mathrm{P}_2 = 6$ ways of doing this:

$$A,\ B \qquad A,\ C \qquad B,\ A \qquad B,\ C \qquad C,\ A \qquad C,\ B.$$

However, if we are not interested in the ordering, just in whether A, B or C are chosen then A, B is the same as B, A *etc.* and so the number of selections is just 3:

$$A,\ B \qquad A,\ C \qquad B,\ C.$$

The effect of ignoring the ordering reduces the number of permutations by a factor of ${}^2\mathrm{P}_2 = 2$. In general, the number of combinations of r objects from n objects is

$$\begin{aligned}\frac{\text{number of ordered samples of size } r}{\text{number of orderings of samples of size } r} &= \frac{{}^n\mathrm{P}_r}{{}^r\mathrm{P}_r} \\ &= \frac{{}^n\mathrm{P}_r}{r!} \\ &= \frac{n!}{r!(n-r)!}.\end{aligned}$$

Again, this is a very commonly found expression in combinatorics, so it has its own notation:

$${}^n\mathrm{C}_r = \frac{n!}{r!(n-r)!}.$$

There are other commonly used notations for this quantity: C_r^n and $\begin{pmatrix} n \\ r \end{pmatrix}$. These numbers are known as the *binomial coefficients*.

Now we can see that the number of ways to select 4 retail outlets out of 20 is

$${}^{20}\mathrm{C}_4 = \frac{20!}{4!16!} = 4845.$$

An easy way to calculate binomial coefficients (at least small ones) is to use the fact that

$${}^n\mathrm{C}_r = \frac{n}{r} \times \frac{n-1}{r-1} \times \frac{n-2}{r-2} \times \cdots \times \frac{n-r+1}{1}.$$

For example,

$${}^{20}\mathrm{C}_4 = \frac{20}{4} \times \frac{19}{3} \times \frac{18}{2}\frac{17}{1}.$$

To see how combinations can be used to calculate probabilities, we will look at the UK National Lottery. In this lottery, there are 49 numbered balls, and six of these are selected at random. A seventh ball is also selected, but this is only relevant if you get exactly five numbers correct. The player selects six numbers before the draw is made, and after the draw, counts how many numbers are in common with those drawn. Players win a prize if they select at least three of the balls drawn. The order in which the balls are drawn in is irrelevant.

To begin with, let's calculate the probability that exactly 3 of the 6 numbers we select are drawn. First we need to count the number of possible draws (the number of different sets of 6 numbers), and then how many of those draws correspond to getting exactly three numbers correct. The number of possible draws is the number of ways of choosing 6 objects from 49. This is

$$^{49}\mathrm{C}_6 = 13,983,816.$$

The number of drawings corresponding to getting exactly three right is calculated as follows. Of the 49 balls from which the draw is made, 6 correspond to your selected numbers, and 43 correspond to other numbers. We want to know how many ways there are of choosing 3 of your selected numbers and 3 other numbers. This is the number of ways of choosing 3 from 6, multiplied by the number of ways of choosing 3 from 43. That is, there are

$$^{6}\mathrm{C}_3\ ^{43}\mathrm{C}_3 = 246,820$$

ways of choosing exactly 3 of your selected numbers. So, the probability of matching exactly 3 numbers is

$$\frac{^{6}\mathrm{C}_3\ ^{43}\mathrm{C}_3}{^{49}\mathrm{C}_6} = \frac{246,820}{13,983,816} \simeq 0.0177.$$

Similarly, we can calculate the probability of getting other prize-winning outcomes:

$$\begin{aligned}
P(\text{match exactly 6 correct numbers}) &= \frac{^{6}\mathrm{C}_6}{^{49}\mathrm{C}_6} = \frac{1}{13,983,816} \simeq 7 \times 10^{-8} \\
P(\text{match exactly 5 correct numbers plus bonus ball}) &= \frac{^{6}\mathrm{C}_5\ ^{1}\mathrm{C}_1}{^{49}\mathrm{C}_6} = \frac{6}{13,983,816} \simeq 4 \times 10^{-7} \\
P(\text{match exactly 5 correct numbers}) &= \frac{^{6}\mathrm{C}_5\ ^{43}\mathrm{C}_1}{^{49}\mathrm{C}_6} = \frac{258}{13,983,816} \simeq 2 \times 10^{-5} \\
P(\text{match exactly 4 correct numbers}) &= \frac{^{6}\mathrm{C}_4\ ^{43}\mathrm{C}_2}{^{49}\mathrm{C}_6} = \frac{13545}{13,983,816} \simeq 1 \times 10^{-4}.
\end{aligned}$$

These outcomes are not very likely and so the prizes are chosen to reflect how likely you are to win. For example, in a recent lottery draw, the prizes were

Number of balls matched	Prize
6	£2.4M
5 plus bonus	£240K
5	£3K
4	£100
3	£10
< 3	£0

This information allows us to calculate a fair price for such a bet. The expected monetary value of the bet is

$$\begin{aligned}\mathrm{EMV} &= P(\text{match 6 balls}) \times \mathrm{Prize}(\text{match 6 balls}) \\ &\qquad + P(\text{match 5 balls plus bonus}) \times \mathrm{Prize}(\text{match 5 balls plus bonus}) \\ &\qquad\qquad + \ldots \\ &\qquad\qquad\qquad + P(\text{match 3 balls}) \times \mathrm{Prize}(\text{match 3 balls}) \\ &= 2.4M \times \frac{1}{13,983,816} + 240K \times \frac{6}{13,983,816} + \ldots + 10 \times \frac{246,820}{13,983,816} \\ &= 0.6176.\end{aligned}$$

Therefore, a fair price for a ticket in this particular lottery is around $62p$. This difference between this and the standard £1 charge for a ticket goes to “good causes” and, of course, Camelot’s profits.

7.3 Probability Distributions

7.3.1 Introduction

In Chapter 1 we saw how surveys can be used to get information on population quantities. For example, we might want to know voting intentions within the UK just before a General Election. Why does this involve *random* variables? In most cases, it is not possible to measure the variables on every member of the population and so some sampling scheme is used. This means that there is uncertainty in our conclusions. For example, if the true proportion of Labour voters were 40%, in a survey of 1,000 voters, it would be possible to get 400 Labour voters, but it would also be possible to get 350 Labour voters or 430 Labour voters. The fact that we have only a sample of voters introduces uncertainty into our conclusions about voting intentions in the population as a whole. Sometimes experiments themselves have inherent variability, for example, the toss of a coin. If the coin were tossed 1000 times and heads occurred only 400 times, would it be fair to conclude that the coin was a biased coin? The subject of Statistics has been developed to understand such variability and, in particular, how to draw correct conclusions from data which are subject to experimental and sampling variability.

Before we can make inferences about populations, we need a language to describe the uncertainty we find when taking samples from populations. First, we represent a random variable by X (capital X) and the probability that it takes a certain value x (small x) as $P(X = x)$.

The *probability distribution* of a discrete random variable X is the list of all possible values X can take and the probabilities associated with them. For example, if the random variable X is the outcome of a roll of a die then the probability distribution for X is

x	$P(X = x)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$
sum	1

Just as with sample data, it is useful to have some summary information about probability distributions. For example, what is the average value of the random variable? How much variation is there in this distribution?

7.3.2 Expectation and the population mean

The mean of a quantitative random variable is a weighted sum of its possible values, where each weight is the probability of the value occurring. This is known as the expected value of the random variable or the population mean of the random variable and is usually written as $\mathrm{E}(X)$ or μ. Therefore, for a discrete random variable,

$$\mathrm{E}(X) = \mu = \sum x\, P(X = x).$$

Previously we have seen a similar calculation when determining the expected monetary value

$$EMV = \sum P(\text{Event}) \times \text{Monetary value of Event}.$$

The expected value is the average value which we would get in an infinitely long sequence of identical experiments.

For example, suppose that the population of interest is this class and that it contains N students. Suppose that we are interested in the number of times that students have bought a particular product (e.g. a cinema ticket) in the last month. Clearly the population mean is just the average of this variable in the class:

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

where x_i is the number of times student i has bought the product. We can also write this as

$$\mu = \frac{1}{n}\sum_{j=0}^{\infty} j f_j = \sum_{j=0}^{\infty} j \frac{f_j}{n}$$

where f_j is the frequency of $x = j$ in the population and f_j/n is the relative frequency. If we choose a student at random from the class then the probability that we choose a student with $x = j$ is

$$P(X = j) = \frac{f_j}{n}$$

the relative frequency and so

$$\mu = \sum_{j=0}^{\infty} jP(X = j).$$

It is also clear that this is the average which we would get if we kept on sampling, with replacement, for a very long time.

For the die-rolling experiment, the average number of spots we would get if we repeated the experiment an "infinite" number of times is

$$E(X) = \sum x\,P(X = x) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \ldots + 6 \times \frac{1}{6} = 3.5.$$

This concept can be generalised to calculate the expected value of any function of X. For instance, in the lottery example discussed previously, the prize was determined by the number of matches. In the die-rolling experiment, we could consider a prize worth the square of the number showing: £1 for a 1, £4 for a 2, £9 for a 3, and so on. In this case the expected prize money is

$$\begin{aligned} E\left(X^2\right) &= \sum x^2\,P(X = x) \\ &= 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + \ldots + 36 \times \frac{1}{6} \\ &= \frac{91}{6} \simeq £15.17. \end{aligned}$$

7.3.3 Population variance and standard deviation

In addition to having the population mean as a measure of location, it is also useful to know about the spread of the random variable about this value. The variance of a random variable is denoted $\mathrm{Var}(X)$ or sometimes σ^2 and is determined by

$$\mathrm{Var}(X) = \sigma^2 = E\left[(X - \mu)^2\right].$$

It is simply the average squared deviation from the mean. Note that this is the same sort of calculation as with sample variances. The larger the value for the variance, the larger the spread.

Referring again back to the die-rolling experiment, if X is the number of spots, we can calculate the variance (using $\mu = 3.5$):

x	$P(X = x)$	$(x - \mu)^2$	$(x - \mu)^2 P(X = x)$
1	1/6	6.25	1.0417
2	1/6	2.25	0.3750
3	1/6	0.25	0.0417
4	1/6	0.25	0.0417
5	1/6	2.25	0.3750
6	1/6	6.25	1.0417
sum	1		2.9167

Hence

$$\mathrm{Var}(X) = 2.9167.$$

As with sample variances, there is an alternative way of calculating population variances, using

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2.$$

Using this formula with the above example gives

x	$P(X = x)$	x^2	$x^2 P(X = x)$
1	1/6	1	0.1667
2	1/6	4	0.6667
3	1/6	9	1.5000
4	1/6	16	2.6667
5	1/6	25	4.1667
6	1/6	36	6.0000
sum	1		15.1667

and so

$$\mathrm{Var}(X) = \sum x^2 P(X = x) - \mu^2 = 15.1667 - 3.5^2 = 2.9167.$$

The standard deviation of a random variable is

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

In this example, $\mathrm{SD}(X) = \sqrt{2.9167} = 1.7078$.

7.4 Exercises 7

1. Consider a lottery that is slightly different to the National Lottery in that there are 48 balls instead of 49. What is the probability of winning the jackpot in this lottery? (That is, you choose six balls and exactly these six are drawn).

2. A market survey has identified 10 desirable features for a new product. However, due to cost constraints, only four of these features can be included. If the features are selected randomly, what is the probability that your four favourites are chosen in your preferred ordering?

3. If you dial 7 digits at random on a (non-mobile) telephone in Newcastle, what is the probability you dial Dr. Farrow's office number (which has 7 digits)?

4. A sample of four mass-produced items is examined for quality control purposes. Each item can be either satisfactory (S) or unsatisfactory (U). Each item has a probability of 0.2 of being unsatisfactory and each item is independent of every other item

   (a) Consider the sequence of 4 items. In how many different sequences can we get

      i. no unsatisfactory items?
      ii. exactly 1 unsatisfactory item?
      iii. exactly 2 unsatisfactory items?
      iv. exactly 3 unsatisfactory items?
      v. four unsatisfactory items.

   (b) Find the probability of a particular sequence containing

      i. no unsatisfactory items.
      ii. exactly 1 unsatisfactory item.
      iii. exactly 2 unsatisfactory items.
      iv. exactly 3 unsatisfactory items.
      v. four unsatisfactory items.

   (c) Hence find the probability that we get

      i. no unsatisfactory items.
      ii. exactly 1 unsatisfactory item.
      iii. exactly 2 unsatisfactory items.
      iv. exactly 3 unsatisfactory items.
      v. four unsatisfactory items.

   (d) Find the mean number of unsatisfactory items.

   (e) Find the variance and standard deviation of the number of unsatisfactory items.

Chapter 8

The Binomial and Poisson Distributions

8.1 The Binomial Distribution

8.1.1 Introduction

In many surveys and experiments we collect data in the form of counts. For example, the number of people in the survey who bought a CD in the past month, the number of people who said they would vote Labour at the next election, the number of defective items in a sample taken from a production line, and so on. All these variables have common features:

1. Each person/item has only two possible (exclusive) responses (Yes/No, Defective/Not defective etc)
   – this is referred to as a *trial* which results in a *success* or *failure*

2. The survey/experiment takes the form of a random sample
   – the responses are independent.

Further suppose that the true probability of a success in the population is p (in which case the probability of a failure is $1 - p$). We are interested in the random variable X, the total number of successes out of n trials. This random variable has a probability distribution in which the probability that $X = r$, that is we get r successes in our n trials, is

$$P(X = r) = {}^{n}\mathrm{C}_{r}\, p^{r}(1 - p)^{n-r}, \quad r = 0, 1, \ldots, n.$$

These probabilities describe how likely we are to get r out of n successes from independent trials, each with success probability p. Note that any number raised to the power zero is one, for example, $0.3^0 = 1$ and $0.654^0 = 1$.

This distribution is known as the *binomial distribution* with index n and probability p. We write this as $X \sim Bin(n, p)$.

8.1.2 Calculating probabilities

For example, we can calculate the probability of getting r threes from 4 rolls of a die as follows. Each roll of the die is a trial which gives a three (success) or "not a three" (failure). The probability of a success is $p = P(\text{three}) = 1/6$. We have $n = 4$ independent trials (rolls of the die). If X is the number of threes obtained then $X \sim Bin(4, 1/6)$ and so

$$\begin{aligned}
P(X = 0) &= {}^4\mathrm{C}_0 \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^4 = \left(\frac{5}{6}\right)^4 = 0.4823 \\
P(X = 1) &= {}^4\mathrm{C}_1 \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^3 = 4 \times \frac{1}{6} \times \left(\frac{5}{6}\right)^3 = 0.3858 \\
P(X = 2) &= {}^4\mathrm{C}_2 \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^2 = 6 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2 = 0.1157 \\
P(X = 3) &= {}^4\mathrm{C}_3 \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^1 = 4 \times \left(\frac{1}{6}\right)^3 \times \frac{5}{6} = 0.0154 \\
P(X = 4) &= {}^4\mathrm{C}_4 \left(\frac{1}{6}\right)^4 \left(1 - \frac{1}{6}\right)^0 = \left(\frac{1}{6}\right)^4 = 0.0008.
\end{aligned}$$

This probability distribution shows that most of the time we would get either 0 or 1 successes but, for example, 4 successes would be quite rare.

Consider another example. A salesperson has a 50% chance of making a sale on a customer visit and she arranges 6 visits in a day. What are the probabilities of her making 0,1,2,3,4,5 and 6 sales? Let X denote the number of sales. Assuming the visits result in sales independently, $X \sim Bin(6, 0.5)$ and

No. of sales	Probability	Cumulative Probability
r	$P(X = r)$	$P(X \leq r)$
0	0.015625	0.015625
1	0.093750	0.109375
2	0.234375	0.343750
3	0.312500	0.656250
4	0.234375	0.890625
5	0.093750	0.984375
6	0.015625	1.000000
sum	1.000000	

The formula for binomial probabilities enables us to calculate values for $P(X = r)$. From these, it is straightforward to calculate cumulative probabilities such as the probability of making no more than 2 sales:

$$\begin{aligned}
P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\
&= 0.015625 + 0.09375 + 0.234375 = 0.34375.
\end{aligned}$$

These cumulative probabilities are also useful in calculating probabilities such as that of making more than 1 sale:

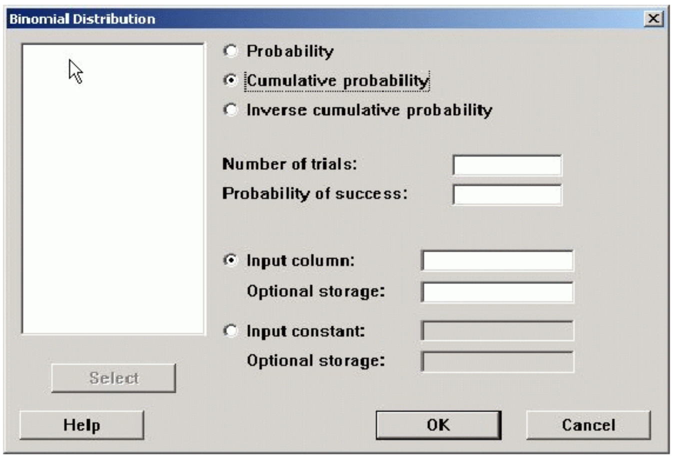
$$P(X > 1) = 1 - P(X \leq 1) = 1 - 0.109375 = 0.890625.$$

Fortunately, binomial probabilities can be found in sets of Statistical Tables or calculated using MINITAB.

Probabilities of binomial events can be calculated in MINITAB as follows. If $X \sim Bin(n,p)$ then probabilities $P(X = r)$ and cumulative probabilities $P(X \leq r)$ can be obtained using the following commands:

```
Calc > Probability Distributions > Binomial
```

This opens the following dialogue box



Binomial Distribution

○ Probability
● Cumulative probability
○ Inverse cumulative probability

Number of trials:
Probability of success:

● Input column:
Optional storage:
○ Input constant:
Optional storage:

1. Select Probability for $P(X = r)$ or Cumulative Probability for $P(X \leq r)$.
2. Enter the Number of trials (n).
3. Enter the Probability of success (p).
4. Check the `Input constant:` button
5. Enter the Input constant (r)
6. Click OK.

8.1.3 Mean and variance

If X is a random variable with a binomial $Bin(n,p)$ distribution then its mean and variance are

$$E(X) = np, \qquad Var(X) = np(1-p).$$

For example, if $X \sim Bin(4, 1/6)$ then

$$E(X) = np = 4 \times \frac{1}{6} = \frac{2}{3} = 0.6667$$

and

$$Var(X) = np(1-p) = 4 \times \frac{1}{6} \times \frac{5}{6} = \frac{5}{9} \simeq 0.5556.$$

Also

$$SD(X) = \sqrt{Var(X)} = \sqrt{\frac{5}{9}} = 0.7454.$$

8.2 The Poisson Distribution

8.2.1 Introduction

The *Poisson distribution* is a very important discrete probability distribution which arises in many different contexts. We can think of a Poisson distribution as what becomes of a binomial distribution if we keep the mean fixed but let n become very large and p become very small, i.e. a large number of trials with a small probability of success in each. In general, it is used to model data which are counts of (random) events in a certain area or time interval, without a known fixed upper limit.

For example, consider the number of calls made in a 1 minute interval to an Internet service provider (ISP). The ISP has thousands of subscribers, but each one will call with a very small probability. If the ISP knows that on average 5 calls will be made in the interval, the actual number of calls will be a Poisson random variable, with mean 5.

If X is a random variable with a Poisson distribution with parameter λ (Greek lower case *lambda*) then the probability that $X = r$ is

$$P(X = r) = \frac{\lambda^r \, e^{-\lambda}}{r!}, \quad r = 0, 1, 2, \ldots \; .$$

We write $X \sim Po(\lambda)$. The parameter λ has a very simple interpretation as the rate at which events occur. The distribution has mean and variance

$$E(X) = \lambda, \qquad Var(X) = \lambda.$$

8.2.2 Calculating probabilities

Returning to the ISP example, suppose we want to know the probabilities of different numbers of calls made to the ISP. Let X be the number of calls made in a minute. Then $X \sim P(5)$ and, for example, the probability of receiving 4 calls is

$$P(X = 4) = \frac{5^4 e^{-5}}{4!} = 0.1755.$$

We can use the formula for Poisson probabilities to calculate the probability of all possible outcomes:

	Probability	Cumulative Probability
r	$P(X=r)$	$P(X \leq r)$
0	0.0067	0.0067
1	0.0337	0.0404
2	0.0843	0.1247
3	0.1403	0.2650
4	0.1755	0.4405
5	0.1755	0.6160
6	0.1462	0.7622
7	0.1044	0.8666
8	0.0653	0.9319
9	0.0363	0.9682
10	0.0181	0.9863
$\vdots$	$\vdots$	$\vdots$
sum	1.000000	

Therefore the probability of receiving between 2 and 8 calls is

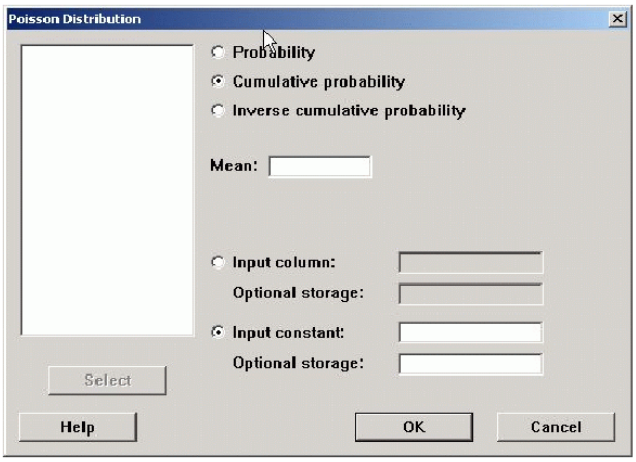
$$P(2 \leq X \leq 8) = P(X \leq 8) - P(X \leq 1) = 0.9319 - 0.0404 = 0.8915$$

and so this event is very likely. Probability calculations such as this enable ISPs to assess the likely demand for their service and hence the resources they need to provide the service.

Probabilities of Poisson events can be calculated in MINITAB as follows. If $X \sim Po(\lambda)$ then probabilities $P(X = r)$ and cumulative probabilities $P(X \leq r)$ can be obtained using the following commands:

```
Calc > Probability Distributions > Poisson
```

This opens the following dialogue box



Poisson Distribution

○ Probability
● Cumulative probability
○ Inverse cumulative probability

Mean:

○ Input column:
Optional storage:
● Input constant:
Optional storage:

1. Select Probability for $P(X = r)$ or Cumulative Probability for $P(X \leq r)$.

2. Enter the Mean (λ).

3. Check the `Input constant:` button

4. Enter the Input constant (r)

5. Click OK.

8.2.3 The Poisson distribution as an approximation to the binomial distribution

When we want to calculate probabilities in a binomial distribution with large n and small p it is often convenient to approximate the binomial probabilities by Poisson probabilities. We match the means of the distributions: $\lambda = np$.

For example, an insurance company has 1,000 customers. In a particular month, each customer has a probability of 0.003 of making a claim and all customers are independent. The distribution of the number of claims (assuming no customer will make more than one claim in a month) is then Bin$(1000,\ 0.003)$. This distribution has mean $1000 \times 0.003 = 3$. We can calculate approximate probabilities using the Poisson(3) distribution. For example, the probability that there are no claims in a month is approximately

$$P(X = 0) = \frac{3^0 e^{-3}}{0!} = e^{-3} = 0.050.$$

8.3 Exercises 8

1. An operator at a call centre has 20 calls to make in an hour. History suggests that they will be answered 85% of the time. Let X be the number of answered calls in an hour.

   (a) What probability distribution does X have?

   (b) What are the mean and standard deviation of X?

   (c) Calculate the probability of getting a response exactly 9 times.

   (d) Calculate the probability of getting fewer than 2 responses.

2. Calls are received at a telephone exchange at random times at an average rate of 10 per minute. Let X be the number of calls received in one minute.

   (a) What probability distribution does X have?

   (b) What are the mean and standard deviation of X?

   (c) Calculate the probability that there are 12 calls in one minute.

   (d) Calculate the probability there are no more than 2 calls in a minute.

3. If X_1 and X_2 are independent Poisson random variables with means λ_1 and λ_2 respectively, then $X_1 + X_2$ is a Poisson random variable with mean $\lambda_1 + \lambda_2$.

   The number of sales made by a small business in a day is a Poisson random variable with mean 2. The number of sales made on one day is independent of the number of sales made on any other day.

   (a) What is the distribution of the total number of sales in a 5-day period?

   (b) What is the probability that the business makes more than 12 sales in a 5-day period?

4. A machine is used to produce components. Each time it produces a component there is a chance that the component will be defective. When the machine is working correctly the probability that a component is defective is 0.05. Sometimes, though, the machine requires adjustment and, when this is the case, the probability that a component is defective is 0.2. Given the state of the machine, components are independent of each other. At the time in question there is a probability of 0.1 that the machine requires adjustment. Components produced by the machine are tested and either accepted or rejected. A component which is not defective is accepted with probability 0.97 and (falsely) rejected with probability 0.03. A defective component is (falsely) accepted with probability 0.15 and rejected with probability 0.85. Given the state of the machine, the acceptance of one component is independent of the acceptance of another component.

   (a) Find the conditional probability that a component is accepted given that the machine is working correctly.

   (b) Find the conditional probability that a component is rejected given that the machine is working correctly.

   (c) Find the conditional probability that a component is accepted given that the machine requires adjustment.

(d) Find the conditional probability that a component is rejected given that the machine requires adjustment.

(e) Find the conditional probability that, out of a sample of 5 components, 2 are accepted and 3 are rejected, given that the machine is working correctly.

(f) Find the conditional probability that, out of a sample of 5 components, 2 are accepted and 3 are rejected, given that the machine requires adjustment.

(g) Find the probability that the machine is working correctly and, out of a sample of 5 components, 2 are accepted and 3 are rejected.

(h) Find the probability that the machine requires adjustment and, out of a sample of 5 components, 2 are accepted and 3 are rejected.

(i) Find the probability that, out of a sample of 5 components, 2 are accepted and 3 are rejected.

(j) Find the conditional probability that the machine is working correctly given that, out of a sample of 5 components, 2 are accepted and 3 are rejected.

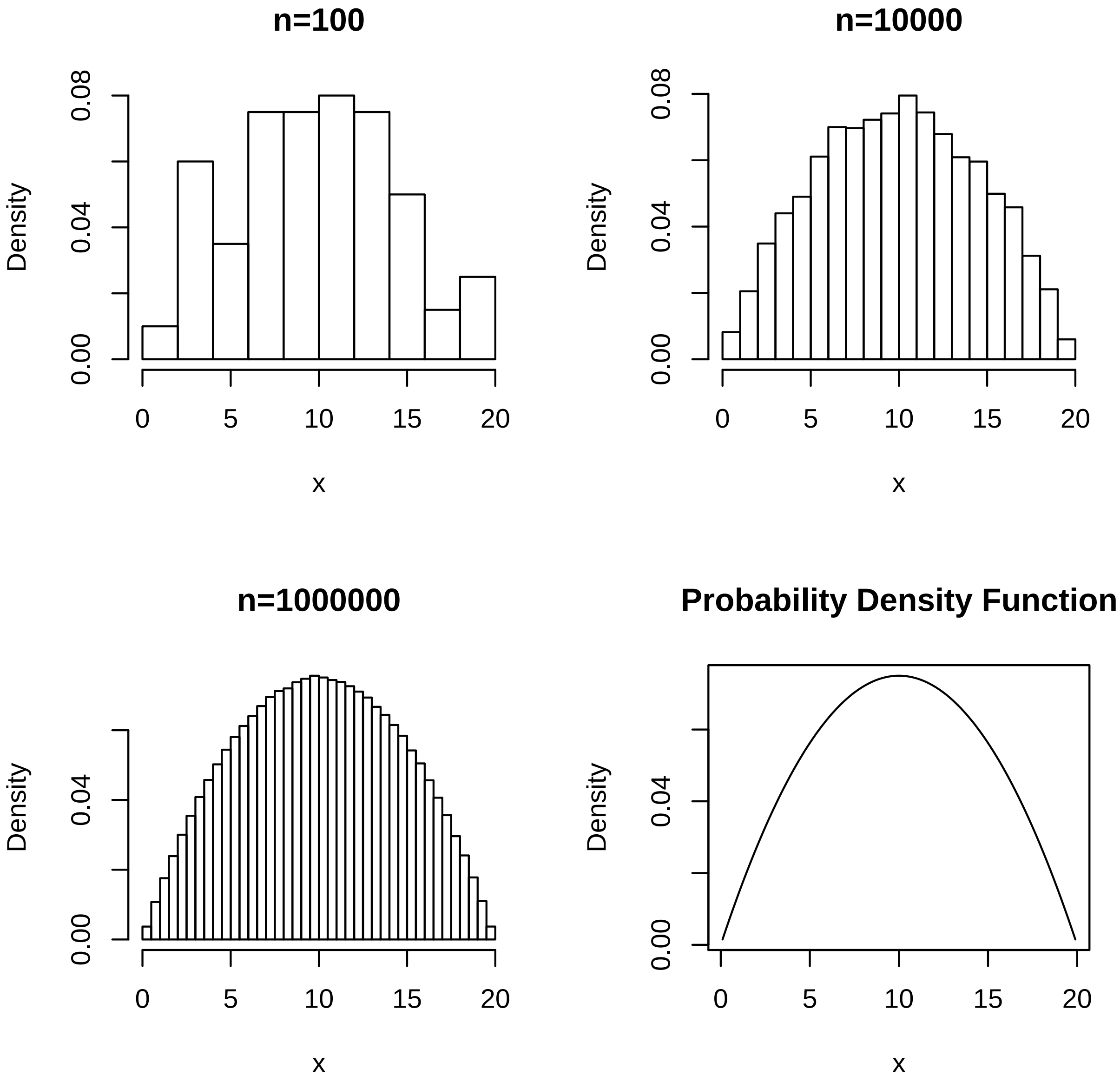
Chapter 9

Continuous Probability Models

9.1 Introduction

We have seen how discrete random variables can be modelled by discrete probability distributions such as the binomial and Poisson distributions. We now consider how to model continuous random variables. A variable is discrete if it takes a *countable* number of values, for example, $r = 0, 1, 2, \ldots, n$ or $r = 0, 1, 2, \ldots$ or $r = 0, 0.1, 0.2, \ldots, 0.9, 1.0$. In contrast, the values which a continuous variable can take form a continuous scale. One simple example of a continuous variable is height. Although in practice we might only record height to the nearest cm, if we could measure height exactly (to billions of decimal places) we would find that everyone had a different height. This is the essential difference between discrete and continuous variables. Therefore, if we could measure the exact height of every one of the n people on the planet, we would find that, for any height x, the proportion of people of height x is either $1/n$ or 0. And if we imagine the number of people on the planet growing over time ($n \to \infty$), this proportion tends to zero. This feature poses a problem for modelling continuous random variables as we can no longer use the methods we have seen work for discrete random variables.

The solution can be found by considering a (relative frequency) histogram of a sample of values taken by the continuous random variable, and thinking about what happens to the histogram as the sample size increases. For example, consider the following graphs which show histograms for samples of 100, 10000 and 1000000 observations made on a continuous random variable which can take values between 0 and 20. The final graph shows what happens when the sample size becomes infinitely big. This final graph is called the probability density function.



As the population sizes gets larger, the histogram intervals get smaller and the jagged profile of the histogram smooths out to become a curve. We call this curve the *probability density function (pdf)* and it is usually written as $f(x)$. Note that probabilities such as $P(X < x)$ can be determined using the pdf as they equate to areas under the curve.

The key features of pdfs are

1. pdfs never take negative values
2. the area under a pdf is one: $P(-\infty < X < \infty) = 1$
3. areas under the curve correspond to probabilities
4. $P(X \leq x) = P(X < x)$ since $P(X = x) = 0$.

We now consider some particular probability distributions that are often used to describe continuous random variables.

9.2 The Uniform Distribution

The *uniform distribution* is the most simple continuous distribution. As the name suggests, it describes a variable for which all possible outcomes are equally likely. For example, suppose you manage a group of Environmental Health Officers and need to decide at what time they should inspect a local hotel. You decide that any time during the working day (9.00 to 18.00) is okay but you want to decide the time "randomly". Here randomly is a short-hand for "a random time, where all times in the working day are equally likely to be chosen". Let X be the time to their arrival at the hotel measured in terms of minutes from the start of the day. Then X is a uniform random variable between 0 and 540:

The total area (base × height) under the pdf must equal one. Therefore, as the base is 540, the height must be $1/540$. Hence the probability density function (pdf) for the continuous random variable X is

$$f(x) = \begin{cases} \dfrac{1}{540} & \text{for } 0 \leq x \leq 540 \\ 0 & \text{otherwise.} \end{cases}$$

In general, we say that a random variable X which is equally likely to take any value between a and b has a uniform distribution on the interval a to b. The random variable has probability density function (pdf)

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and probabilities can be calculated using the formula

$$P(X \leq x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b. \end{cases}$$

Therefore, for example, the probability that the inspectors visit the hotel in the morning (within 180 minutes after 9am) is

$$P(X \leq 180) = \frac{180 - 0}{540 - 0} = \frac{1}{3}.$$

The probability of a visit during the lunch hour (12.30 to 13.30) is

$$\begin{aligned} P(210 \leq X \leq 270) &= P(X \leq 270) - P(X < 210) \\ &= \frac{270 - 0}{540 - 0} - \frac{210 - 0}{540 - 0} \\ &= \frac{270 - 210}{540} \\ &= \frac{60}{540} \\ &= \frac{1}{9}. \end{aligned}$$

9.2.1 Mean and Variance

The mean and variance of a continuous random variable can be calculated in a similar manner to that used for a discrete random variable. However the specific techniques required to do this are outside the scope of this course and so we will simply state the results.

If X is a uniform random variable on the interval a to b then its mean and variance are

$$E(X) = \mu = \frac{a + b}{2}, \qquad Var(X) = \sigma^2 = \frac{(b - a)^2}{12}.$$

In the above example, we have

$$E(X) = \frac{a + b}{2} = \frac{0 + 540}{2} = 270,$$

so that the mean arrival of the inspectors is 9am $+ 270$ minutes $= 13.30$. Also

$$Var(X) = \frac{(540 - 0)^2}{12} = 24300$$

and therefore $SD(X) = \sqrt{Var(X)} = \sqrt{24300} = 155.9$ minutes.

9.3 The Exponential Distribution

The *exponential distribution* is another common distribution that is used to describe continuous random variables. It is often used to model lifetimes of products and times between “random” events such as arrivals of customers in a queueing system or arrivals of orders. The distribution has one parameter, λ. Its probability density function is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

and probabilities can be calculated using

$$P(X \leq x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } x > 0. \end{cases}$$

The main features of this distribution are:

1. an exponentially distributed random variable can only take positive values
2. larger values are increasingly unlikely – exponential decay
3. the value of λ fixes the rate of decay – larger values correspond to more rapid decay.

Consider an example in which the time (in minutes) between successive users of a pay phone can be modelled by an exponential distribution with $\lambda = 0.3$. The probability of the gap between phone users being less than 5 minutes is

$$P(X < 5) = 1 - e^{-0.3 \times 5} = 1 - 0.223 = 0.777.$$

Also the probability that the gap is more than 10 minutes is

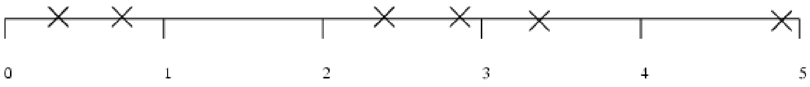
$$P(X > 10) = 1 - P(X \leq 10) = 1 - \left(1 - e^{-0.3 \times 10}\right) = e^{-0.3 \times 10} = 0.050$$

and the probability that the gap is between 5 and 10 minutes is

$$P(5 < X < 10) = P(X < 10) - P(X \leq 5) = 0.950 - 0.777 = 0.173.$$

One of the main uses of the exponential distribution is as a model for the times between events occurring randomly in time. We have previously considered events which occur at random points in time in connection with the Poisson distribution. The Poisson distribution describes probabilities for the number of events taking place in a given time period. The exponential distribution describes probabilities for the times between events. Both of these concern events occurring randomly in time (at a constant average rate, say λ). This is known as a *Poisson process*.

Consider a series of randomly occurring events such as calls at a credit card call centre. The times of calls might look like



There are two ways of viewing these data. One is as the number of calls in each minute (here 2, 0, 2, 1 and 1) and the other as the times between successive calls. For the Poisson process,

- the number of calls has a Poisson distribution with parameter λ, and
- the time between successive calls has an exponential distribution with parameter λ.

9.3.1 Mean and Variance

The mean and variance of the exponential distribution can be shown to be

$$E(X) = \mu = \frac{1}{\lambda}, \qquad Var(X) = \sigma^2 = \frac{1}{\lambda^2}.$$

9.4 Exercises 9

1. An express coach is due to arrive in Newcastle from London at 23.00. However in practice it is equally likely to arrive anywhere between 15 minutes early to 45 minutes late, depending on traffic conditions. Let the random variable X denote the amount of time (in minutes) that the coach is delayed.

   (a) Sketch the pdf.

   (b) Calculate the mean and standard deviation of the delay time.

   (c) What is the probability that the coach is less than 5 minutes late?

   (d) What is the probability that the coach is more than 20 minutes late?

   (e) What is the probability that the coach arrives between 22.55 and 23.20?

   (f) What is the probability that the coach arrives at 23.00?

   (g) What is the probability that the coach arrives at 0.00?

   (h) Do you think that this is a good model for the coach's arrival time?

2. A network server receives incoming requests according to a Poisson process with rate $\lambda = 2.5$ per minute.

   (a) What is the expectation of the time between arrivals of requests?

   (b) What is the probability that the time between requests is less than 2 minutes?

   (c) What is the probability that the time between requests is greater than 1 minute?

   (d) What is the probability that the time between requests is between 30 seconds and 50 seconds?

3. As Production Manager, you are responsible for buying a new piece of equipment for your company's production process. A salesman from one company has told you that he can supply you with equuipment for which the time to first breakdown (in months) follows an exponential distribution with $\lambda = 0.11$. Another salesman (from another company) has told you that the time to first breakdown of their machines is also exponentially distributed but with $\lambda = 0.01$. It is very important that the equipment you purchase does not break down for at least six months. Calculate the probability of this outcome for both suppliers and make a recommendation to the company board about which machine should be bought.

   How might you take into account a difference between the prices for the machines?

Chapter 10

The Normal Distribution

10.1 Introduction

The *normal* distribution is possibly the best known and most used continuous probability distribution. It provides a good model for data in very many different applications, for example, the yields of crops, the heights of people, students' marks. The outcomes of many production processes also follow normal distributions and hence it is used widely in industry.

The normal distribution has two parameters: the mean μ and the variance σ^2. The standard deviation $\sigma = \sqrt{\sigma^2}$ but we usually use the variance to specify the parameters. The probability density function of a normal distribution is often said to be "bell shaped" :

The formula for the pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

There is no simple formula for calculating probabilities. However, they can be determined using tables or statistical packages such as Minitab.

There are four important characteristics of the normal distribution:

1. It is symmetrical about its mean, μ.

2. The mean, median and mode all coincide.
3. The area under the curve is equal to 1.
4. The curve extends for ever in both directions to infinity (i.e. to $\pm\infty$).

Below is a plot of the pdf of normal distributions for different values of μ and σ.

Note that the mean μ locates the distribution on the x–axis and the standard deviation σ affects the spread of the distribution, with larger values giving flatter and wider curves.

10.2 Notation

If a random variable X has a normal distribution with mean μ and variance σ^2, then we write

$$X \sim N\left(\mu, \sigma^2\right).$$

For example, a random variable X which follows a normal distribution with mean 10 and variance 25 is written as $X \sim N\left(10, 25\right)$ or $X \sim N\left(10, 5^2\right)$. It is important to note that the second parameter in this notation is the variance and not the standard deviation.

10.3 Some properties of normal variables

An important property concerns addition (or subtraction) of normal random variables. If X_1 and X_2 both have normal distributions and $Y = X_1 + X_2$ then Y also has a normal distribution. If $Z = X_1 - X_2$ then Z also has a normal distribution. If $X_1 \sim N(\mu_1,\ \sigma_1^2)$ and $X_2 \sim N(\mu_2,\ \sigma_2^2)$ then Y has mean $\mu_y = \mu_1 + \mu_2$ and Z has mean $\mu_z = \mu_1 - \mu_2$. If X_1 and X_2 are *independent* then Y and Z both have variance $\sigma_y^2 = \sigma_z^2 = \sigma_1^2 + \sigma_2^2$.

For example, if wagonloads of material have weights which are normally distributed with mean 3 tonnes and standard deviation 0.5 tonnes then the total weight from two wagonloads is normally distributed with mean $3 + 3 = 6$ tonnes and standard deviation $\sqrt{0.5^2 + 0.5^2} = 0.7071$ tonnes.

If $X \sim N(\mu,\ \sigma^2)$ and a and b are fixed numbers and $W = a + bX$ then $W \sim N(a + b\mu,\ k^2\sigma^2$. For example, if the weight of a load of material has a normal distribution with mean 3 tonnes and standard deviation 0.5 tonnes and we take exactly 1 tonne out then the weight of the remainder has a normal distribution with mean 2 tonnes and standard deviation 0.5 tonnes. If the price we can expect to get for a certain quantity of our product is £X and $X \sim N(1000,\ 50^2)$ and a pound is worth 1.6 euros then the price in euros is $Y \sim N(1600,\ 80^2)$.

An important consequence of this is that, if $X \sim N(\mu, \sigma^2)$, and

$$Z = \frac{X - \mu}{\sigma}$$

then $Z \sim N(0,\ 1)$.

10.3.1 Probability calculations and the standard normal distribution

All probabilities for the normal distribution can be expressed in terms of those for a normal distribution with mean 0 and variance 1. Usually, a random variable with this *standard normal distribution* is called Z, that is

$$Z \sim N\left(0, 1\right).$$

Probabilities for a random variable $X \sim N(\mu, \sigma^2)$ can be determined in terms of those for $Z \sim N(0, 1)$ using the formula

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

and values for $P(Z \leq z)$ can be found in tables (see the end of this chapter). Therefore, the probability $P(X \leq x)$ can be found by looking up (in standard normal tables) the probability corresponding to

$$z = \frac{x - \mu}{\sigma}.$$

We now look at a series of examples to illustrate how to calculate normal probabilities using these tables.

1. We first look at how to determine probabilities for the standard normal distribution.

   (a) The probability that the random variable Z is less than -1.46 is $P(Z < -1.46)$. Therefore we look for the probability in tables corresponding to $z = -1.46$: row labelled -1.4, column headed -0.06. This gives $P(Z < -1.46) = 0.0721$.

   (b) The probability that the random variable Z is less than -0.01 is $P(Z < -0.01)$. Therefore we look for the probability in tables corresponding to $z = -0.01$: row labelled 0.0, column headed -0.01. This gives $P(Z < -0.01) = 0.4960$.

   (c) The probability that the random variable Z is less than 0.01 is $P(Z < 0.01)$. Therefore we look for the probability in tables corresponding to $z = 0.01$: row labelled 0.0, column headed 0.01. This gives $P(Z < 0.01) = 0.5040$.

   (d) The probability that the random variable Z is greater than 1.5 is $P(Z > 1.5)$. Now $P(Z > 1.5) = 1 - P(Z \leq 1.5)$. Therefore we look for the probability in tables corresponding to $z = 1.5$: row labelled 1.5, column headed 0.00. This gives $P(Z < 1.5) = 0.9332$ and therefore $P(Z > 1.5) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668$.

(e) The probability that the random variable Z lies between than -1.2 and 1.5 is

$$\begin{aligned} P(-1.2 < Z < 1.5) &= P(Z < 1.5) - P(Z \leq -1.2) \\ &= 0.9332 - 0.1151 \\ &= 0.8181. \end{aligned}$$

We now consider how to determine probabilities for general normal distributions.

2. Suppose we are interested in the IQ of 18-20 year olds and that IQs follow a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$. Mathematically we let the random variable X denote the IQ of a randomly chosen person from this age group and then $X \sim N(100, 15^2)$. Probability statements about IQs can be made as follows.

   (a) The probability that an 18-20 year old has an IQ less than 85 is $P(X < 85)$. Using the formula

   $$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

   we need to calculate

   $$z = \frac{85-\mu}{\sigma} = \frac{85-100}{15} = -1$$

   and from tables we obtain $P(Z < -1) = 0.1587$. Therefore

   $$P(X < 85) = 0.1587.$$

   (b) The probability that an 18-20 year old has an IQ greater than 142 is $P(X > 142)$. Now $P(X > 142) = 1 - P(X \leq 142)$ and

   $$z = \frac{142-\mu}{\sigma} = \frac{142-100}{15} = 2.8.$$

   Using tables, we see that $P(Z \leq 2.8) = 0.9974$ and so

   $$P(X > 142) = 1 - 0.9974 = 0.0026.$$

3. Suppose that the vitamin C content per $100g$ tin of tomato juice is normally distributed with mean $\mu = 20mg$ and standard deviation $\sigma = 4mg$. Let X be the vitamin C content of a randomly chosen tin.

   (a) The probability that the tin has less than $25mg$ of vitamin C is $P(X < 25)$. Now

   $$z = \frac{25-\mu}{\sigma} = \frac{25-20}{4} = 1.25$$

   and from tables we obtain $P(Z < 1.25) = 0.8944$. Therefore

   $$P(X < 25) = 0.8944.$$

   (b) The probability that the tin has more than $25mg$ of vitamin C is $P(X > 25)$. Now $P(X > 25) = 1 - P(X \leq 25)$ and so

   $$P(X > 25) = 1 - 0.8944 = 0.1056.$$

(c) The probability that the tin has between $18mg$ and $25mg$ of vitamin C is

$$Pr(18 < X < 25) = P(X < 25) - P(X \leq 18).$$

We can determine $P(X \leq 18)$ from tables using

$$z = \frac{18 - \mu}{\sigma} = \frac{18 - 20}{4} = -0.5,$$

giving $P(X \leq 18) = P(Z < -0.5) = 0.3085$. Therefore

$$Pr(18 < X < 25) = 0.8944 - 0.3085 = 0.5859.$$

We can also use the tables in reverse. For example, we might want to know below what value are 95% of the population. This is equivalent to determining the value of z that satisfies $P(Z < z) = 0.95$. From tables, we can see that $P(Z < 1.64) = 0.9495$ and $P(Z < 1.65) = 0.9505$. Therefore the value we want for z lies between 1.64 and 1.65. If a more accurate value is needed we can interpolate between these values: 0.95 is half-way between 0.9495 and 0.9505 and so we take $z = 1.645$. This is a more accurate answer and sufficient in most cases. However, the exact value for z can be found from more detailed tables or via a computer package such as Minitab. Here are some more examples.

1. Below what value does 10% of the standard normal population fall? From tables we get

$$P(Z < -1.28) = 0.1003 \quad \text{and} \quad P(Z < -1.29) = 0.0985$$

and so we take

$$\begin{aligned} z &= -1.29 + \frac{0.1 - 0.0985}{0.1003 - 0.0985} \times \{-1.28 - (-1.29)\} \\ &= -1.29 + \frac{0.0015}{0.0018} \times 0.01 \\ &= -1.2817. \end{aligned}$$

In other words $P(Z < -1.2817) = 0.1$ and so 10% of the standard normal population falls below -1.2817.

2. A similar calculation can be used to calculate the IQ that identifies the 10% of 18-20 year olds with the smallest IQ. We need the value of x, where $P(X < x) = 0.1$. Now this population has $\mu = 100$ and $\sigma = 15$. Also

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

and so we need x so that

$$P\left(Z \leq \frac{x - 100}{15}\right) = 0.1.$$

We know (from earlier) that $P(Z < -1.2817) = 0.1$ and therefore we solve

$$\frac{x - 100}{15} = -1.2817,$$

that is

$$\begin{aligned}x &= 100 - 1.2817 \times 15 \\ &= 100 - 19.2255 \\ &= 80.7745.\end{aligned}$$

Notice that the calculation that transforms the z–value onto the x–scale is

$$x = \mu + z\sigma.$$

3. What is the IQ that identifies the 1% of 18-20 year olds with the greatest IQ? Again, we first determine the value z that identifies the top 1% of a standard normal population and then translate this into an IQ. So we need the value z that satisfies $P(Z > z) = 0.01$. This is the same value as satisfies $P(Z < z) = 0.99$. A quick examination of tables gives the two key probabilities as

   $$P(Z < 2.32) = 0.9898 \quad \text{and} \quad P(Z < 2.33) = 0.9901$$

   and so we take

   $$\begin{aligned}z &= 2.32 + \frac{0.99 - 0.9898}{0.9901 - 0.9898} \times \{2.33 - 2.32\} \\ &= 2.32 + \frac{0.0002}{0.0003} \times 0.01 \\ &= 2.3267.\end{aligned}$$

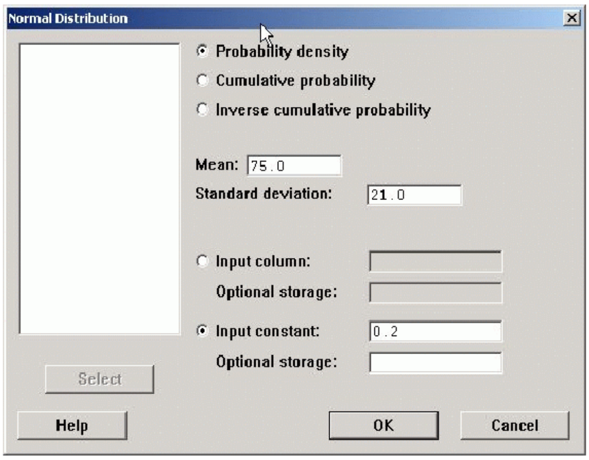
   In other words $P(Z < 2.3267) = 0.99$ and so 1% of the standard normal population lies above $z = 2.3267$. Moving back to the IQ scale, we need the value x such that $P(X > x) = 0.01$ and so we take

   $$\begin{aligned}x &= \mu + z\sigma \\ &= 100 + 2.3267 \times 15 \\ &= 134.9.\end{aligned}$$

Minitab is very helpful with calculating normal probabilities. The following commands will calculate probabilities $P(X < x)$ and also values of x that satisfy $P(X < x) = p$:

1. `Calc > Probability Distributions > Normal`

   opens up dialogue box

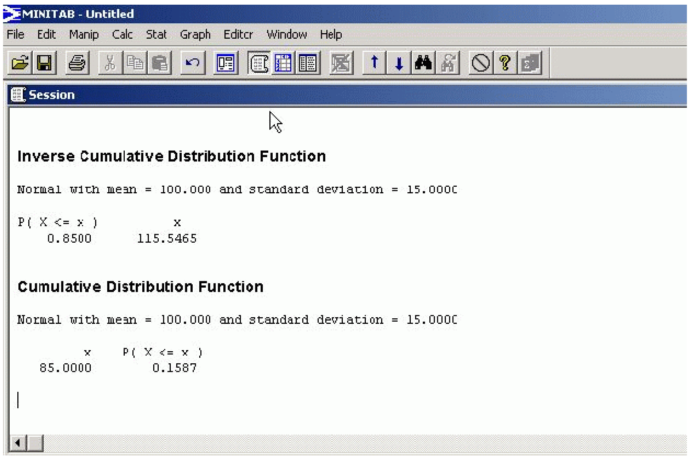


● Probability density
○ Cumulative probability
○ Inverse cumulative probability

Mean: 75.0
Standard deviation: 21.0

○ Input column:
Optional storage:
● Input constant: 0.2
Optional storage:

2. Select Cumulative probability for $P(X < x)$ or Inverse cumulative probability for the value of x satisfying $P(X < x) = p$
3. Enter the Mean (μ) and the Standard Deviation (σ)
4. Select Input Constant and enter the value for x or p (as appropriate)
5. Click OK
6. The answer is displayed in the Session Window:



```
Inverse Cumulative Distribution Function

Normal with mean = 100.000 and standard deviation = 15.000C

P( X <= x )          x
    0.8500      115.5465


Cumulative Distribution Function

Normal with mean = 100.000 and standard deviation = 15.000C

        x      P( X <= x )
  85.0000         0.1587
```

10.4 The normal approximation to the binomial distribution

The normal distribution can be used as an approximation to the binomial distribution $\text{Bin}(n, p)$ for large n and medium p. (Say if both np and $n(1 - p)$ are greater than about 7). To approximate a Bin(n, p) distribution we use a normal distribution with the same mean and variance. That is $\mu = np$ and $\sigma^2 = np(1-p)$.

Example. We plan to do a market research survey in which people will be asked whether or not they would buy a new product. A random sample of 600 people will be asked. Suppose that the true proportion of people in the population who would buy the product is 40%, i.e. $p = 0.4$. Find the probability that, in our survey, between 220 and 260 (inclusive) answer "Yes."

In this case $n = 600$ and $p = 0.4$ so $np = 240$ and $n(1-p) = 360$. Therefore we can use the approximation. We use the normal distribution with mean $\mu = np = 240$ and variance $\sigma^2 = np(1-p) = 144$. That is $N(240, 144)$. The standard deviation is $\sigma = \sqrt{144} = 12$.

If $X \sim N(240, 144)$ then

$$\begin{aligned} \Pr(X < 220) &= \Pr\left(\frac{X-240}{12} < \frac{220-240}{12}\right) \\ &= \Pr(Z < -1.67) \\ &= 0.0475. \end{aligned}$$

So $\Pr($ fewer than 220 "Yes" $) \approx 0.0475$.

We can sometimes obtain a better approximation by using a *continuity correction.* Since $B(n,p)$ is a discrete distribution, $\Pr($ number of "Yes" $= 220) > 0$. Using the continuity correction, we count as "220" everything between 219.5 and 220.5 so, for $\Pr($ number of "Yes" $< 220)$ we would use

$$\Pr(X < 219.5) = \Pr\left(Z < \frac{219.5-240}{12}\right) = \Pr(Z < -1.71) \approx 0.0436.$$

Here the continuity correction makes a noticeable difference. Sometimes it does not.

In the same way

$$\Pr(X < 260.5) = \Pr\left(Z < \frac{260.5-240}{12}\right) = \Pr(Z < 1.71) \approx 0.9564 = 1 - 0.0436.$$

So the probability that, in our survey, between 220 and 260 (inclusive) people say "Yes" is approximately $0.9564 - 0.0436 = 0.9128 \approx 0.91$.

10.5 Exercises 10

1. The weights of bags of animal feed made in a mill follow a normal distribution with mean $\mu = 8.1$ kg and standard deviation $\sigma = 0.07$kg.

   (a) What is the probability that the weight of a bag is over 8.25kg?

   (b) What is the probability that the weight of a bag is between 8.0kg and 8.25kg?

   (c) A customer requires bags which weigh no less than 8.0kg. What percentage of the mill's output can be used to supply this customer?

   (d) The mill is trying to negotiate a new contract with this customer. It is in the mill's interests to be able to supply 98% of its output to the customer. What is the largest weight which achieves this requirement?

2. A drinks machine is regulated by its manufacturer so that it discharges an average of $200ml$ per cup. However, the machine is not particularly accurate and actually discharges an amount that has a normal distribution with standard deviation $15ml$.

   (a) What percentage of cups contain below the minimum permissible volume of $170ml$?

   (b) What percentage of cups contain over $225ml$?

   (c) What is the probability that the cup contains between $175ml$ and $225ml$?

   (d) How many cups would you expect to overflow if $250ml$ cups are used for the next 10000 drinks?

3. A company promises delivery within 20 working days of receipt of order. However in reality they deliver according to a normal distribution with a mean of 16 days and a standard deviation of 2.5 days.

   (a) What proportion of customers receive their order late?

   (b) What proportion of customers receive their orders between 10 and 15 days of placing their order?

   (c) How many days should the delivery promise be adjusted to if only 3% of orders are to be late?

   (d) A new order processing system promises to reduces the standard deviation of delivery times to 1.5 days. If this system is used, what proportion of customers will receive their deliveries within 20 days?

4. Bananas of a certain variety have weights, in kg, which are independent and normally distributed with mean 0.15 and variance 0.0025. Find the probability that

   (a) three bananas weigh more than 0.5kg.

   (b) four bananas weigh more than 0.5kg.

   In a shop bananas are put on the scales one at a time until the total weight becomes more than 0.5kg. Find the probability that exactly four bananas are needed.

   Explain why the first sentence in this question can only be approximately, not exactly, true.

5. Bags of produce have a nominal weight of 1kg. In fact the weights have a normal distribution with mean 1064g and standard deviation 50g. A bag which weighs less than 1kg is considered to be underweight.

   (a) Find the probability that a bag is underweight.

   (b) Assuming that the weights of bags are independent, find an approximate value for the probability that, in a batch of 100 bags, more than 15 are underweight.

Probability Tables for the Standard Normal Distribution

The table contains values of $Pr(Z < z)$, where $Z \sim N(0, 1)$.

z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	0.00
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986