



## Statistics Skills:

# Data summaries and some basic statistical inference

Dr. Lee Fawcett

School of Mathematics & Statistics

## Introduction

Data are the key to answering many important questions in transport. Should a local authority invest in traffic calming measures, such as speed ramps or road narrowing? Is a recently-installed speed camera really that effective? Does the cost associated with a new road traffic safety measure represent good value for money? What can we expect the daily traffic flow to be along a new bypass? Has the introduction of a new park-and-ride scheme had an impact on the volume of traffic in a nearby town centre? Has the new tram system in a city eased traffic congestion? These are all questions that can be answered with data. But how do we get our data? And once we've got it, what do we do with it? We will start off by thinking about some basic data collection techniques, before considering some simple methods for summarising data. In tomorrow's session, we will consider more sophisticated methods of data analysis, including techniques for testing simple hypotheses. Before we start, it's worth being familiar with some key words which will crop up from time-to-time...

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 time it takes children to get to a particular school from their house in the morning, then this is our random variable; a particular child's time would be an observation and if we recorded the time for each child in a particular class, those would be our data, which would form a *sample* from the population of all children in the school. As a shorthand we often use a capital letter (say  $X$ ) to represent our random variable, and the lower case, with subscripts (e.g.  $x_1, x_2, \dots$ ), to represent observations *in a general sense*.

In practice it is difficult to observe whole populations, unless we are interested in a very limited population, e.g. just the children in a particular class in one school. In reality we usually observe a subset of the population, and this raises questions about who or what goes into our sample; we will briefly consider various commonly-used sampling techniques soon.

Once we have our data, it is important to understand what *type* it is – so we can figure out exactly what to do with it. The two main categories are *quantitative* and *qualitative* (in other words, numeric and non-numeric, respectively). Quantitative data can be further subdivided into categories referred to as *discrete* and *continuous*. Discrete data moves in 'steps' – for example, whole numbers, shoe sizes etc. Continuous data can usually take *any* value, although there may be restrictions to the set of positive numbers (for example height, weight) and there may be issues of accuracy depending on the equipment being used to record the data (and so sometimes continuous data might 'look' discrete!). There are other words that are often used within these categories, such as 'ordinal', and 'categorical'.

The following diagrams should help to pull all of this together.



# 1 Collecting data

In this section, we consider the most commonly-used methods for collecting data. Here, we assume that we do not have the resources to record data on *all* items in the population – that is, we assume we cannot take a *census* – and so essentially we need to think about a strategy for taking a sample from our population. We want the people/items in our sample to ‘mirror’ the people/items in the population as a whole, and so this needs to be done carefully; we would like our sample to be *representative* of the population, and we would like to avoid a *biased* sample.

## 1.1 Simple random sampling

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

## 1.2 Stratified sampling

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

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

### 1.3 Systematic sampling

This is a form of sampling which can be used where the population is clearly structured. For example, if you were interested in obtaining a 10% sample from a batch of bicycle components being manufactured, you would select the first component at random; after that, you pick every tenth item to come off the production line. The simplicity of selection makes this a particularly easy sampling scheme to implement, especially in a production setting. The disadvantages of this method are that it is not completely 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.

### 1.4 Multi-stage sampling

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

### 1.5 Judgemental sampling

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

### 1.6 Accessibility sampling

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

## 1.7 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.

## 1.8 Sample size

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

## 2 Presenting data

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; graphical displays can also be used for a more visually arresting summary.

### 2.1 Tabular presentations

There are many types of tables. For example, we have all seen tables listing sales of cars by type, or exchange rates, or the financial performance of companies. These types of tables can be very informative. However, they can also be difficult to interpret, especially those which contain vast amounts of data.

Frequency tables are amongst the most commonly-used tables and are perhaps the most easily understood. They can be used with continuous, discrete, categorical and ordinal data. Frequency tables are a useful starting point for constructing some of the graphical summaries we will look at shortly.

#### 2.1.1 Frequency tables for categorical data

The following table presents the modes of transport used daily by 30 students to get to and from University (survey date: 3rd August 2013).

| 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 | 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                    |
| <b>Total</b> | 30        | 100                    |

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*.

### 2.1.2 Frequency tables for count data

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

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

These data can be presented in a relative frequency table by number of days on which different numbers of cars were sold. A tally count can be used to keep track of the frequencies (although probably not necessary here because the frequencies in each category are fairly small); we can then convert these frequencies into relative frequencies as before. The corresponding table is shown below.

| Cars Sold     | Tally | Frequency | Relative Frequency % |
|---------------|-------|-----------|----------------------|
| 1             |       | 0         | 0                    |
| 2             |       | 0         | 0                    |
| 3             |       | 0         | 0                    |
| 4             | I     | 1         | 7.14                 |
| 5             | I     | 1         | 7.14                 |
| 6             | II    | 2         | 14.29                |
| 7             | II    | 2         | 14.29                |
| 8             | III   | 3         | 21.43                |
| 9             | II    | 2         | 14.29                |
| 10            | II    | 2         | 14.29                |
| 11            | I     | 1         | 7.14                 |
| <b>Totals</b> | 14    | 14        | 100                  |

### 2.1.3 Frequency tables for continuous data

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

| Class Interval               | Tally      | Frequency | Relative Frequency % |
|------------------------------|------------|-----------|----------------------|
| $175 \leq \text{time} < 180$ | I          | 1         | 2                    |
| $180 \leq \text{time} < 185$ | III        | 3         | 6                    |
| $185 \leq \text{time} < 190$ | III        | 3         | 6                    |
| $190 \leq \text{time} < 195$ | III I      | 6         | 12                   |
| $195 \leq \text{time} < 200$ | III III    | 10        | 20                   |
| $200 \leq \text{time} < 205$ | III III II | 12        | 24                   |
| $205 \leq \text{time} < 210$ | III III    | 8         | 16                   |
| $210 \leq \text{time} < 215$ | III        | 3         | 6                    |
| $215 \leq \text{time} < 220$ | III        | 3         | 6                    |
| $220 \leq \text{time} < 225$ | I          | 1         | 2                    |
| <b>Totals</b>                |            | 50        | 100                  |

Some notes:

- The width of the class intervals can be tricky to get right – too wide, and there won’t be enough intervals to unravel any patterns in the data, too narrow and there could be lots of intervals with very small frequencies
- A good rule of thumb: Aim for between 10–15 intervals
- Make sure the interval widths are sensible – 5’s, 10’s, 20’s, 100’s etc.
- The intervals don’t all have to be the same width!

## 2.2 Graphical methods for presenting data

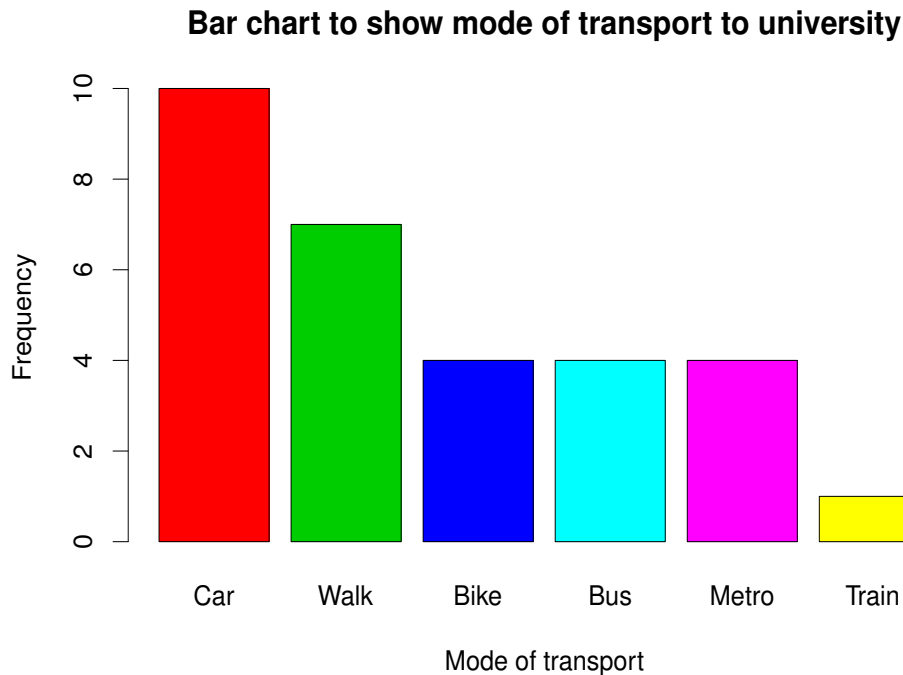
We have looked at ways of collecting data and then collating them into tables. Frequency tables are useful methods of presenting data; they do, however, have their limitations. With large amounts of data graphical presentation methods are often clearer to understand. Here, we look at some of the most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

### 2.2.1 Bar charts

Bar charts are a commonly-used and clear way of presenting categorical data or any ungrouped discrete frequency observations. For example, recall our frequency table showing students' modes of transport:

| Mode         | Frequency |
|--------------|-----------|
| Car          | 10        |
| Walk         | 7         |
| Bike         | 4         |
| Bus          | 4         |
| Metro        | 4         |
| Train        | 1         |
| <b>Total</b> | 30        |

We can then present this information as a bar chart:



This bar chart clearly shows that the most popular mode of transport is the car and that the metro, bus and cycling are all equally popular (in our small sample). Bar charts provide a

method of quickly spotting simple patterns of popularity within a discrete data set. Some things to note:

- You *must* leave equal-width spaces between each bar;
- All bars *must* be the same width!
- Colours can be used to make the chart more interesting.

### 2.2.2 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 best used when there are only a handful of categories to display. Consider the following frequency table which shows the numbers of different vehicles travelling under a motorway bridge during rush hour.

| Vehicle      | Frequency  |
|--------------|------------|
| Car          | 76         |
| Lorry        | 40         |
| Motorbike    | 17         |
| Coach        | 6          |
| Caravan      | 3          |
| <b>Total</b> | <b>142</b> |

The corresponding pie chart is shown below. The full circle in the pie chart represents 100% of the data. The segments represent the proportions of each group of the dataset. When drawing these by hand, you need to work out the angle subtended for each category, as a proportion of the  $360^\circ$  in the circle as a whole. For example, for ‘cars’ we would have

$$\text{Angle subtended} = \frac{76}{142} \times 360^\circ = 192.7^\circ.$$

### 2.2.3 Histograms

Bar charts have their limitations; for example, they cannot be used to present continuous data. When dealing with continuous random variables a different kind of graph is required. One such graph is 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 rectangles that is proportional to their frequency.

The frequency table for the data on service times for a budget airline call centre (see Section 2.1.3) was:

| Service time                 | Frequency |
|------------------------------|-----------|
| $175 \leq \text{time} < 180$ | 1         |
| $180 \leq \text{time} < 185$ | 3         |
| $185 \leq \text{time} < 190$ | 3         |
| $190 \leq \text{time} < 195$ | 6         |
| $195 \leq \text{time} < 200$ | 10        |
| $200 \leq \text{time} < 205$ | 12        |
| $205 \leq \text{time} < 210$ | 8         |
| $210 \leq \text{time} < 215$ | 3         |
| $215 \leq \text{time} < 220$ | 3         |
| $220 \leq \text{time} < 225$ | 1         |
| <b>Total</b>                 | 50        |

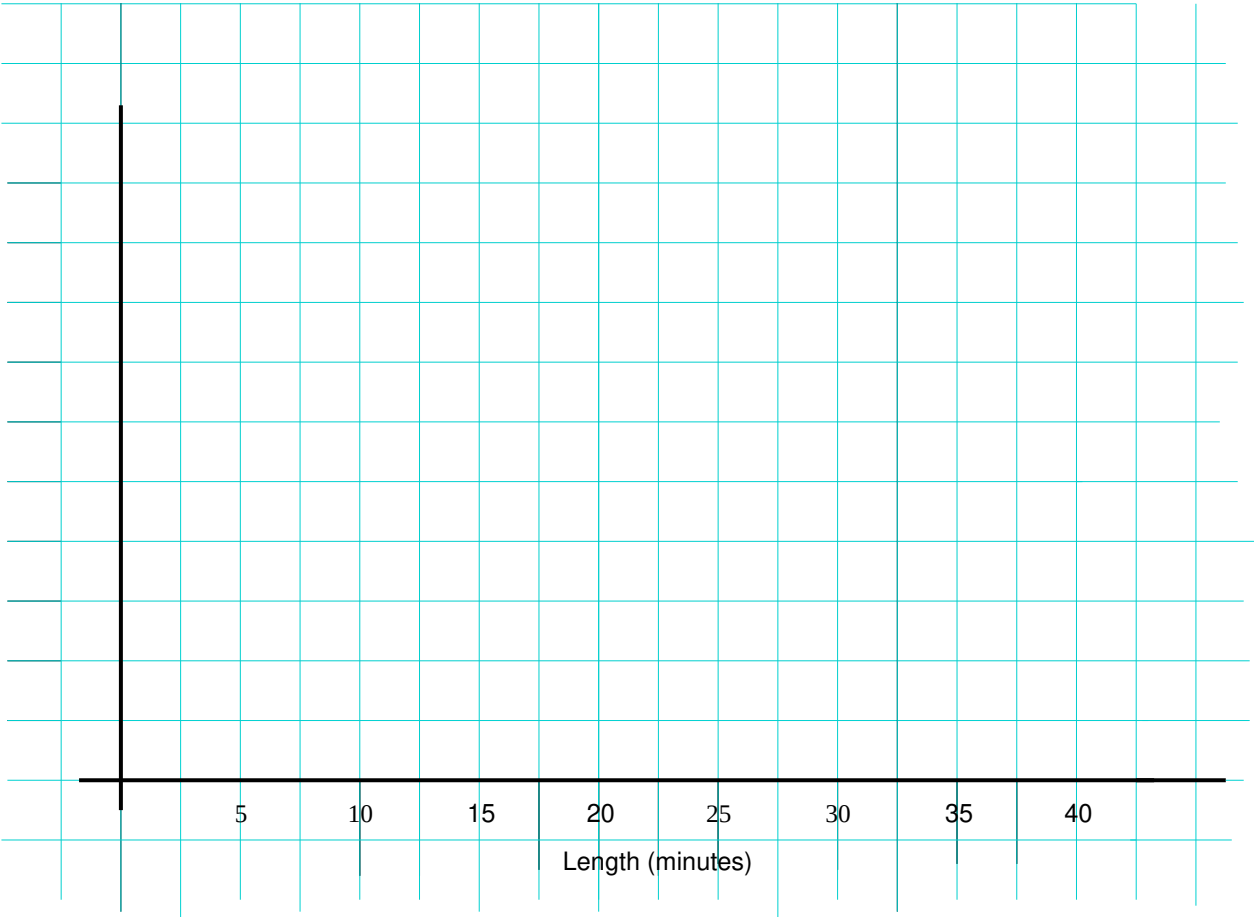
Notice that all the class intervals are the same width, and so the histogram for these data is:

Example

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

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

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

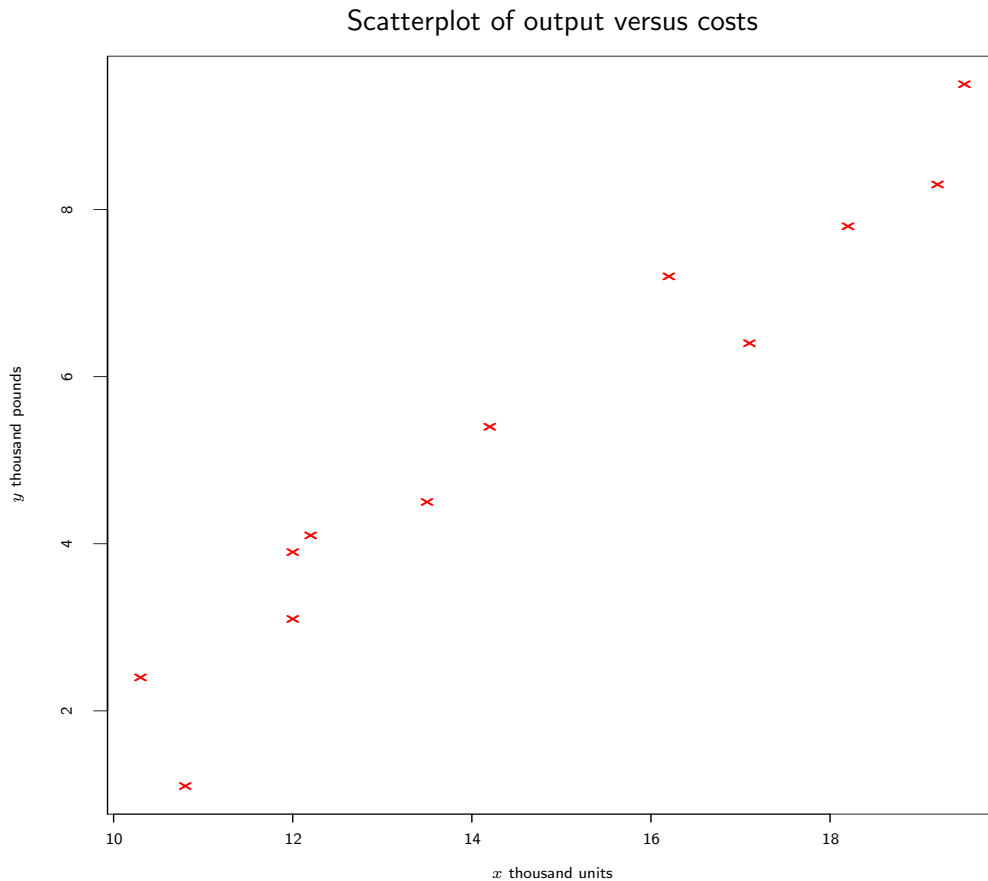


### 2.2.4 Scatter plots

Scatter plots are used to plot two variables which you believe might be related, for example, height and weight, advertising expenditure and sales, or age of machinery and maintenance costs. Consider the following data for monthly output ( $X$  thousand units) and total costs ( $Y$  thousand pounds) at *Cycle Centre UK*.

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

If you were interested in the relationship between the cost of production and the number of units produced you could easily plot this by hand, using the  $x$  and  $y$  values as pairs of co-ordinates; for example, you would plot the first pair as a point on a graph using the co-ordinates (10.3, 2.4) – the point would be placed at “10.3 along, 2.4 up”. The other points would be plotted in the same way, and can be seen in the scatterplot below.



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. A downwards sloping trend would imply that as output increased, total costs declined, an unlikely scenario. In the next section, we will look at how to *quantify* such relationships, where a *linear* relationship seems apparent.

### 3 Numerical summaries of data

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

#### 3.1 Mathematical notation

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

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

|                 |       |       |       |
|-----------------|-------|-------|-------|
| 1st sample:     | 1     | 5     | 7     |
| 2nd sample:     | 2     | 0     | 3     |
| ⋮               | ⋮     | ⋮     | ⋮     |
| typical sample: | $x_1$ | $x_2$ | $x_3$ |

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

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

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

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

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

## 3.2 Measures of average

These are also referred to as measures of *location* or, more commonly, just *averages*. In general terms, they tell us the value of a ‘typical’ observation. There are three measures which are commonly used: the (arithmetic) *mean*, the *median*, and the *mode*. We will consider these in turn. When these are obtained from a sample of data, rather than the entire population, we often prefix with the word ‘sample’.

### 3.2.1 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  $\bar{x}$  (“ $x$  bar”). In general, the mean is calculated using the formula

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

or equivalently as

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

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

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

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

The mean number of cars sold per day is

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

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                   |
| <b>Total</b> ( $n$ )    | 14                  |

The sample mean can be calculated from these data as

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

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

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

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

### 3.2.2 The median

The median is occasionally used instead of the mean, particularly when the data have an asymmetric profile or when there are outliers (as indicated by graph, perhaps). 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. For larger data sets, the calculation is more easily done using a computer package such as R.

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

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

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

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

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

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

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

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

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

### 3.2.3 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. For example, in a survey of 12 students, 4 said they read the *Metro* newspaper, 5 said they read *The Sun* and 3 said they read *The Times*. Thus, the mode is *The Sun*, as it is the most popular newspaper. It is possible to refer to modal classes with grouped data. This is simply the class with the greatest frequency of observations. For example, the modal class of

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

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

## 3.3 Measures of spread

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

### 3.3.1 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.

### 3.3.2 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:

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

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

Just as with the median, these quartiles might not correspond to actual observations. For example, in a dataset with  $n = 20$  values, the lower quartile is the  $5\frac{1}{4}$ th largest observation, that is, a quarter of the way between the 5th and 6th largest observations. This calculation is essentially the same process we used when calculating the median. Consider the following data, which show the distance (in km) cycled by 20 randomly selected students in a 30 minute period:

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

### 3.3.3 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. Strictly speaking, the sample variance measures deviation about a value calculated from the data (the sample mean) and so we use an  $n - 1$  divisor rather than  $n$ . That is, we use the formula

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

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

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

or equivalently as

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

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

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

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

**Example**

Consider again the cycling distance data; however two new students from the University's Athletics Club have now also taken part, with distances of 24.3km and 26.2km. The amended dataset is shown below, written in ascending order.

|     |     |     |      |      |      |      |      |      |      |      |
|-----|-----|-----|------|------|------|------|------|------|------|------|
| 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 | 24.3 | 26.2 |

- (a) Find the mean and standard deviation.
- (b) Why might the mean and standard deviation not be the most appropriate summaries to use here? Obtain more appropriate numerical summaries.



3.4 Box-and-whisker plots

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

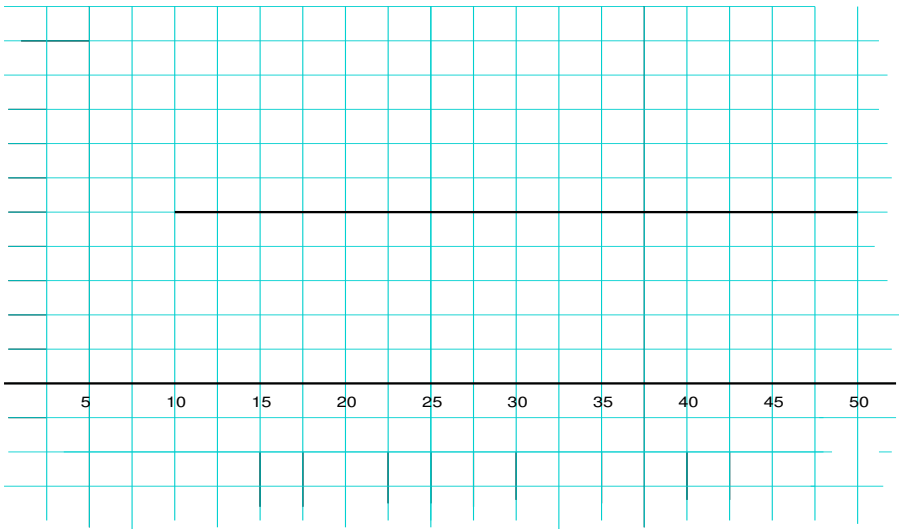
The plot is constructed as follows. After laying out an  $x$ -axis for the full range of the data, a rectangle is drawn with ends at the the upper and lower quartiles. The rectangle is split into two at the median. This is the “box”. Finally, lines are drawn from the box to the minimum and maximum values – these are the “whiskers”. Often computer packages will also display outliers on box-and-whisker plots – that is, points which seem extremely large or small, relative to the rest of the data.

Example

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

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

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

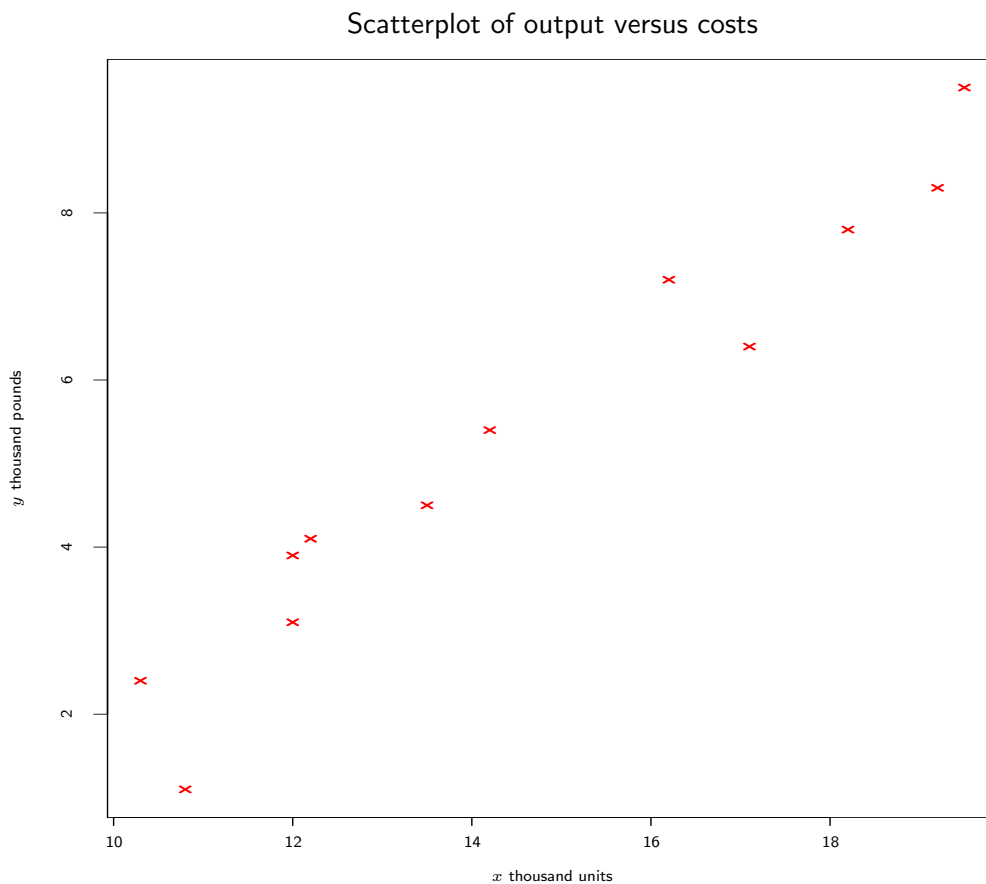


### 3.5 Correlation

In Section 2.2.4 we considered scatter plots as a useful graphical tool for identifying relationships between a pair of variables. When we have pairs of observations on each person/item in our sample, we say we have *bivariate* data, and the *correlation coefficient* is a useful tool for quantifying the degree of linear dependence between the two variables.

Recall the data from Section 2.2.4 on monthly output ( $X$  thousand units) and total costs ( $Y$  thousand pounds) at *Cycle Centre UK*, and the corresponding scatterplot:

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



There appears to be a linear relationship between  $x$  and  $y$  – that is, the points in the plot all lie close to a straight line. In fact, because the line would give an uphill slope as we look at the plot from left to right, we say that  $X$  and  $Y$  are *positively correlated*; a downhill slope would give a *negative* correlation. We might even go so far as to say that this positive correlation is *strong*, as it is clear that most of the points would lie fairly close to a straight line. The correlation coefficient, often abbreviated with the letter  $r$ , quantifies this linear association; the following sketches help illustrate its interpretation.



The correlation coefficient can be calculated using the following formula:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}},$$

where

$$\begin{aligned} S_{xy} &= \sum xy - n\bar{x}\bar{y}, \\ S_{xx} &= \sum x^2 - n\bar{x}^2 \quad \text{and} \\ S_{yy} &= \sum y^2 - n\bar{y}^2. \end{aligned}$$

This looks worse than it really is! In our example, we have

$$\bar{x} = \frac{10.3 + 12 + \dots + 19.2}{12} = 14.6 \quad \text{and} \quad \bar{y} = \frac{2.4 + 3.9 + \dots + 8.3}{12} = 5.31.$$

Also,

$$\begin{aligned} \sum xy &= (10.3 \times 2.4) + (12 \times 3.9) + \dots + (19.2 \times 8.3) = 1020.7 \\ \sum x^2 &= 10.3^2 + 12^2 + \dots + 19.2^2 = 2678.44 \\ \sum y^2 &= 2.4^2 + 3.9^2 + \dots + 8.3^2 = 410.79. \end{aligned}$$

This gives

$$\begin{aligned} S_{xy} &= 1020.7 - 12 \times 14.6 \times 5.31 = 90.388 \\ S_{xx} &= 2678.44 - 12 \times 14.6^2 = 120.52 \\ S_{yy} &= 410.79 - 12 \times 5.31^2 = 72.4368, \end{aligned}$$

giving

$$r = \frac{90.388}{\sqrt{120.52 \times 72.4368}} = 0.967.$$

This is as we'd expect, since the scatterplot shows a strong, positive linear relationship between  $X$  and  $Y$ .

## General exercises: sections 1→3

1. Identify the type of data described in each of the following examples:
  - (a) An opinion poll was taken asking people which party they would vote for in a general election.
  - (b) In a steel production process the temperature of the molten steel is measured and recorded every 60 seconds.
  - (c) A market researcher stops you in Northumberland Street and asks you to rate between 1 (disagree strongly) and 5 (agree strongly) your response to opinions presented to you.
  - (d) The hourly number of units produced by a beer bottling plant is recorded.
  
2. Rashid, Ellen and Loucas all work as business analysts for a leading clothing manufacturer. The design team want to know how successful a new pair of jeans will be, and so they ask the analysts to conduct some research.
  - (a) Loucas is lazy. He goes home that night and asks his four flatmates to give the new jeans a score from 1–5, and then reports his findings back to the design team the next day.  
What sort of sampling scheme has Loucas adopted, and why might it be flawed?
  - (b) Ellen thinks she will try to collect a simple random sample of people who might wear the jeans; she will then try to elicit their opinions. Why might a simple random sample be difficult to obtain here?
  - (c) Rashid is more thoughtful than both Loucas and Ellen. He thinks about the target population for the new jeans: females in the age group 21–30. He then spends a full day walking up and down Northumberland Street, stopping people he thinks fulfil this criteria and asking them to decide whether or not they would buy the new jeans.  
What sort of sampling scheme has Rashid adopted?
  
3. The following table shows daily traffic flow (thousands of vehicles) for a section of dual carriageway in South Tyneside. The data were recorded over a two week period from September 2nd 2013.

|       |      |       |       |       |       |       |
|-------|------|-------|-------|-------|-------|-------|
| 10.04 | 9.38 | 9.75  | 11.23 | 10.80 | 10.45 | 10.00 |
| 12.48 | 9.80 | 10.46 | 10.05 | 10.32 | 10.70 | 10.50 |

- (a) Find the mean daily traffic flow. Also find the standard deviation.
- (b) Find the median daily traffic flow. Also find the interquartile range.
- (c) Construct a box-and-whisker plot, and comment.

4. Consider the following data for daily ticket sales (in €) at a selection of *Deutsche Bahn* ticket offices in Germany, before and after a local advertising campaign.

| Daily Sales                      | Before | After |
|----------------------------------|--------|-------|
| $1000 \leq \text{sales} < 2000$  | 10     | 7     |
| $2000 \leq \text{sales} < 3000$  | 30     | 10    |
| $3000 \leq \text{sales} < 4000$  | 40     | 25    |
| $4000 \leq \text{sales} < 5000$  | 20     | 35    |
| $5000 \leq \text{sales} < 6000$  | 15     | 37    |
| $6000 \leq \text{sales} < 7000$  | 12     | 40    |
| $7000 \leq \text{sales} < 8000$  | 10     | 20    |
| $8000 \leq \text{sales} < 9000$  | 8      | 10    |
| $9000 \leq \text{sales} < 10000$ | 0      | 5     |
| <b>Totals</b>                    | 145    | 189   |

- (a) Calculate the percentage relative frequency for before and after.
- (b) Plot the relative frequency histograms for before and after, and comment.
5. A market researcher asked 650 students what their favourite daily newspaper was. The results are summarised below. Represent these data in an appropriate graphical manner.

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

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

30 28 27 33 35 32 25 27 30 1

- (a) Estimate the average number of passenger planes arriving, per hour, using (i) the mean and (ii) the median.
- (b) Summarise the spread of these data by calculating (i) the range; (ii) the inter-quartile range and (iii) the standard deviation.
- (c) Runway East 2 was closed for most of the final hour during the observation period due to a snowstorm. Given this information, comment on the dataset above and suggest which of your summaries of location and spread would be the most suitable. Which would be the *least* suitable? Why?
7. The following observations show average observed vehicular speed ( $X$  mph) and the number of road traffic casualties ( $Y$ ) at 6 potential accident 'blackspots' in Tyne and Wear over the period of a year. Obtain the correlation coefficient for these data.

|     |      |      |      |      |      |      |
|-----|------|------|------|------|------|------|
| $x$ | 42.3 | 30.1 | 67.7 | 52.4 | 33.4 | 72.8 |
| $y$ | 1    | 1    | 2    | 4    | 0    | 7    |

## 4 Hypothesis tests

### 4.1 Introduction and motivating example

Karlsruhe, in Germany, has a fully integrated public transport system consisting of several tram routes, many local bus services and local and inter-city train services. The authorities are interested in the viability of two tram routes in the city: routes M4 and M7.

To maintain these routes, authorities suggest a target of €2150 per day, on average, in ticket sales. For 30 randomly selected days in 2013, daily ticket sales on line M4, in €, were:

|      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|
| 2182 | 2083 | 1419 | 2544 | 1666 | 2197 | 2000 | 2182 | 2450 | 1982 |
| 2213 | 2162 | 2201 | 1925 | 1317 | 2423 | 2271 | 1869 | 2108 | 1609 |
| 1895 | 2740 | 1325 | 2515 | 2089 | 2344 | 2047 | 2251 | 1351 | 1503 |

Should the authorities consider withdrawing tram service M4?

Clearly, revenue will vary day-by-day. Some days, revenue might greatly exceed the authorities' target (e.g. day 22 above has a revenue of €2740); however, on other days revenue might fall way short (e.g. on day 15 revenue dropped to just €1317). To get an overall impression of the situation of line M4, we might work out *average* revenue for our sample above:

$$\bar{x} = \frac{2182 + 2083 + \dots + 1503}{30} = \text{€}2028.77,$$

a little short of the authorities' target (incidentally, the standard deviation is  $s = 380.22$  – try this yourself!). But these 30 observations represent a very small sample from the population as a whole, and the value of  $\bar{x}$  is bound to change for different samples. We need to account for this *sampling variability* of the mean  $\bar{x}$ , and the most common way of doing this is to perform a hypothesis test.

### 4.2 The one-sample $t$ test

#### 4.2.1 The basics

The one-sample  $t$ -test can be useful when we are interested in how the mean of a set of sample observations compares to some target value. The mean in our sample, as always, is denoted by  $\bar{x}$ . The standard notation in Statistics for the *population* mean is the Greek symbol  $\mu$  (pronounced “mu”). Obviously,  $\bar{x}$  is our sample estimate of  $\mu$ .

In the example above about trams in Karlsruhe, we'd like to know whether or not the target daily revenue of €2150 is being met. Hence, we'd like to know how likely it is that  $\mu = 2150$ , and this assertion is made in the *null hypothesis*, denoted by  $H_0$  and often written down as follows:

$$H_0 : \mu = 2150.$$

We usually test against a general *alternative hypothesis*  $H_1$ , in this case written down as

$$H_1 : \mu \neq 2150,$$

which says “mu is *not* equal to €2150”.

When performing the hypothesis test, we assume  $H_0$  to be true. We then ask ourselves the question:

“How likely is it that we would observe the data we have, or indeed anything more extreme than this, if the null hypothesis is true?”

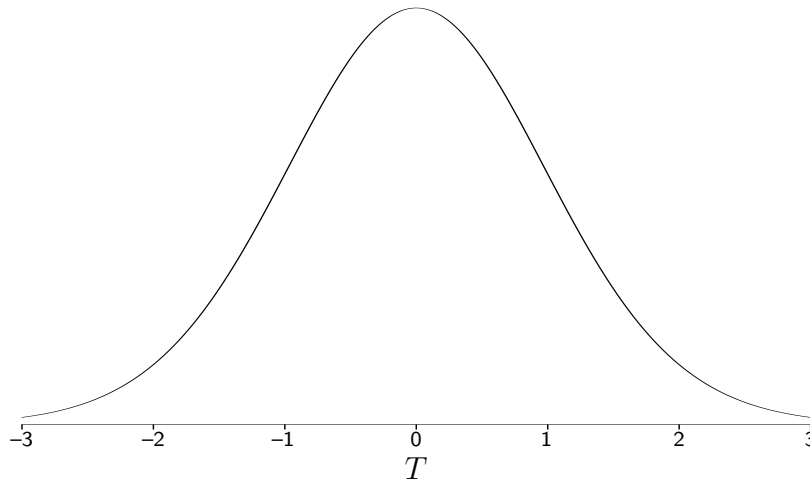
We can get a handle on this question thanks to the *Central Limit Theorem* in Statistics. Although we will not go into the details here, this result tells us that the quantity

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows a known statistical distribution: Student’s  $t$  distribution, hence the name of this particular hypothesis test. In this formula:

- $\bar{x}$  is our sample mean
- $\mu$  is the assumed value of the population mean under  $H_0$
- $s$  is the standard deviation from our sample
- $n$  is the sample size

Student’s  $t$  distribution has the following shape, centred around zero:

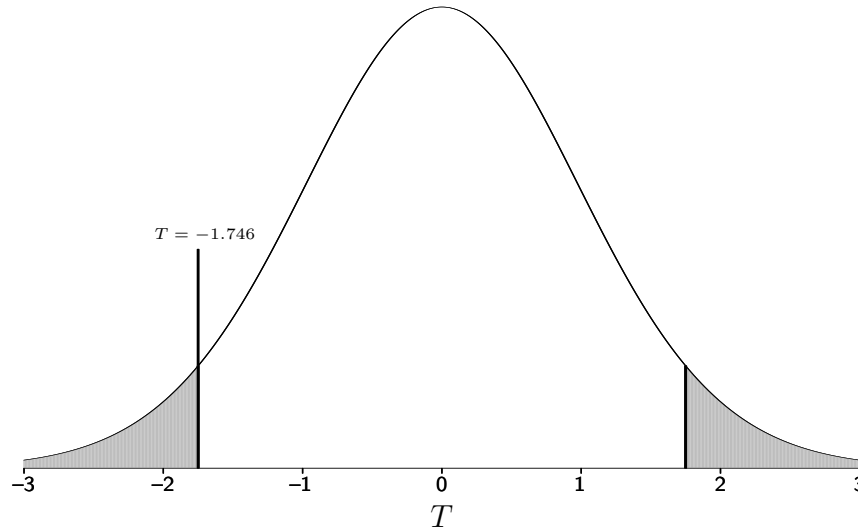


The area under this curve is 1, since this is a *probability density function*. Clearly, the most likely value to occur here is 0, as this is where the density function peaks; we are less likely to see values that are further away from zero than values which are closer to zero. If the null hypothesis is true, then  $\mu = 2150$ , and so our *test statistic*, using the sample data, is

$$T = \frac{2028.77 - 2150}{380.22/\sqrt{30}} = -1.746.$$

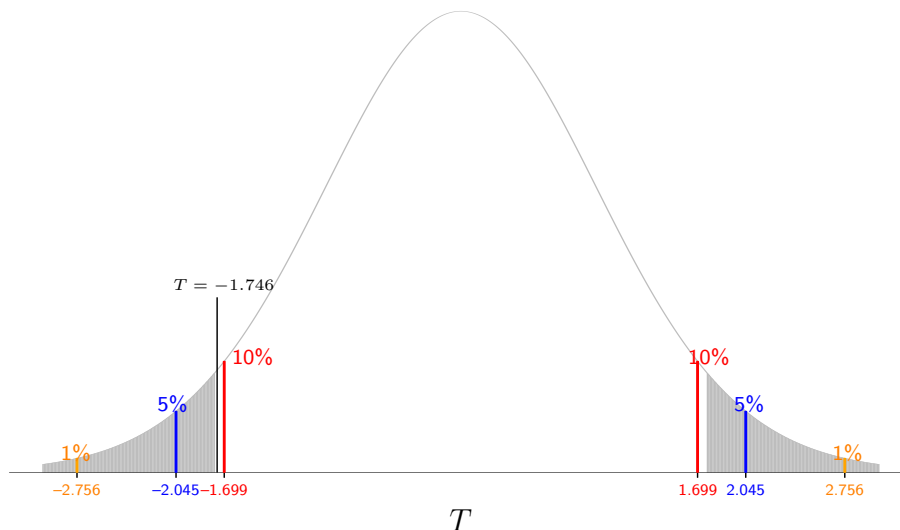
How plausible is it that this value comes from the distribution shown above? The further away it is from zero, the less plausible this is, and so the more evidence we have against the null hypothesis.

Since the  $t$  distribution is symmetric,  $T = 1.746$  is just as extreme as  $T = -1.746$ , and so the shaded region in the following diagram illustrates *the  $p$ -value* – the probability of observing the data we have, or anything more extreme than this, if the null hypothesis is true. In other words, the answer to our earlier question!



The closer the area of the shaded region (the  $p$ -value) is to 0%, the less plausible it is that we would observe the data we have if the null hypothesis is true, that is, the more evidence we have to reject  $H_0$ . So, we need to work out the area of the shaded region under the curve in the diagram above. Luckily for us, standard statistical tables (see page 31) can be used to help us figure this out.

For example, the critical values in Student's  $t$  distribution which would give a shaded area of 10% in our example would be  $\pm 1.699$ : we are performing a 'two-tailed test' here (since our alternative hypothesis is  $\neq$ ); in Student's  $t$  distribution,  $\nu = n - 1$  which gives  $30 - 1 = 29$  in this example; thus, according to the table, the corresponding critical value is 1.699. In fact, for our example we are able to update our diagram to give:



So, we can see from our latest diagram that the shaded area is somewhere between 10% and 5% – in other words, our  $p$ -value is between 5 and 10 percent. Earlier, we said that the smaller this  $p$ -value is, the more evidence we have to reject  $H_0$ . The question now, is:

“What constitutes a  $p$ -value small enough to reject  $H_0$ ?”

The convention (but by no means a hard-and-fast cut-off) is to reject  $H_0$  if the  $p$ -value is smaller than 5%. Thus, here we would say:

- Our  $p$ -value is greater than 5% (in fact, it’s somewhere between 5% and 10% – a computer can tell us that it’s exactly 9.132%)
- Thus, we do *not* reject  $H_0$  – in fact, we *retain*  $H_0$
- There is insufficient evidence to suggest a real deviation from the proposed population mean of €2150

Care should be taken not be too strong in our conclusions. For example, we should not say “We have proven that the population mean is €2150”, or even “We have proven that tram line M4 should not be withdrawn”. Rather, we can say something like “From our sample, there is not enough evidence to suggest that tram line M4 should be withdrawn”. If we take a larger sample, our conclusions might change! More generally,

*Absence of evidence is not evidence of absence*

### Summary: One-sample $t$ test

1. Formulate your null hypothesis ( $H_0$ ) and your alternative hypothesis ( $H_1$ ):

$$H_0 : \mu = \text{target or proposed value}$$

$$H_1 : \mu \neq \text{target or proposed value}$$

2. Calculate your test statistic:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

3. Use your test statistic, and statistical tables, to obtain a range for your  $p$ -value. Remember – the row to use in the table is  $\nu = n - 1$ . A diagram will also help here!
4. Form your conclusion – the convention is to reject  $H_0$  if your  $p$ -value is less than 5%.

|                       | One-tailed test | 10%      | 5%       | 2.5%     | 1%       | 0.5%     |
|-----------------------|-----------------|----------|----------|----------|----------|----------|
|                       | Two-tailed test | 20%      | 10%      | 5%       | 2%       | 1%       |
|                       | 1               | 3.078    | 6.314    | 12.706   | 31.821   | 63.657   |
|                       | 2               | 1.886    | 2.920    | 4.303    | 6.965    | 9.925    |
|                       | 3               | 1.638    | 2.353    | 3.182    | 4.541    | 5.841    |
|                       | 4               | 1.533    | 2.132    | 2.776    | 3.747    | 4.604    |
|                       | 5               | 1.476    | 2.015    | 2.571    | 3.365    | 4.032    |
|                       | 6               | 1.440    | 1.943    | 2.447    | 3.143    | 3.707    |
|                       | 7               | 1.415    | 1.895    | 2.365    | 2.998    | 3.449    |
| $\nu = n - 1$         | 8               | 1.397    | 1.860    | 2.306    | 2.896    | 3.355    |
| (one-sample test)     | 9               | 1.383    | 1.833    | 2.262    | 2.821    | 3.250    |
|                       | 10              | 1.372    | 1.812    | 2.228    | 2.764    | 3.169    |
| $\nu = n_1 + n_2 - 2$ | 11              | 1.363    | 1.796    | 2.201    | 2.718    | 3.106    |
| (two-sample test)     | 12              | 1.356    | 1.782    | 2.179    | 2.681    | 3.055    |
|                       | 13              | 1.350    | 1.771    | 2.160    | 2.650    | 3.012    |
|                       | 14              | 1.345    | 1.761    | 2.145    | 2.624    | 2.977    |
|                       | 15              | 1.341    | 1.753    | 2.131    | 2.602    | 2.947    |
|                       | 16              | 1.337    | 1.746    | 2.120    | 2.583    | 2.921    |
|                       | 17              | 1.333    | 1.740    | 2.110    | 2.567    | 2.898    |
|                       | 18              | 1.330    | 1.734    | 2.101    | 2.552    | 2.878    |
|                       | 19              | 1.328    | 1.729    | 2.093    | 2.539    | 2.861    |
|                       | 20              | 1.325    | 1.725    | 2.086    | 2.528    | 2.845    |
|                       | 21              | 1.323    | 1.721    | 2.080    | 2.518    | 2.831    |
|                       | 22              | 1.321    | 1.717    | 2.074    | 2.508    | 2.819    |
|                       | 23              | 1.319    | 1.714    | 2.069    | 2.500    | 2.807    |
|                       | 24              | 1.318    | 1.711    | 2.064    | 2.492    | 2.797    |
|                       | 25              | 1.316    | 1.708    | 2.060    | 2.485    | 2.787    |
|                       | 26              | 1.315    | 1.706    | 2.056    | 2.479    | 2.779    |
|                       | 27              | 1.314    | 1.703    | 2.052    | 2.473    | 2.771    |
|                       | 28              | 1.313    | 1.701    | 2.048    | 2.467    | 2.763    |
|                       | 29              | 1.311    | 1.699    | 2.045    | 2.462    | 2.756    |
|                       | $\vdots$        | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
|                       | $\infty$        | 1.282    | 1.645    | 1.960    | 2.326    | 2.576    |

*Table of critical values from Student's  $t$  distribution.*

**Example**

Daily ticket sales for 1st June 2013–30th June 2013 on tram line M7 in Karlsruhe are shown below.

|      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|
| 1405 | 1761 | 1912 | 2327 | 2000 | 2368 | 2282 | 2295 | 1675 | 1287 |
| 1968 | 1756 | 1687 | 1848 | 1967 | 1879 | 1924 | 1919 | 1626 | 1488 |
| 1724 | 1852 | 1234 | 1668 | 1915 | 1244 | 2086 | 1319 | 1329 | 1909 |

The corresponding sample mean and standard deviation are €1788.47 and €322.42, respectively. Perform a two-tailed hypothesis test to investigate the financial viability of tram route M7. Work at the 5% level of significance.



### 4.2.2 Further considerations

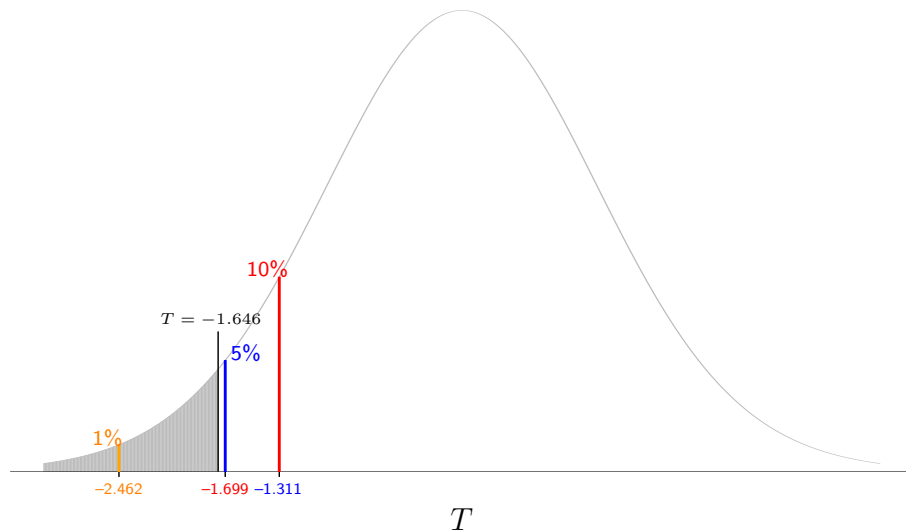
#### *One-tailed tests*

Occasionally, we might perform a ‘one-tailed test’. In the context of the one-sample  $t$  test, this would require a change to the alternative hypothesis. If there is an *a priori* belief that the population mean might be less than, or perhaps greater than, the target/proposed value, we could then specify a one-tailed alternative hypothesis as opposed to the general two-tailed alternative set up in the previous section.

For example, the authorities might believe that tram line M4 is not reaching the target of €2150. Thus, prior to any data collection, we might specify the following alternative hypothesis:

$$H_1 : \mu < 2150,$$

which reads as “mu is less than 2150”. The null hypothesis stays the same; as does the calculation of the test statistic. The only thing that changes is the way in which we obtain the  $p$ -value. Now, we only consider the area under the curve to the *left* of the test statistic  $T$ . When using tables to obtain the range for our  $p$ -value, we use the columns for ‘One-tailed test’ instead of ‘Two-tailed test’, giving the following updated diagram:



Now we see that:

- Our  $p$ -value lies between 1% and 5% (in fact, a computer tells us that it's exactly 4.566%)
- We now *reject*  $H_0$  in favour of  $H_1$
- There is evidence to suggest that the population mean might be less than the target of €2150

Now we might conclude that there is evidence to suggest the withdrawal of tram line M4.

So how do we know if we should use a one-tailed test or a two-tailed test? If you perform a one-tailed test, you are completely disregarding the possibility of an effect in the other direction. In our tram example, we are not allowing for the possibility that average ticket sales on line M4 might *exceed* the target – this could be the wrong thing to do given that on almost half of the days in our sample, ticket sales do indeed exceed €2150. If you don't care about missing such an effect, then a one-tailed test might be OK. However, careful consideration should be given to the consequences of effects in the 'opposite direction'. For example, average daily ticket sales exceeding our target value could mean services are too busy. *The default position is to use a two-tailed test, unless there are compelling reasons against this. In any case, this decision should always be made before looking at the data.*

### Errors

A *Type I Error* occurs when the null hypothesis is true but is wrongly rejected. This is often referred to as a 'false hit', or a false positive (e.g. when a diagnostic test indicates the presence of a disease, when in fact the patient does not have the disease).

The rate of a Type I Error is known as the *size* of the test and is usually denoted by the Greek symbol  $\alpha$ , pronounced "alpha", and usually equals the significance level of the test – that is, the *p*-value beyond which we have decided to reject  $H_0$  (e.g. 5% or 0.05 in our earlier examples).

A *Type II Error* occurs when the null hypothesis is false, but erroneously fails to be rejected. Hence, we fail to assert what is present, and so this is often referred to as a 'miss'.

The rate of the Type II Error is usually denoted by the Greek symbol  $\beta$ , pronounced "beta", and is related to the *power* of a test (which equals  $1 - \beta$ ).

### Power

The *power* of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true; thus, the greater the power, the better! A *power analysis* can be readily performed on a computer, prior to any data collection, to indicate the sample size necessary to achieve adequate power – the standard for which is usually set at around 0.8 or 80%. An under-powered study can often allow different conclusions to be reached depending on the significance level chosen (e.g. 1% or 5%), which might be undesirable.

## 4.3 The two-sample *t* test

Suppose now that the authorities in Karlsruhe want to know if there's any real difference between average daily ticket sales on tram lines M4 and M7. We can never know the answer for certain, as the data we have will likely always be from *samples* of daily ticket sales – that is, the true population mean ticket sales for the two lines will be unknown. However, as with the one-sample *t* test, we can use information given in sample data to infer things about the population as a whole.

As with the one-sample test, we must start by setting up our hypotheses. In a two-sample *t* test, the null hypothesis is always that the population means for the two groups are the same.

Thus, here we might write

$$H_0 : \mu_{M4} = \mu_{M7},$$

where  $\mu_{M4}$  and  $\mu_{M7}$  represent the population mean daily ticket sales for tram lines M4 and M7, respectively. The general two-tailed alternative would be

$$H_1 : \mu_{M4} \neq \mu_{M7},$$

although, as discussed earlier, we could opt for one-tailed alternatives – for example,

$$H_1 : \mu_{M4} < \mu_{M7} \quad \text{or}$$

$$H_1 : \mu_{M4} > \mu_{M7}.$$

To allow for both possibilities, the usual approach is to adopt a general two-tailed alternative.

The test statistic for the two-sample  $t$  test is given by

$$T = \frac{\bar{x}_{M4} - \bar{x}_{M7}}{s \times \sqrt{1/n_{M4} + 1/n_{M7}}},$$

where

- $\bar{x}_{M4}$  and  $\bar{x}_{M7}$  are the mean daily ticket sales for the two tram lines
- $n_{M4}$  and  $n_{M7}$  are the sample sizes (which don't have to be equal!)
- $s$  is the *pooled standard deviation*, given in terms of the two sample standard deviations for tram lines M4 and M7,  $s_{M4}$  and  $s_{M7}$  (respectively):

$$s = \sqrt{\frac{(n_{M4} - 1)s_{M4}^2 + (n_{M7} - 1)s_{M7}^2}{n_{M4} + n_{M7} - 2}}.$$

Of course, more generally, we can replace the labels M4 and M7 with anything – usually, they are just written using the labels “1” and “2”.

In our example, we have

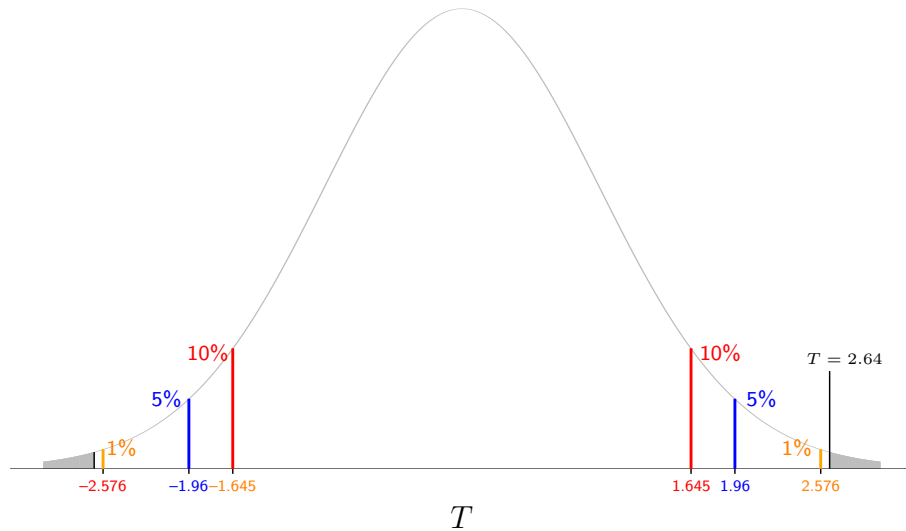
$$\begin{aligned} s &= \sqrt{\frac{(30 - 1) \times 380.22^2 + (30 - 1) \times 322.42^2}{30 + 30 - 2}} \\ &= \sqrt{\frac{7207135.239}{58}} = 352.507. \end{aligned}$$

Now we can compute the test statistic:

$$T = \frac{2028.77 - 1788.47}{352.507 \times \sqrt{1/30 + 1/30}} = 2.640.$$

As in the one-sample test, the theory tells us that  $T$  also follows Student's  $t$  distribution, but this time we have  $\nu = n_{M4} + n_{M7} - 2$ , which gives 58 in our example.

We now look at tables to obtain our critical values from Student's  $t$  distribution. You'll notice that we don't have a row for  $\nu = 58$ ; for values of  $\nu \geq 30$ , we use the  $\infty$  row at the bottom of the table, giving us the following diagram:



Locating our test statistic  $T = 2.64$  on this diagram reveals a  $p$ -value smaller than 1%. If we work at the 5% level of significance, as usual, then

- We would reject  $H_0$  and accept the alternative  $H_1$
- There is evidence to suggest a real difference in average daily ticket sales between the two tram lines
- Since  $\bar{x}_{M4} > \bar{x}_{M7}$ , we might conclude that daily ticket sales are significantly higher from line M4 than line M7, i.e. line M4 is significantly more profitable.

### Summary: Two-sample $t$ test

1. Formulate your null hypothesis ( $H_0$ ) and your alternative hypothesis ( $H_1$ ):

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

2. Calculate your test statistic:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s \times \sqrt{1/n_1 + 1/n_2}},$$

where

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

3. Use your test statistic, and statistical tables, to obtain a range for your  $p$ -value. Remember – the row to use in the table is  $\nu = n_1 + n_2 - 2$ . A diagram will also help here!
4. Form your conclusion – the convention is to reject  $H_0$  if your  $p$ -value is less than 5%.

## 4.4 Assumptions

In the hypothesis tests covered so far the following are assumed:

1. Our sample(s) are drawn from a Normal distribution – that is, the population(s) have a bell-shaped profile similar to those shown in this section for Student's  $t$  distribution. This assumption can be checked by using various graphical procedures – at a very basic level, a histogram of the sample data can be inspected. In fact,  $t$  tests are rather robust to departures from Normality, although if the data indicate severe non-normality, transformations can be applied, or even *non-parametric tests* used which have no distributional assumptions (e.g. the Wilcoxon test).
2. Data within our sample(s) are *independent* – that is, there is no association between any within-sample observations and so we can pretty much treat our observations as having occurred ‘randomly’. For the data collected for tram line **M4** this might be a reasonable assumption, as we were told that the data were obtained from “... 30 randomly selected days in 2013”. However, this is not the case for the data for tram line **M7**: the 30 observations were taken from consecutive days in June 2013. A spell of warm weather could encourage more people to leave their cars at home and use public transport, giving rise to higher-than-normal takings on this tram line for several consecutive days. Indeed, this could well be the case – 5 of the largest observations occur on the dates June 4th–June 8th (inclusive).
3. In a two-sample  $t$  test, it is assumed that the two populations are independent of each other. For example, had the data from the two tram lines been observed on exactly the same days, then the assumption of independent populations could be violated: observations from the two samples now essentially appear in pairs, linked by day, and so could depend on one another. In fact, Karlsruhe has an annual “leave your car at home” day on June 21st; thus, we might expect users of both tram lines to increase simultaneously on this day. When there is a natural pairing between observations in our two samples, a *paired comparisons  $t$  test* should instead be used (see the next section).
4. In a two-sample  $t$  test, it is also assumed that the two population standard deviations are equal. Of course, we usually don't know the value of the population standard deviations. However, the *sample* standard deviations provide estimates of these quantities. Many different methods have been proposed for comparing these sample standard deviations – a rough-and-ready check is that one is no more than twice the other. This seems to be the case in our worked example in the last section:  $s_{M4} = 380.22$  and  $s_{M7} = 322.42$ . If we cannot assume equal population standard deviations, a modified version of the test, called a *Welch's  $t$  test*, is available.

## 4.5 Paired comparisons $t$ test

Average vehicular speed (in mph) at 10 known accident blackspots in Newcastle was recorded over a two week period before, and after, the implementation of a speed camera; the results are shown in the table below.

|        | Site number |      |      |      |      |      |      |      |      |      |
|--------|-------------|------|------|------|------|------|------|------|------|------|
|        | 1           | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
| Before | 36.2        | 33.3 | 29.7 | 29.9 | 34.1 | 38.6 | 27.0 | 29.9 | 35.1 | 33.0 |
| After  | 34.1        | 33.2 | 30.0 | 26.5 | 29.1 | 36.1 | 22.4 | 29.6 | 30.1 | 30.1 |

Have the speed cameras been effective in reducing speed at these blackspots?

You might be tempted to say “yes” – at 9 of the 10 sites, average speed is lower in the after period than in the before period. However, is this a real reduction, or just random fluctuation? In other words, is the reduction *significant*?

We can summarise the two samples numerically, using the labels B and A for before and after, respectively:

$$\bar{x}_B = 32.68 \quad s_B = 3.53$$

$$\bar{x}_A = 30.12 \quad s_A = 3.87$$

It would now be wrong to use a two-sample  $t$  test to compare before and after, as assumption (3) listed in Section 4.4 is violated: the before and after values for each site form pairs of observations, linked by the site itself, and so we cannot assume independence between our two populations. For example, the speed limit at sites 1 and 6 is 40mph; therefore, we might always expect these two sites to observe higher speeds than the rest, regardless of the period (before/after). More generally, we might expect a reduction in speed after speed cameras have been installed – but the reduction in speed might be different at each site, depending on other site-specific factors.

The way forward here is to perform a *paired comparisons  $t$  test*.

We can set up our null and alternative hypotheses as in the two-sample  $t$  test:

$$H_0 : \mu_B = \mu_A$$

$$H_1 : \mu_B \neq \mu_A.$$

We *could* use a one-tailed alternative to check whether not there is evidence for a reduction in average speed (probably the intention of the speed cameras), giving:

$$H_1 : \mu_B > \mu_A.$$

However, this will not allow us to check whether or not the speed cameras could have had the opposite effect – that is, an *increase* in average speed. Although the figures in the table above suggest this is unlikely, remember that the hypotheses should be set up before any data have been collected – and in this example it might make sense to check for both situations.

In a paired comparisons  $t$  test we now work with paired differences – that is, we calculate

$$d_i = x_{B,i} - x_{A,i}, \quad i = 1, \dots, 10,$$

and then calculate the following test statistic:

$$T = \frac{\bar{d}}{s_d/\sqrt{n}},$$

where  $\bar{d}$  and  $s_d$  are the mean and standard deviation for the paired differences  $d_i$ .

The differences in average speed for each site are:

$$\begin{aligned} d_1 &= 36.2 - 34.1 = 2.1 \\ d_2 &= 33.3 - 33.2 = 0.1 \\ d_3 &= 29.7 - 30.0 = -0.3 \\ &\vdots \\ d_{10} &= 33.0 - 30.1 = 2.9 \end{aligned}$$

Now the mean of the  $d_i$  is

$$\bar{d} = \frac{2.1 + 0.1 - 0.3 + \dots + 2.9}{10} = \frac{25.6}{10} = 2.56;$$

similarly, the standard deviation can be shown to be

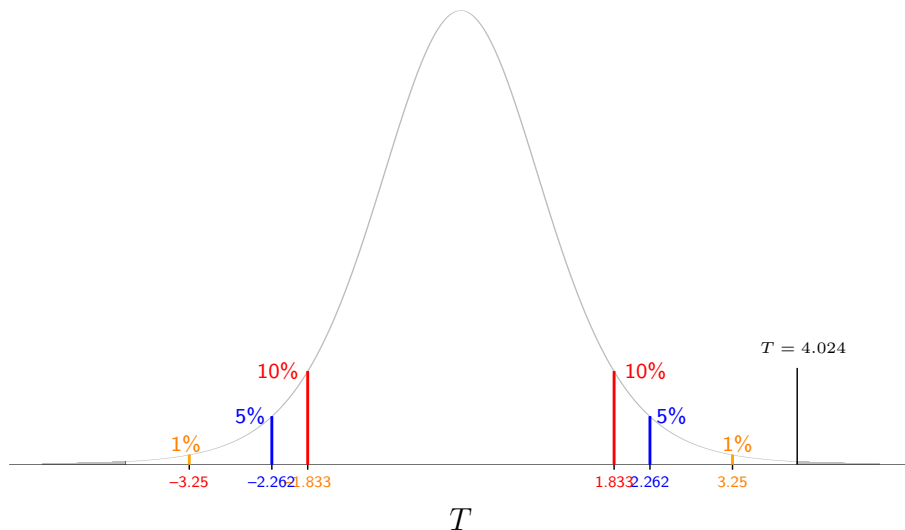
$$s_d = 2.012.$$

Therefore, the test statistic is

$$T = \frac{2.56}{2.012/\sqrt{10}} = 4.024.$$

Now that we are working with a single set of differences, rather than the two samples independently, this is essentially a one-sample problem; thus, we proceed to find our  $p$ -value in exactly the same way we did in the examples in Section 4.2, using  $\nu = n - 1$  in the  $t$  table, where  $n$  is the number of pairs (and so here we use  $\nu = 10 - 1 = 9$ ).

Since we are using a two-tailed alternative, we obtain the following diagram:



From this, it is clear that our  $p$ -value is less than 1%. So:

- We reject  $H_0$  and accept the alternative  $H_1$
- There is evidence to suggest a real difference in average speed between the before and after periods
- Since  $\bar{x}_A < \bar{x}_B$ , we might conclude that, since the introduction of speed cameras, average vehicular speed at these blackspot sites has significantly reduced

### *A note of caution*

From this, we might be tempted to say that “speed cameras have reduced speed”; those in favour of speed cameras have been known to take this further, and make claims such as “speed cameras save lives”. However, there is no implication of *causation* here – other factors might have influenced motorists’ speed at the same time as the introduction of the speed cameras (e.g. “*Speed Kills*” advertising campaigns, diverted traffic etc.).

### **Summary: Paired comparisons $t$ test**

1. Use in place of a two-sample  $t$  test when the observations in the two samples are paired.
2. Formulate your null hypothesis ( $H_0$ ) and your alternative hypothesis ( $H_1$ ):

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

3. Calculate your test statistic:

$$T = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where:

- you first find the raw differences  $d_i$  between each pair
  - $\bar{d}$  is then the mean of these difference
  - $s_d$  is the standard deviation of these differences
  - $n$  is the number of pairs
4. Use your test statistic, and statistical tables, to obtain a range for your  $p$ -value. Remember – the row to use in the table is  $\nu = n - 1$ . A diagram will also help here!
  5. Form your conclusion – the convention is to reject  $H_0$  if your  $p$ -value is less than 5%.

Example

Seven married couples took part in a public transport satisfaction survey. The questions given to each participant in the survey resulted in an overall score out of 100 for their level of satisfaction with local public transport facilities. The results are shown in the table below.

|         | Couple |      |      |      |      |      |      |
|---------|--------|------|------|------|------|------|------|
|         | 1      | 2    | 3    | 4    | 5    | 6    | 7    |
| Husband | 51.3   | 68.3 | 39.8 | 75.2 | 69.9 | 81.5 | 68.2 |
| Wife    | 54.2   | 44.3 | 18.5 | 62.3 | 55.2 | 76.3 | 70.0 |

Perform an appropriate hypothesis test to see if there is a significant difference between male and female levels of satisfaction.



## 5 Confidence intervals

Confidence intervals provide an alternative to hypothesis tests for assessing questions about the population mean (or population means in two sample problems), although often they are used *alongside* hypothesis tests. Due to lack of time, we will not cover the general theory supporting the construction of confidence intervals, although this theory is the same as that discussed in Section 4 for hypothesis tests and revolves around the *Central Limit Theorem*. The aim is to use the information in our sample to obtain an interval within which we might expect the true population mean to lie, *with a specified level of confidence*.

Recall from Section 4 that the sample mean  $\bar{x}$  is an *estimate* of the population mean  $\mu$ . The population mean does exist; we'd like to know what it is, but usually this isn't possible (unless we are able to take a census). The next best thing is to take a sample from the population and find  $\bar{x}$  – if our sample is representative of the population as a whole, we might feel confident that  $\bar{x}$  is a good estimator for  $\mu$ , and might be close to the true value.

The problem is, if we were to take many samples from the population, and so calculate many  $\bar{x}$ 's, they are all likely to be different to each other. Which one would we trust the most? In reality, we don't have the resources to take many samples, and so we'd like a way of capturing the variability of  $\bar{x}$  based on the information in our single sample. This is where statistical theory can help us out!

### 5.1 Construction

Most confidence intervals are symmetric and two-sided. Thus, to obtain a confidence interval for the population mean we:

1. Find the mean in our sample,  $\bar{x}$
2. Subtract some amount from  $\bar{x}$  to obtain the *lower bound* of our confidence interval
3. Add the same amount in (2) to our sample mean  $\bar{x}$  to obtain the *upper bound* of our confidence interval

The amount we add and subtract from  $\bar{x}$  is a function of the sample standard deviation  $s$  and the sample size  $n$ ; it also depends on just how 'confident' we want to be of capturing the population mean  $\mu$  within our interval! To be 100% confident, the lower and upper bounds would need to extend to the smallest and largest values possible for our random variable, and so it makes no sense to find such an interval; the standard is to find a *95% confidence interval*.

The formula for a symmetric, two-sided 95% confidence interval for the population mean  $\mu$  is

$$\left( \bar{x} - t \times \frac{s}{\sqrt{n}}, \quad \bar{x} + t \times \frac{s}{\sqrt{n}} \right),$$

often condensed to just

$$\bar{x} \pm t \times \frac{s}{\sqrt{n}};$$

as in Section 4,  $t$  is a critical value from Student's  $t$  distribution and can be found from statistical tables.

Let's return to the example from Section 4 regarding daily ticket sales on tram line M4 in the city of Karlsruhe in Germany. Recall the following summaries:

$$\bar{x}_{M4} = \text{€}2028.77 \quad s_{M4} = \text{€}380.22 \quad n_{M4} = 30$$

Thus, the 95% confidence interval for the *population* mean daily revenue is:

$$2028.77 \pm t \times \frac{380.22}{\sqrt{30}}$$

$$2028.77 \pm t \times 69.418$$

So now we just need to find  $t$ !

- We are attempting to find a two-sided confidence interval (which is generally the case), and so this is equivalent to performing a two-tailed hypothesis test; hence, we are looking for a critical value in the table on page 31 relating to a “Two-tailed test”
- A 95% confidence interval is equivalent to testing a hypothesis at the 5% level of significance; hence, we use the “5%” column
- As in a one-sample  $t$  test,  $\nu = n - 1$ ; so in this example we use row 29 in the table

Combining the information in the bullet points above gives us the critical value

$$t = 2.045,$$

and so we can now finish off our calculations:

$$2028.77 \pm 2.045 \times 69.418 \longrightarrow (\text{€}1886.81, \text{€}2170.73).$$

### Example

Obtain a 95% confidence interval for daily ticket sales on tram line M7 in Karlsruhe.



## 5.2 Interpretation

Care should be taken when interpreting confidence intervals. Take the 95% interval for the population mean ticket sales on tram line M4:

$$(\text{€}1886.81, \text{€}2170.73).$$

The correct way to interpret this is to say “This interval will either capture the true population mean, or it will not. If we were to take 100 samples, and obtain 100 such confidence intervals, about 95 of these intervals would capture the true mean”. Notice that the interval above contains the authorities’ target value of €2150. Thus, we can also conclude that there is no evidence to suggest mean ticket sales for this tram line are any different to the target value; hence, there is no evidence to suggest this tram line isn’t economically viable. Notice that this ties in with our conclusions in the two-tailed hypothesis test, where we retained  $H_0$ .

How can you interpret your confidence interval for mean ticket sales on tram line M7?



For two samples from independent populations, we can check for significant differences in the two population means by checking to see if the corresponding intervals overlap. If they do not overlap, this could indicate a significant difference in the means of the two populations. Similarly, for paired data, we can construct a confidence interval for the differences  $d_i$  and check to see whether or not this includes zero.

Returning to the results of the transport satisfaction survey on page 41: recall we had the following set of differences  $d_i$  between scores for husband and wife (we used results for  $n = 7$  couples):

$$-2.9 \quad 24.0 \quad 21.3 \quad 12.9 \quad 14.7 \quad 5.2 \quad -1.8$$

For these we obtain

$$\bar{d} = 10.48571 \quad \text{and} \quad s_d = 10.65543.$$

Using  $\bar{d}$  and  $s_d$  in place of  $\bar{x}$  and  $s$  in the formula on page 42 gives:

$$\begin{aligned} 10.48571 \pm t \times \frac{10.65543}{\sqrt{7}} \\ 10.48571 \pm t \times 4.02737. \end{aligned}$$

For a 95% confidence interval we use the two-tailed 5% column in the table on page 31; also,  $\nu = n - 1 = 7 - 1 = 6$ , giving  $t = 2.447$ , and so the 95% confidence interval for the population mean difference is

$$10.48571 \pm 2.447 \times 4.02737 \longrightarrow (0.631, 20.341).$$

Notice that this interval does *not* include zero. This would indicate that the population mean difference in scores for males and females is significantly different from zero – that is, there is a real difference in the scores that males and females give! This is in agreement with the results of the paired comparisons  $t$  test we performed on page 41.

Sometimes, more or less ‘confidence’ is required. Although the 95% confidence interval is the ‘standard’, sometimes 90% or 99% confidence intervals are constructed. The only thing that changes here is the critical value used from Student’s  $t$  distribution: for a 90% interval, we would refer to the “10%” column in the table on page 31; for a 99% interval, we would use the “1%” column. Greater confidence would increase the width of the confidence interval.

## General exercises: sections 4→5

1. The following are the number of cars caught speeding each day on one speed camera over a two week period.

|    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|
| 10 | 12 | 15 | 9  | 8  | 12 | 11 |
| 6  | 15 | 17 | 12 | 10 | 9  | 7  |

How does this camera compare with the whole Northumbria police average of 8 per day per camera? Use an appropriate hypothesis test *and* construct a confidence interval to help you here.

2. The mean age of students attending a college is 19.5 years. The ages of a sample of 7 students who have been offered college accommodation are given below. Test the hypothesis that younger students are favoured.

|      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|
| 17.9 | 18.2 | 19.1 | 20.3 | 17.8 | 17.4 | 17.8 |
|------|------|------|------|------|------|------|

3. An airline company, EasyAir, advertises 5 hour flights from Glasgow to Cairo. Another company flying the same route, RyanJet, suspects EasyAir of false advertising, and samples 20 of EasyAir's flights between Glasgow and Cairo. The mean flight time from this sample is 5 hours and 10 minutes with a standard deviation of 20 minutes. Are RyanJet's suspicions supported?
4. A credit card company wants to determine the mean income of its card holders. It also wants to find out if there are any differences in mean income between males and females. A random sample of 225 male card holders and 190 female card holders was drawn, and the following results obtained:

|         | Mean    | Standard deviation |
|---------|---------|--------------------|
| Males   | £16 450 | £3675              |
| Females | £13 220 | £3050              |

Calculate 95% confidence intervals for the mean income for males and females. Is there any evidence to suggest that, on average, males' and females' incomes differ? If so, describe this difference.

5. Newcastle City Council are investigating the speed of cars on roads in a housing estate. The following data show the average speeds (in mph) for 8 roads in the morning (9am–12 noon) and in the afternoon (12 noon–3pm):

|           | Road |      |      |      |      |      |      |      |
|-----------|------|------|------|------|------|------|------|------|
|           | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    |
| Morning   | 18.4 | 19.3 | 15.2 | 17.1 | 25.4 | 35.8 | 18.9 | 26.4 |
| Afternoon | 20.5 | 15.8 | 15.1 | 16.5 | 22.4 | 35.0 | 15.1 | 27.0 |

Perform an appropriate hypothesis test to check for a difference in average speed between the morning and afternoon periods.

## Solutions to general exercises

### Sections 1—3 (pages 25–26)

1. Data types:

- (a) Qualitative  $\rightarrow$  Categorical
- (b) Quantitative  $\rightarrow$  Continuous
- (c) Quantitative  $\rightarrow$  Discrete  $\rightarrow$  Ordinal
- (d) Quantitative  $\rightarrow$  Discrete

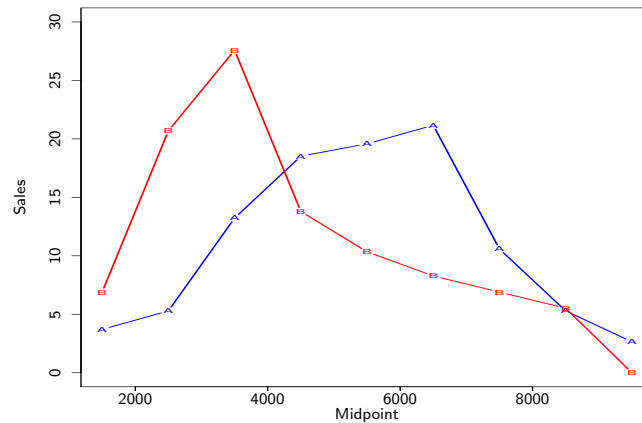
2. (a) Accessibility sampling. Could be flawed as it is non-random and is prone to bias.  
 (b) Difficult – if not impossible – to get a complete list of everyone in the target population.  
 (c) Judgemental sampling.

3. (a) Mean = 10.426 = 10,426 vehicles per day. Standard deviation = 0.76 = 760 vehicles per day.  
 (b) Median = 10.385 = 10,385 vehicles per day. The lower quartile is 9.95 (in position 3.75); the upper quartile is 10.725 (in position 11.25); therefore, the IQR = 0.775 = 775 vehicles per day.  
 (c) A box-and-whisker plot would consist of a horizontal ‘whisker’ extending from 9.38  $\rightarrow$  12.48, and vertical lines to form the ‘box’ at 9.95, 10.385 and 10.725.

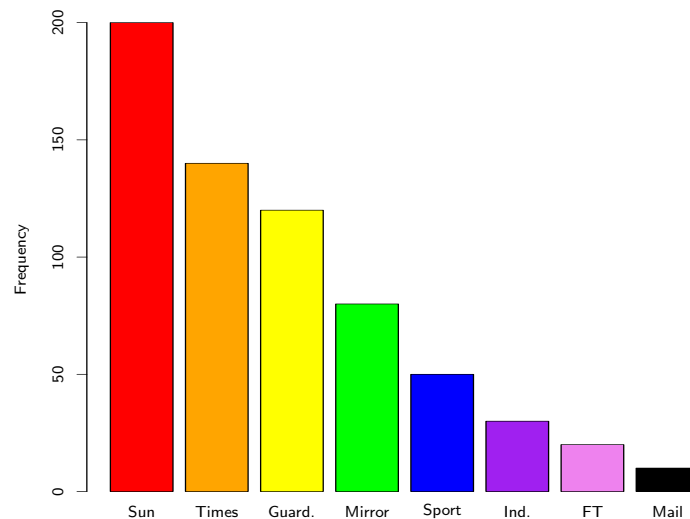
4. (a) The percentage relative frequencies are:

| Daily Sales                      | Before | After |
|----------------------------------|--------|-------|
| $1000 \leq \text{sales} < 2000$  | 6.9    | 3.7   |
| $2000 \leq \text{sales} < 3000$  | 20.7   | 5.3   |
| $3000 \leq \text{sales} < 4000$  | 27.6   | 13.2  |
| $4000 \leq \text{sales} < 5000$  | 13.8   | 18.5  |
| $5000 \leq \text{sales} < 6000$  | 10.3   | 19.6  |
| $6000 \leq \text{sales} < 7000$  | 8.3    | 21.2  |
| $7000 \leq \text{sales} < 8000$  | 6.9    | 10.6  |
| $8000 \leq \text{sales} < 9000$  | 5.5    | 5.3   |
| $9000 \leq \text{sales} < 10000$ | 0.0    | 2.6   |
| <b>Totals</b>                    | 145    | 189   |

- (b) Plotting the histograms, where “B”=Before and “A”=After, gives the graph overleaf. Notice that instead of drawing the bars, we have joined up the mid-points of the tops of the bars to create *polygons*. Polygons are often used instead of standard histograms when comparing two or more groups, as they are more easily overlaid on the same graph.



5. A simple bar chart or pie chart will do here:



6. (a)  $\bar{x} = 26.8$  planes per hour; Median = 29 planes per hour.  
 (b) Range = 34; IQR = 5.75; standard deviation = 9.566 planes per hour.  
 (c) The last observation is clearly an outlier, since the runway was closed for most of the hour and so it wasn't possible to observe a "typical" number of planes. Thus, the median and IQR might be the most suitable summaries of location and spread to use here (respectively), as these are not distorted by extremely large/small observations. The least suitable measure of location would be the mean, which can be distorted by outliers; the range would be the least suitable measure of spread, for the same reasons.
7. Correlation coefficient  $r = 0.786$  – positive association between speed and casualties.

**Sections 4 → 5 (page 46)**

1. One-sample  $t$  test.  $H_0 : \mu = 8$  versus  $H_1 : \mu \neq 8$ ;  $T = 3.453$ ;  $p$ -value less than 1%; reject  $H_0$  in favour of  $H_1$  – there is evidence to suggest a departure from the Northumbria police force average of 8 cars per day.  
95% confidence interval, using  $t = 2.160$ : (8.935, 12.922) cars per day – doesn't 'capture' the target value of 8, and so we would conclude that there is a significant departure from the force average of 8 per day – as in the hypothesis test.
2. One-sample  $t$  test.  $H_0 : \mu = 19.5$  versus  $H_1 : \mu < 19.5$  (one-tailed as we're asked to check whether younger students are favoured);  $T = -2.98$ ;  $p$ -value between 1% and 5%; reject  $H_0$  in favour of  $H_1$ ; there is evidence to suggest that younger students are being favoured!
3. One-sample  $t$  test.  $H_0 : \mu = 300$  minutes versus  $H_1 : \mu > 300$  minutes;  $T = 2.236$ ;  $p$ -value between 1% and 5%; reject  $H_0$  in favour of  $H_1$ ; *RyanJet's suspicions are supported!*
4. 95% Confidence intervals. Male: (£15969.80, £16930.20); female: ( pounds 12786.31, £13653.69). There is a clear gap between these intervals (i.e. they do not overlap), suggesting a significant difference between the average income of males and females. The interval for females is lower than that for males, suggesting also that average male income is higher than that for females.
5. Paired comparisons  $t$  test.  $H_0 : \mu_M = \mu_A$  versus  $H_1 : \mu_M \neq \mu_A$ ;  $\bar{d} = 1.1375$ ,  $s_d = 2.1084$ , giving  $T = 1.5259$ ;  $p$ -value greater than 5%; retain  $H_0$ ; there is insufficient evidence to suggest a difference between speeds in the morning and afternoon periods.