STAT0002 Introduction to Probability and Statistics

Dr Paul Northrop

2020-10-17

Contents

The purpose of these notes

These notes supplement the teaching materials available from the STAT0002 Moodle page. The teaching events in STAT0002 will follow the general order of the topics covered in these notes.

Please see the Module overview section of the STAT0002 Moodle page for important general information about STAT0002.

Chapter 1

Introduction

We will introduce core ideas in Probability and Statistics. These ideas will be introduced informally and the mathematical level will be kept as elementary as possible. Examples of real investigations will be used to motivate discussion of the ideas and to illustrate simple statistical methods. In the course STAT0003 the material in STAT0002 will be revisited in a more formal way and more advanced concepts and methods will be introduced.

1.1 Real statistical investigations

We will spend some lecture time looking at examples of real investigations. The first of these is introduced in Section 1.2 and will be used as an worked example for your Meet your Professor In-course Assessment (ICA). Most of these are real investigations which have been described in real research papers. We will also use some much simpler teaching examples to illustrate statistical ideas and methods. However, teaching examples can give the impression that all statistical analyses are straightforward. In practice they are not.

> "Most real-life statistical problems have one or more nonstandard features. There are no routine statistical questions; only questionable statistical routines." David Cox

The vast majority of real investigations have at least one non-standard feature which means that we cannot simply throw the data into a computer and get it to spit out the answer. Statistical analyses require a lot of careful thought.

Ideally a statistician should be consulted **before** any data are collected. Often this is not the case. Commonly the statistician is presented with a set of data, with little explanation of its meaning or context. Sometimes the researcher has processed the raw data in some way before giving it to the statistician, perhaps removing information that seems, to them, to be unimportant. It is not uncommon for a researcher to ask a statistician to "calculate a p-value for me". Real statistical analyses are never this simple.

Before starting a formal statistical analysis it is important to consider carefully the context of the problem. Data are not just numbers. They are recorded values of known **variables**, such as height or weight; they have **units** and an **interpretation**. Ask lots of questions of the people who produced the data, clarify the main objectives of the analysis and check for problems with the data. As we shall see in the in first example on the Space Shuttle, making a careless mistake early in an analysis can have dire consequences.

Many of the real-life problems we will consider required quite complicated data analyses reported in long research papers. I have summarised and simplified the details where necessary so that they are easier to understand. However, the main ideas and findings are unchanged. Some examples contain concepts

Table 1.1: Space shuttle data available at meeting. Number of O-rings (out of a total of 6) with thermal distress (damage) for launches at a given temperature

	flight	date	damaged	temperature
2	2	12/11/1981	1	70
9	9	03/02/1984	1	57
10	10	06/04/1984	1	63
11	11	30/08/1984	1	70
14	14	24/01/1985	3	53
21	21	30/10/1985	2	75
23	23	21/01/1986	1	58
24	24	28/01/1986	?	31

and words which we will not define until later in STAT0002 or STAT0003. However, their meaning should be clear from the context of the problem. I hope that these investigations will convince you of the importance of the subject of Statistics.

1.2 Challenger Space Shuttle Catastrophe

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989) Risk analysis of the space shuttle: Pre-Challenger prediction of failure. J. Amer. Statist. Assoc. **84**(408), 945–957.

On 28 January 1986 the space shuttle Challenger exploded shortly after its launch, killing the seven astronauts on board.

> Could this accident have been predicted and therefore prevented?

The accident was caused by gas leaking from one of the fuel tanks into the intense heat produced by the booster rockets. Usually such leaks are prevented by rubber seals called O-rings. It was known that O-ring failure would destroy Challenger and its crew. Subsequent investigation revealed that O-rings do not seal properly at low temperatures. O-rings need to expand to fill the gaps through which fuel can leak. At low temperatures O-rings lose elasticity, cannot expand, and therefore cannot fill the gaps.

The night before the launch some engineers expressed concern about the possible effect of low temperature on O-ring peformance - the temperature forecast for the launch was 31°F (-0.5°C), much lower than on previous shuttle launches and below the temperature at which the O-rings were designed to work effectively. A meeting was called at which data (in table Table 1.1) resulting from the 23 previous launches were discussed.

The third column contains the number of O-rings which showed some damage due to thermal distress after the flight. What do you notice about these numbers?

Figure 1.1, which illustrates these data graphically, was examined at the meeting. Despite some of the people present at the meeting suggesting that the launch should be postponed until the temperature reached the lowest temperature experienced in previous launches, the meeting concluded that there was no evidence of a temperature effect on the performance of O-rings and the launch went ahead.

Flights giving zero incidents of thermal distress were not included in the graph. This was because it was felt that these flights did not contribute any information about the temperature effect. When the complete dataset (see Table 1.2 and Figure 1.2) is examined it is clear that these flights **do** contribute extra information.

The enquiry into the Challenger accident concluded that a more careful analysis of the O-ring data would have revealed the apparent effect of temperature on O-ring performance.

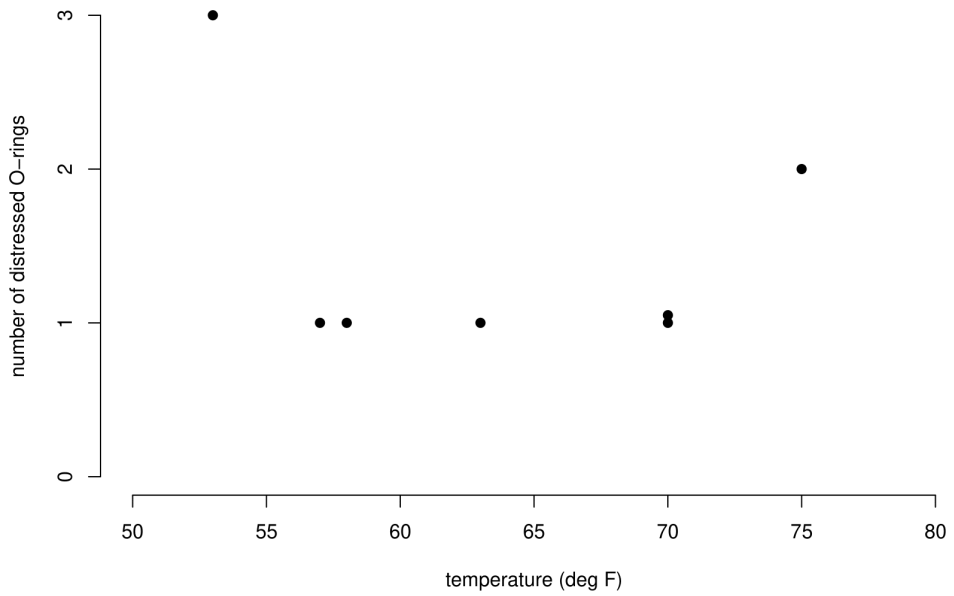

Figure 1.1: Number of damaged O-rings plotted against temperature, for flights prior to 28/01/1986. Flights showing no incidents of distress have been omitted. No clear association between the number of distressed O-rings and temperature is evident.

Table 1.2: Complete space shuttle data. Number of damaged O-rings (out of a total of 6) for launches at a given temperature

flight	date	damaged	temperature
1	21/04/1981	0	66
2	12/11/1981	1	70
3	22/03/1982	0	69
4	11/11/1982	0	68
5	04/04/1983	0	67
6	18/06/1983	0	72
7	30/08/1983	0	73
8	28/11/1983	0	70
9	03/02/1984	1	57
10	06/04/1984	1	63
11	30/08/1984	1	70
12	05/10/1984	0	78
13	08/11/1984	0	67
14	24/01/1985	3	53
15	12/04/1985	0	67
16	29/04/1985	0	75
17	17/06/1985	0	70
18	29/07/1985	0	81
19	27/08/1985	0	76
20	03/10/1985	0	79
21	30/10/1985	2	75
22	26/11/1986	0	76
23	21/01/1986	1	58
24	28/01/1986	?	31

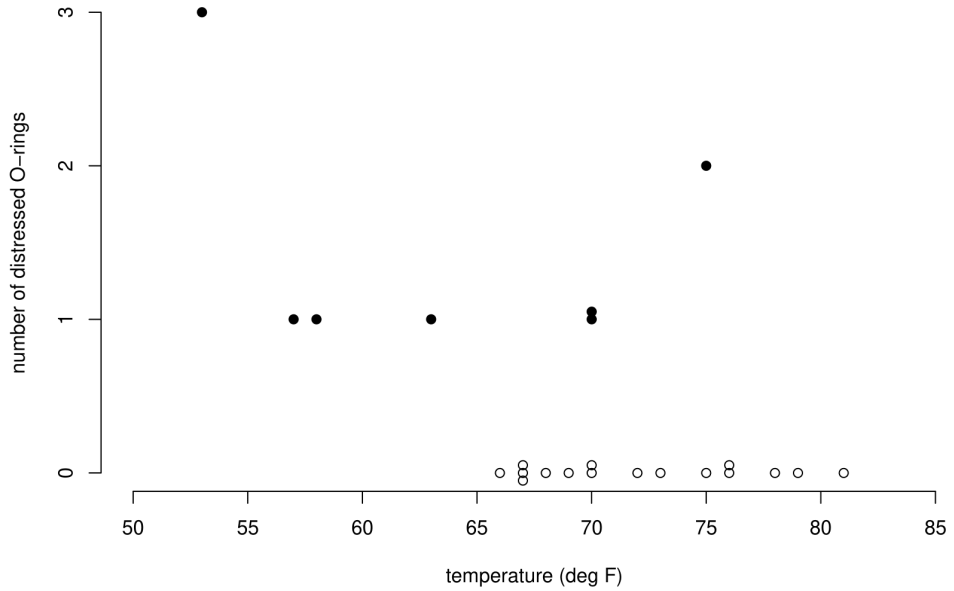

Figure 1.2: Number of damaged O-rings plotted against temperature, for flights prior to 28/01/1986. Flights showing no incidents of distress have been included as hollow circles. A clear negative association between the number of distress O-rings and temperature is evident.

What can we learn from this example?

- Data analyses can have life and death consequences. Statisticians can be very important people!
- Statistical analyses should use **all** the data. In this example only a non-random sample of the data are used. Removing some of the data had dire consequences. Values of zero are still data.
- It is dangerous to extrapolate beyond the range of your data. No data were available below 50°F. The forecast temperature of 31°F was much lower than this.

After the accident Dalal et al. (1989) estimated the probability of a catastropic O-ring failure (that is, one that would cause an explosion) at 31°F to be at least 0.13, which is large considering that seven lives were at stake. [To quantify their **uncertainty** they estimate that the probability is 90% certain to be between 0.03 and 0.37.] However, it should certainly be made clear that this estimate may not be at all reliable. For example, it could be that as temperature decreases below 50°F the risk of an accident increases much more quickly than the statistical analysis suggests.

The plot in Figure 1.3 gives you an idea of one of the analyses that Dalal et al. (1989) carried out. The sample proportions of O-rings showing thermal distress (number of O-rings showing distress divided by 6) are plotted against temperature. Also plotted is a smooth curve fitted to these data. [This analysis is beyond scope of STAT0002/0003. You may study this type of model in STAT0023.]

1.2.1 Uncertainty

Suppose there is a true curve, of the same general type as the one in Figure 1.3, which describes how the probability that an O-ring is damaged depends on temperature. We use the NASA test flight data to guess, or **estimate** the exact shape of this true curve. The curve in figure 1.3 is **not** the true curve, it is an **estimate** of the true curve based on these data.

If NASA repeated their launches, at exactly the same temperatures, these new data on the number of damaged O-rings would not be the same as the old data and the shape of the new estimated curve would be different from the shape of the old estimated curve. It may be that these 2 curves are quite similar or it could be that they are very different. We could ask them, very politely, repeat this process

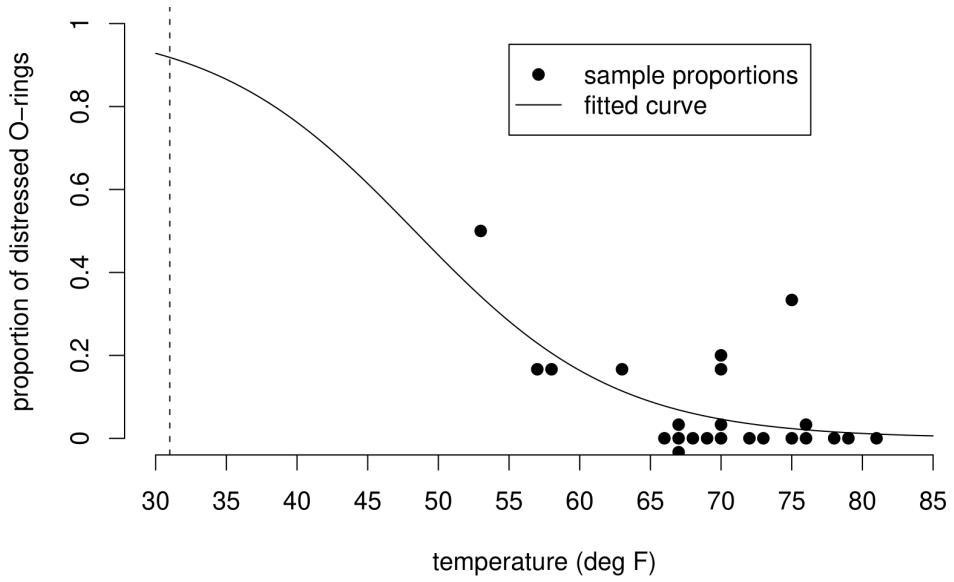

Figure 1.3: Proportion of O-rings showing some thermal distress plotted against temperature, with fitted logistic curve. The fitted curve reflects the apparent negative association between this proportion and temperature.

many times to get a large number of different sets of data. Each set of data produces an estimated curve.

The dataset (and its estimated curve) we have is just one of many possible datasets that could be produced. We can imagine picking this dataset (and curve) at random from a big bag of possible datasets (and curves). These ideas are summarised in Figure 1.4.

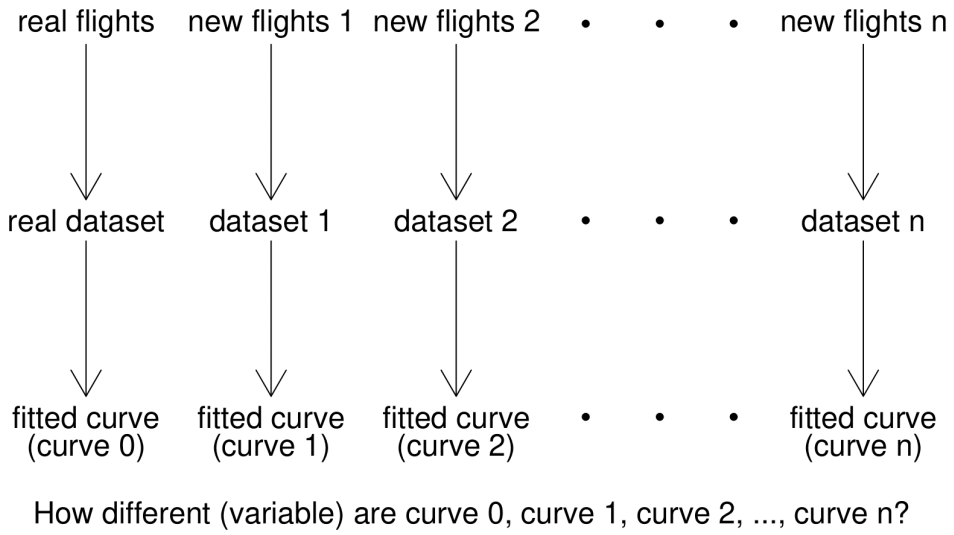

Figure 1.4: Diagram to illustrate the idea of repeating an experiment many times. Each simulated set of flights leads to its own dataset and fitted logistic curve.

Suppose that the estimated curves from the possible datasets are very similar to each other. We say that their **variability** is small. If this is the case then it doesn't matter much which dataset we picked from the big bag of possible datasets: the results are similar for all datasets. Therefore, we can be fairly certain that the results we got from the dataset we have are close to the truth. Therefore the **uncertainty** surrounding the results is small.

On the other hand, if the estimated curves from the possible datasets are very different to each other then their **variability** is large. If this is the case then the results will be very different depending on

which dataset we pick. Therefore, it is possible that the results we got from the dataset are very far from the truth. Therefore the **uncertainty** about the results is large.

We can see that **variability** and **uncertainty** are closely related. Small variability tends to produce small uncertainty, whereas large variability tends to produce large uncertainty. As we might expect the amount of data, (or, more precisely, the amount of **information** in the data) matters. Large datasets, with lots of information, tend to produce small variability in the results and therefore small uncertainty. Small datasets, with small amounts of information, tend to produce large variability in the results and therefore large uncertainty.

So, how can we quantify how much uncertainty there is in the space shuttle example? It is unlikely that NASA will carry out all their launches again just for us. However, it is possible for us to produce (**simulate**) on a computer our own, fake, datasets using the estimated curve in Figure 1.3. If this curve (and the assumptions used to produce it) are correct, this is equivalent to NASA carrying out more test flights: the simulated datasets have exactly the same statistical properties as the real dataset. In summary, we

- create a large number of fake (**simulated**) datasets;
- for each dataset we estimate a curve to describe how the probability of O-ring damage depends on temperature;
- examine how much the curves, and the estimate of probability at different temperatures vary between the simulated datasets.

Figure 1.5 shows 50 simulated curves and the curve estimated from the real data.

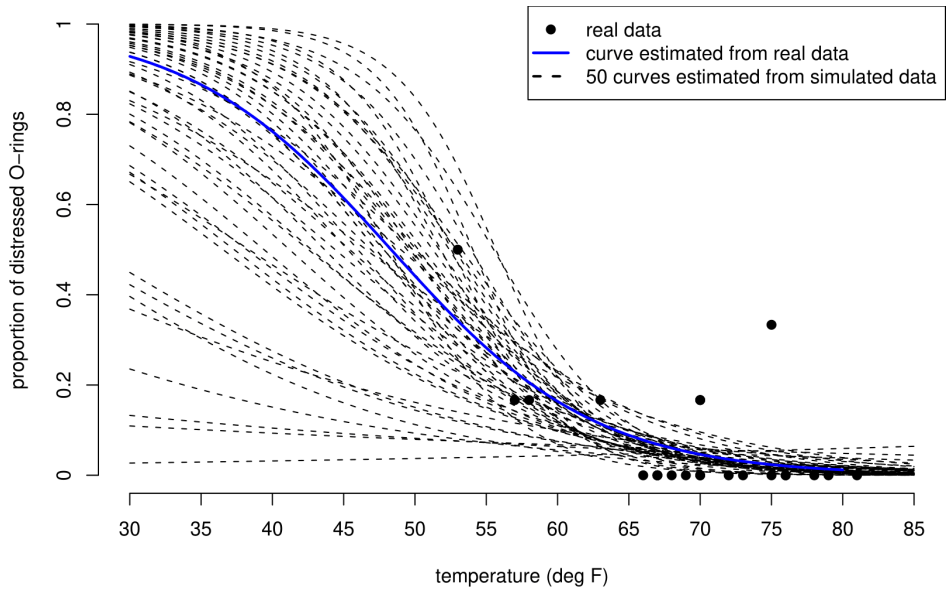

Figure 1.5: 50 curves fitted to simulated shuttle test flight data. The curves are similar over the range of temperatures observed in the data (53 to 81 degrees F), but vary greatly for lower temperatures, such as 31 degrees F.

There is a lot of variability in these curves. Notice that the curves are quite close to each other for high temperatures - where we have some data - but that they are very spread out for low temperatures - where we have no data. This is confirmed by figure 1.6 which shows how the estimated probability of O-ring damage depends on temperature.

There is a large amount of uncertainty about the esimatated probabilities, particularly at 31°F, where it really mattered.

To show the effect of sample size (the size of the dataset) we simulate datasets which are larger than the real dataset and see how much the curves fitted to these data vary between the datasets. Figure

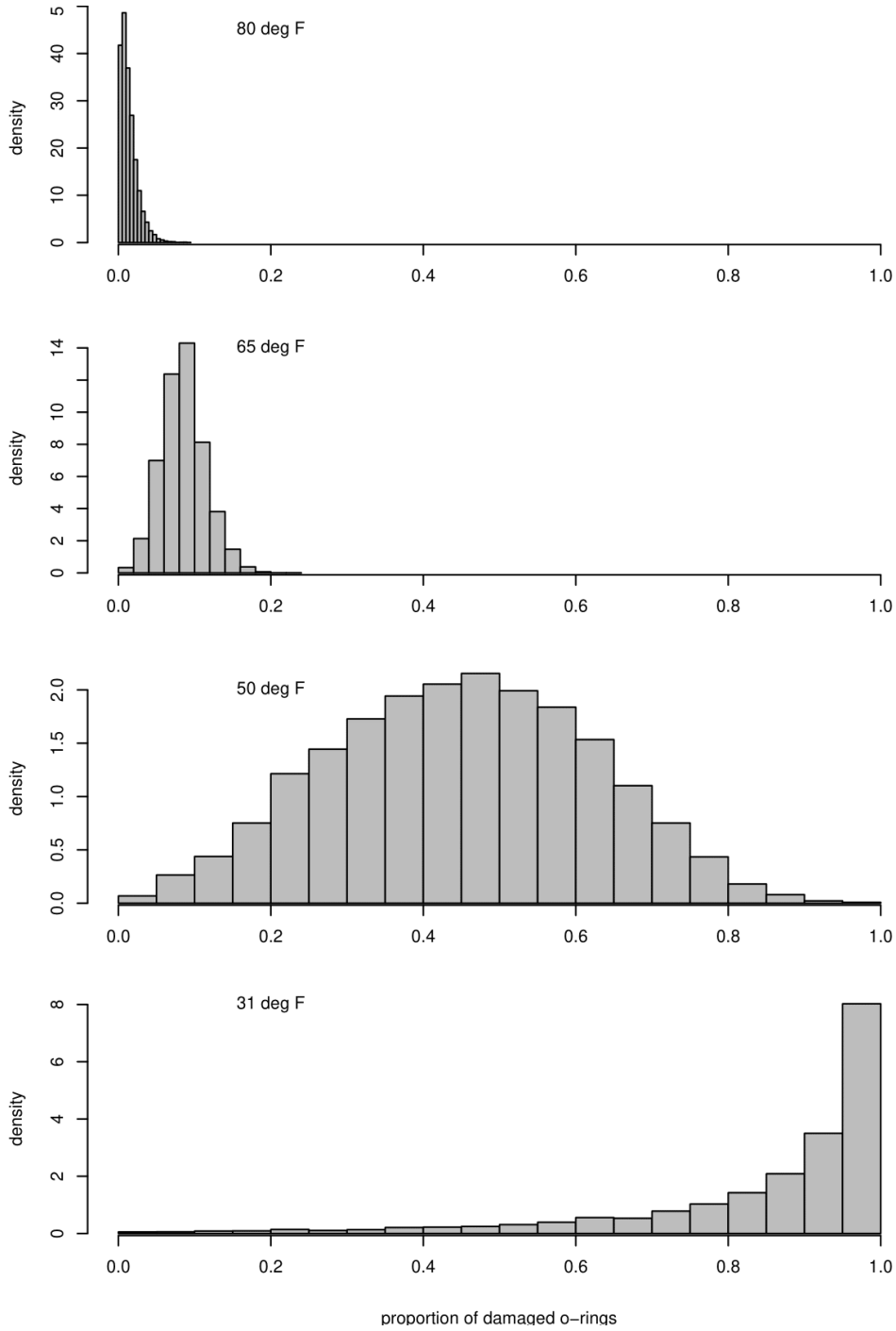

Figure 1.6: Histograms of estimated probabilities of O-ring damage at different temperatures.

1.7 shows the estimated curves from 50 datasets, each of which is 10 times the size of the real dataset. Figure 1.8 shows curves for datasets which are 100 times the size of the original dataset. As the sample size increases the variability decreases and so the uncertainty decreases.

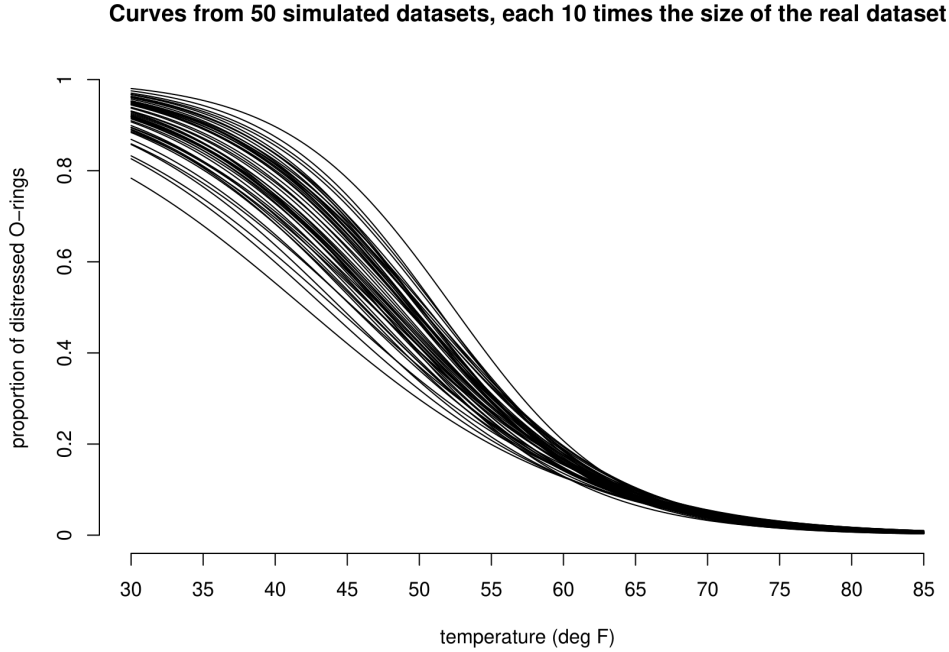

Figure 1.7: 50 curves fitted to simulated shuttle test flight datasets that are each 10 times the size of the real dataset. In comparision to the curves based on the real dataset these curves vary less, but are still most variable for low temperatures.

1.3 A very brief introduction to stochastic simulation

This section contains words that we will not define until later in the course. Further information about stochastic simulation is available from the Stochastic Simulation section of the STAT0002 Moodle page.

In Statistics it is common to assess a statistical method based on how well it would perform if used repeatedly on a large number of new datasets, where we imagine that the new datasets have exactly the same statistical properties as the real data. In some cases it is possible to do this using mathematics. Alternatively, we can use a computer to produce some fake (simulated) datasets from a model that has been fitted to the real data. How can we do this?

Stochastic (stochastic simply means "involving randomness") simulation is based on the ability to generate a random number u between 0 and 1. Stochastic simply means "involving randomness". Your pocket calculator probably has a button to do this, perhaps called RAN#. It is possible to transform this number u so that it looks like it has been drawn from the distribution required, e.g. a binomial distribution or a normal distribution. If we produce a sequence $u_1, u_2, \ldots, u_n$ of random numbers between 0 and 1 and transform them appropriately, then the transformed values will look like a random sample from the distribution required. Of course, because these values are produced by rule implemented by a computer they are not really random. However, if the rule is designed carefully, these values are close enough to being a random sample for our purposes.

For the purposes of the space shuttle experiment we simply need to simulate a 1 (O-ring distressed) with probability p, and a 0 (0-ring not distressed) with probability $1-p$. This is easy. If U is a random number between 0 and 1 then the probability that $U < p$ is p. Therefore, we define

$$X = \begin{cases} 1 & \text{if } U < p, \\ 0 & \text{if } U \geq p. \end{cases} \tag{1.1}$$

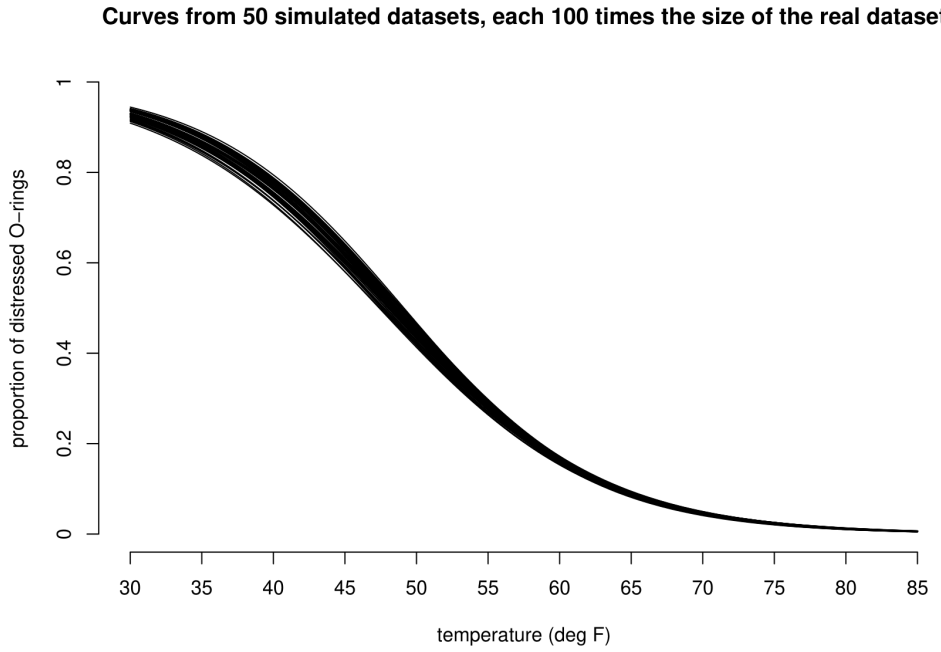

Figure 1.8: 50 curves fitted to simulated shuttle test flight datasets that are each 100 times the size of the real dataset. In comparision to the curves based on the real dataset these curves vary much less, but are still most variable for low temperatures.

For $p = 1/2$ this is like using a computer to flip an unbiased coin.

To simulate a fake space shuttle dataset we do the following for each of the 23 flights:

1. set p to be the value of the fitted curve in Figure 1.3 corresponding to the flight temperature;
2. generate 6 random numbers $u_1, \ldots, u_6$ between 0 and 1;
3. calculate $x_1, \ldots, x_6$ using equation (1.1);
4. calculate $y = x_1 + \cdots + x_6$, the total number of distressed O-rings.

We have assumed that the 6 O-rings have the same probability of becoming distressed and are distressed independently of each other. In Chapter 4 we will see that y is a value simulated from a binomial(6, p) distribution. We will use simulation several times in STAT0002 to study properties of statistical methods. All you need to know is that we can use a computer to produce fake data that look like they come from a certain probability distribution.

Chapter 2

Descriptive Statistics

The first important step in any data analysis is to **describe** the available data. This is often called an **exploratory** or **initial** data analysis. It is normally not possible to just look at the dataset, especially if it is large, and just see any interesting structures. The task of a statistician is therefore to **extract** and **condense** the relevant information – what is relevant will depend on the aim of the analysis. Some of the standard methods to do so are addressed in the next sections. Despite all the technicalities, always remember that the numbers / figures / plots produced for data must be **interpreted with regard to the problem or question at hand**, that is, always ask yourself "what does this number / plot mean?".

Before embarking on a formal statistical analysis of the data we should look at summaries of the data such as graphs, tables and summary statistics. This can be important to

1. reveal problems with, or errors in, the data;
2. get a 'feel' for the data;
3. identify interesting features of the data, e.g. is treatment A very obviously better at treating a disease than treatment B?;
4. suggest how the data should be analysed;
5. present conclusions.

In some cases the data summaries make it very clear what is going on and may make more formal methods of statistical analysis unnecessary.

2.1 Types of data

Before analysing data it is important to consider what **type** they are. This will affect which statistics it is sensible to calculate, which graphs it is sensible to plot and which of the simple distributions we will study in Chapter 5 might be used for these data.

2.1.1 Qualitative or categorical data

Items are assigned to **groups** or **categories** based on some **qualitative** property. Examples:

- Hair colour: blonde, brown, red, black etc.
- Smoking status: smoker, non-smoker;
- Severity of illness: none, mild, moderate, severe;
- Degree class: 3, 2ii, 2i, 1.

The data are **labels**: if numbers are assigned to the categories (e.g. 0=smoker, 1=non-smoker) the numbers chosen do not mean anything in themselves.

Categorial data can be classified as either

- **nominal**: the categories are unordered, e.g. hair colour, smoking status;
- **ordinal**: the categories are ordered, e.g. severity of illness, degree class.

An important special case is **binary** data: categorical data with only 2 categories. These data can be nominal (e.g. male, female) or ordinal (e.g. small, large).

Nominal data: describe by (relative) frequencies. It is sensible to quote the mode, but not the mean or median.

Ordinal data: It is sensible to quote the mode or median, but not the mean.

2.1.2 Quantitative or numerical data

Items are measured in some way based on some **quantitative** property. This produces a one, or more, **numbers**. Examples:

- Time, in hours;
- Height, in cm;
- Age, in years;
- Number of damaged O-rings (see space shuttle investigation);
- Number of births on one day at a particular hospital;
- Number of units passed in first year.

Numerical data can be classified as either

- **Discrete**. Only certain values are possible (there are gaps between the possible values), e.g. number of damaged O-rings, number of births, number of units passed in first year (0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4);
- **Continuous**. In theory, **any** value within an interval of the real line is possible, e.g. time, height, age.

Often discrete data are **counts**. Continuous data usually come from measurement. In practice continuous data are recorded discretely, e.g. to two decimal places.

Interval data and ratio data

Quantitative data can be further classified as **interval** or **ratio**. Both interval data and ratio data have the property that an increase of 1 unit means the same whether it is from, say, 1 to 2 or from 10 to 11. However,

- a ratio scale has a natural zero, for example, temperature measured in degrees Kelvin;
- an interval scale does not have a natural zero, for example temperature measured in Fahrenheit.

Ratios are only meaningful on a ratio scale. For example,

- IQ: A zero IQ does not exist. A person with an IQ of 120 is not twice as intelligent as a person with an IQ of 60. Therefore, IQs are interval data.
- Income: A zero income does exist. A person whose take-home income is £20,000 does earn twice as much as some whose take-home income is £10,000.Therefore, incomes are ratio data.

2.2 Describing distributions

In describing the distribution of one variable the following it is important to examine the following.

1. **Location / average / central tendency of the data**. Where is the centre of the distribution? What is a typical value?

2. **Spread / variability / dispersion / scale**. How **variable** are the data? How far are they spread out?
3. **Shape**. What shape is the distribution of the data? In particular, is it **symmetric** or **skewed**, and if skewed, which way? A long tail to the right is called **positive skew** (or right skew or skewed to the right). A long tail to the left is known as a **negative skew** (or left skew or skewed to the left). Positive skew is much more common than negative skew. Figure 2.1 gives some examples of shapes of symmetric, positive skew and negative skew distributions. In addition to being symmetric the plot in the top left of Figure 2.1 of figure is bell-shaped. This shape is the shape of a **normal distribution** (see Section 5.7). The normal distribution is an important distribution in Statistics. We may wish to decide whether the data look like they have come from a normal distribution.
4. **Outliers**. Are there any outliers, that is, observations that appear to be out of line with the pattern of the rest of the data? This issue can also be hard to judge. For example, with a small number of observations, it is difficult to distinguish between data from a heavily skewed distribution and data from a symmetric distribution with outliers. What constitutes an outlier depends on the context so there is no rigid rule for defining/detecting outliers. The intended statistical analysis also matters. We will consider how to deal with outliers (in the context of linear regression) in Section 8.3.1.
5. **Is there anything else to report?** Note any **unusual features** about the data. Are there particular numbers which appear more often than we could expect? Do the data separate into groups?

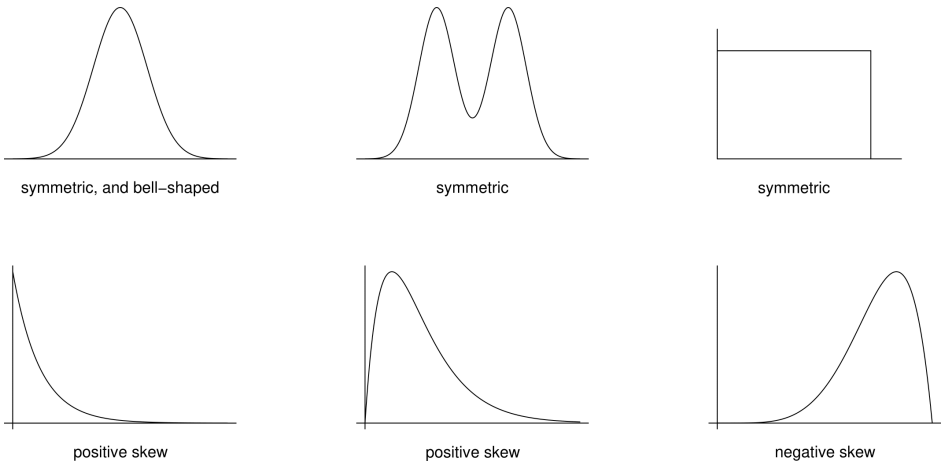

Figure 2.1: Examples of shapes of symmetric, positively skewed and negatively skewed distributions.

We will look at 3 ways basic tools which are used to describe and summarise data: summary statistics, tables and graphs. We can use a combination of these. For example summary statistics may be presented in a table or a graph. Carefully produced graphs are often the best way to describe, explore and summarise a set of data. Summary statistics reduce the data to a small set of numbers. Tables can retain more information but do not work well for datasets which are large or have many variables. In contrast graphs can show most, if not all, the information in the data and reveal complex relationships.

Table 2.1: Time (in hours) spent by each of 95 women giving birth at the John Radcliffe hospital in Oxford, UK, during a particular week.

day1	day2	day3	day4	day5	day6	day7
2.10	4.00	2.60	1.50	2.50	4.00	2.00
3.40	4.10	3.60	4.70	2.50	4.00	2.70
4.25	5.00	3.60	4.70	3.40	5.25	2.75
5.60	5.50	6.40	7.20	4.20	6.10	3.40
6.40	5.70	6.80	7.25	5.90	6.50	4.20
7.30	6.50	7.50	8.10	6.25	6.90	4.30
8.50	7.25	7.50	8.50	7.30	7.00	4.90
8.75	7.30	8.25	9.20	7.50	8.45	6.25
8.90	7.50	8.50	9.50	7.80	9.25	7.00
9.50	8.20	10.40	10.70	8.30	10.10	9.00
9.75	8.50	10.75	11.50	8.30	10.20	9.25
10.00	9.75	14.25		10.25	12.75	10.70
10.40	11.00	14.50		12.90	14.60	
10.40	11.20			14.30		
16.00	15.00					
19.00	16.50					

Example: Oxford births data

Table 2.1 shows the times (in hours) spent by 95 women giving birth in the delivery suite of the John Radcliffe Hospital in Oxford during 1 week. These are ratio data. At first we ignore the fact that the data are recorded on different days.

2.3 Summary Statistics

One way to summarise a dataset is to calculate numerical summaries called **summary statistics**. Summary statistics can be used as indicators of the location, spread and shape of the data (although looking at a plot can be more helpful).

2.3.1 Five number summary

A useful first impression of the distribution of quantitative or ordinal data is given by the a five number summary. As we will see later, the five number summary involves quantities called **sample quantiles**. These are estimates of theoretical quantities that we will study in Chapter 4. There is more than one way to calculate sample quantiles. For example, the R statistical package has 9 options in its `quantile()` function. The particular method given below is just one of these options.

If a dataset of observations, $x_1, x_2, \ldots, x_n$, is arranged in order of size as

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

then the **sample median** is the 'middle' value (halfway between $x_{(1)}$ and $x_{(n)}$), that is,

$$m = x_{(\frac{1}{2}(n+1))}\,.$$

The median is a measure of location.

Informally, we can think of the **sample lower quartile** as the sample median of the lower half of the data, or, equivalently, as the value that divides the lower 25% of the data from the rest of the data. One way to estimate this is

$$q_L = x_{(\frac{1}{4}(n+1))}\,.$$

Similarly, we can think of the **sample upper quartile** as the sample median of the upper half of the data, which we could estimate using

$$q_U = x_{(\frac{3}{4}(n+1))}\,.$$

If m, q_L or q_U do not correspond directly with one of the observations then we can use linear interpolation. Suppose that $n = 44$. Then we could calculate the sample median using

$$x_{(22.5)} = x_{(22)} + \frac{1}{2}\left(x_{(23)} - x_{(22)}\right) = \frac{x_{(22)} + x_{(23)}}{2},$$

the sample lower quartile using

$$x_{(11.25)} = x_{(11)} + \frac{1}{4}\left(x_{(12)} - x_{(11)}\right) = \frac{3}{4}\,x_{(11)} + \frac{1}{4}\,x_{(12)},$$

and the sample upper quartile using

$$x_{(33.75)} = x_{(33)} + \frac{3}{4}\left(x_{(34)} - x_{(33)}\right) = \frac{1}{4}\,x_{(33)} + \frac{3}{4}\,x_{(34)}.$$

This is not the only possibility: you may find that different methods are used in some textbooks and by some computer packages. If the data are ordinal then interpolating may not make sense.

The quartiles q_L, m, q_U (so called because they divide the data into 4 equal parts) are sometimes denoted q_1, q_2 and q_3.

The **five number summary** of the data set is the set of values

$$x_{(1)}, q_L, m, q_U, x_{(n)}\,,$$

that is, the sample minimum, lower quartile, median, upper quartile and maximum.

The **range** is defined as $x_{(n)} - x_{(1)}$ and the **inter-quartile range** (IQR) as $q_U - q_L$. The range and IQR are measures of spread.

More generally, we could could calculate sample **quantiles**, or **percentiles**. The 30% quantile, for example, is the value at or below which 30% of the data lie. The $100p$% sample quantile is $x_{(p(n+1))}$. When $p(n+1)$ is not an integer, $x_{(p(n+1))}$ can be calculated using linear interpolation. Note: the number of quantiles which we can estimate reliably depends on the **sample size** n. For example, if $n = 3$, it doesn't make sense to try to estimate the 10% quantile. In this case $q_L = x_{(1)}, m = x_{(2)}$ and $q_U = x_{(3)}$.

Sometimes the sample size n is added to the five number summary. The sample size can be of interest in its own right, for example when it records the number of times an event of interest occurs in a fixed period of time, for example, the number of births in the delivery suite of the John Radcliffe hospital in Oxford during one week.

2.3.2 Mean and standard deviation

The most well known descriptive measures of **numerical** data are the (arithmetic) mean and the standard deviation.

The **sample mean**, a measure of location, is defined as the arithmetic average

$$\bar{x} \;=\; \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

The **sample variance**, a measure of spread, is

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

The **sample standard deviation**, also a measure of spread, but has the same units as the data, is

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

For example, if the units of the data are metres then the units of the variance are metres^2 and the units of the standard deviation are metres.

The formula (with a different denominator)

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

which is used by some calculators is equal to $s^2(1-1/n)$, **not** s^2. For large n the values of s^2 and $s^2(1-1/n)$ will be close.

For data that are very skewed, or contain outliers, the sample median may be a more appropriate measure of location than the sample mean. This is because the value of the sample mean is strongly influenced by large or small values. For example, if the data are positively skewed the value of the mean may be much larger than where we would judge by eye the centre of the data to be. However, for data which are fairly symmetric there are reasons to prefer the sample mean to the sample median. For example,

- the sample mean is easier to calculate;
- if samples are taken repeatedly the sample mean varies less than the sample median.

We will examine this in more detail in Section 6.5. Similarly, for a measure of spread, the sample standard deviation may be preferred for approximately symmetric data with no outliers, otherwise the IQR is preferable.

2.3.3 Mode

For categorical data or discrete data the mode is the value (or values) which occurs most often. The concept of a mode is relevant to continuous data, but it is less obvious how we might estimate this using data. We return to this in Section 4.5. The mode is a measure of location.

Examples

What are the sample mean, median and mode of the following data?

blonde hair, red hair, red hair, black hair

What are the sample mean, median and mode of the degree classes?

3, 2ii, 2i, 1, 1

What are the sample mean, median and mode of the following numbers?

10, 350

Which measures of location are sensible for different types of data? Consider each case in Table 2.2.

Table 2.2: Types of data and measures of location.

	mean	median	mode
nominal			
ordinal			
numerical			

2.3.4 Symmetry

Many standard statistical methods work best when the data are distributed symmetrically. Looking at a graph is the best way to examine whether this is true. However, the relative values of the sample mean and sample median can give us an idea whether the data are approximately symmetric, as summarised in Table 2.3, but this rule-of-thumb can be misleading.

Table 2.3: Relative values of the sample mean and median and what this **might** suggest in some cases.

mean $<$ median	mean $=$ median	mean $>$ median
negative skew	symmetric	positive skew

Example. Oxford births data

Table 2.4 gives the five-number summary of the Oxford birth times data.

Table 2.4: Sample five-number summary of the Oxford birth times data.

$x_{(1)}$	q_L	m	q_U	$x_{(n)}$
1.50	4.90	7.50	9.75	19.00

Half of the women took between approximately 5 and 10 hours to give birth.The quickest delivery was 90 minutes and the longest 19 hours. The mean $\bar{x}$ is 7.72 hours and the standard deviation is 3.57 hours. The fact that sample mean $>$ sample median suggest that the data are slightly positively skewed, but this is something that we should confirm by looking at a suitable graph (see Section 2.5).

Measures of skewness

Usually the best way to examine the shape of a distribution is to look at a graph see section 2.5. In addition we could calculate summary measures of skewness, such as: the **standardized sample skewness**

$$\text{skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3},$$

where s is the sample standard deviation, and the **sample quartile skewness**

$$\text{quartile skewness} = \frac{(q_U - m) - (m - q_L)}{q_U - q_L},$$

where q_L, m and q_U are the sample quartiles.

These measures are each 0 for perfectly symmetric data, negative for negative skew data and positive for positive skew data. The standardized sample skewness can take any value on the real line. The

quartile skewness must lie in $[-1, 1]$. The quartile skewness has the advantage that it is less sensitive to outliers than the standardized sample skewness.

For the Oxford births data the standardized sample skewness is 0.63 and the sample quartile skewness is -0.072. In this example, the standardized sample skewness suggests that the data are positively skewed, whereas the quartile skewness suggests that the data are (slightly) negatively skewed.

Table 2.5 summarises the summary statistics may be used as measures of location, spread and shape.

Table 2.5: Summary of summary statistics

location	spread	shape
median	inter-quartile range	quartile skewness
mean	standard deviation or variance	skewness
mode		

2.3.5 Correlation

Measures of correlation aim to summarise the strength of the relationship between two variables. Suppose that we have two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ of **paired** data. For example, x_1 and y_1 could be the height and weight of person 1, x_2 and y_2 the height and weight of person 2, etc.

The sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \in [-1, 1]. \tag{2.1}$$

measures the strength of **linear** association between the two variables.

We will look at correlation in detail later in the course (in Chapter 9). We must be careful to use the sample correlation coefficient only when it is appropriate to do so. We will see that it is important to plot the data.

The product-moment correlation coefficient r is not the only possible measure of correlation. An alternative is **Spearman's rank correlation coefficient** r_S. First we rank the $x_1, \ldots, x_n$ values, giving a rank of 1 to the largest x value, a rank of 2 to the second largest, down to a rank of n for the smallest value. This gives ranks $r_1^x, \ldots, r_n^x$. Then we do the same with $y_1, \ldots, y_n$ to produce ranks $r_1^y, \ldots, r_n^y$. [If there are ties then we average the ranks of tied observations, e.g. if the 3rd and 4th largest values are equal then they each get a rank a 3.5.] Then we calculate the product-moment correlation of the paired ranks$(r_i^x, r_i^y), i = 1, \ldots, n$ using equation (2.1). If there are no ties then r_S simplifies to

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where $d_i = r_i^x - r_i^y$ is the difference in the ranks of x_i and y_i. The general idea is to extract from the raw data only the ordering of the data points.

The choice between using r or r_S as a measure of correlation is similar to the choice between using the sample mean or the sample median as a measure of location. In particular, r_S is less sensitive to outliers than r.

We have noted that r measures the strength of **linear** association between two variables. In contrast r_S is a measure of how close the relationship between the variables is to being **monotone**, i.e. either increasing or decreasing but not necessarily linear. If $r_S = 1$ then the data have a perfect monotone increasing relationship. If $r_S = -1$ then the data have a perfect monotone decreasing relationship.

A simple example

Consider the small dataset in Table 2.6.

Table 2.6: A small example dataset.

x_i	rank x_i	y_i	rank y_i	d_i
-2	6	-1.5	6	0
-1	5	-1.1	5	0
0	4	0.2	4	0
1	3	1.1	3	0
2	2	1.6	1	1
10	1	1.5	2	-1

Exercise. Show that for these data $r = 0.70$ and $r_S = 0.94$. Can you explain why $r_S > r$? Looking at a scatter plot of y against x will help you see why.

2.4 Tables

We saw in the Space shuttle investigation (Section 1.2) that data can be presented in a table. We also saw that a graph can be a better way to see relationships and patterns in the data. In this section we look at a table which summarises the distribution of a set of data on one variable. We also look at a graph based on this table.

2.4.1 Frequency distribution

A **frequency distribution** is a tabular summary of a set of data that shows the number of items in each of several non-overlapping classes. To construct a frequency distribution for a sample we need to choose:

- the number of classes;
- the width of classes.

It is common to choose all classes to have the same width, but there may be situations where it makes sense to use classes with different widths. For discrete data each data value usually constitutes a class.

The first and second columns of Table 2.7 show the frequency distribution of the Oxford birth times. The first column defines the classes, the second column gives the number of observations (the **frequency**) which fall into each class. The frequencies sum to 95, the total number of observations. The frequency distribution provides a quick way to summarise the birth times. From the table we can see that the class 6–8 hours has the largest frequency. Therefore, 6–8 hours is called the **modal class**. However, note that the frequency distribution depends on the choice of the classes.

Relative frequency distribution

Column 3 of Table 2.7 contains the **proportion** or **relative frequency** of observations in each class. Column 3 is calculated by dividing column 2 (the frequencies) by the total frequency (95 in this example). This produces the **relative frequency distribution** of the data. Column 3 shows that, for example, 15% of the women took between 2 and 4 hours to give birth. A graphical display of the relative frequency distribution of a set of data is provided by a **histogram** (see Section 2.5.1).

Table 2.7: Frequency table of the Oxford birth times. x–y means $x <$ time $\leq y$.

time (hours)	frequency	relative frequency	cumulative frequency	cumulative relative frequency
0-2	2	0.02	2	0.02
2-4	14	0.15	16	0.17
4-6	14	0.15	30	0.32
6-8	22	0.23	52	0.55
8-10	21	0.22	73	0.77
10-12	12	0.13	85	0.89
12-14	2	0.02	87	0.92
14-16	6	0.06	93	0.98
16-18	1	0.01	94	0.99
18-20	1	0.01	95	1
total	95	1		

Cumulative distribution

Column 4 of Table 2.7 contains the total number of observations with values less than or equal to the upper limit of each class. This is the **cumulative frequency**. We can see that 73 of the women took no longer than 10 hours to give birth. Column 5 contains the proportion of observations with values less than the upper limit of each class. This is the **cumulative relative frequency**. It is calculated by dividing column 4 by 95. We can see that approximately 77% of the women took no longer than 10 hours to give birth. It can be helpful to display the cumulative relative frequencies in a graph as in Figure 2.2.

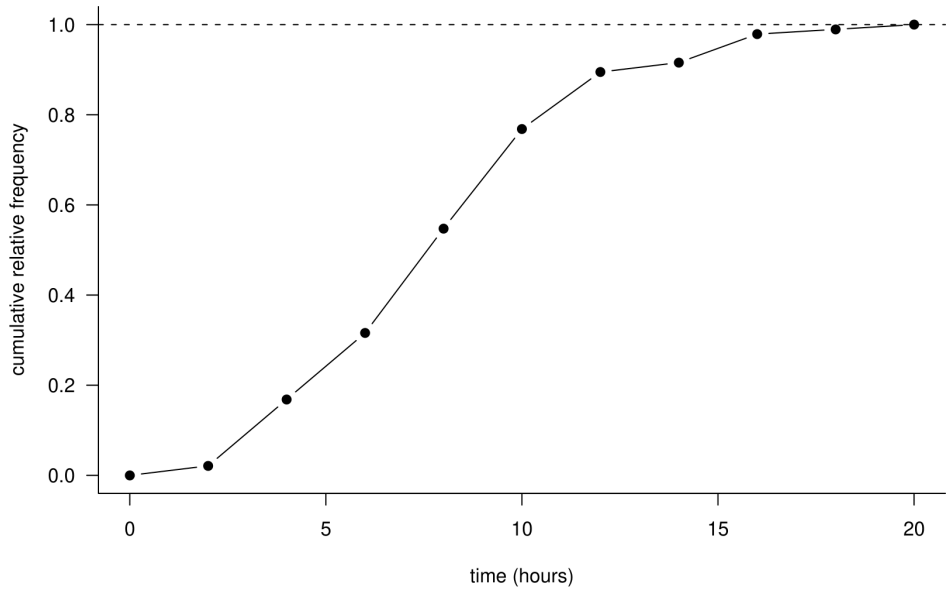

Figure 2.2: A cumulative relative frequency distribution of the Oxford birth times.

The shape of this plot depends on the choice of the classes. We could increase the detail in the plot by increasing the number of classes, that is, by decreasing the class width. In an extreme case we could choose the classes so that there is a class for every unique value in the data. Table 2.8 shows how this could be done.

Figure 2.3 shows the resulting graph. Notice that the shape is similar to Figure 2.2 but it is less smooth. The function (from time on the horizontal axis to cumulative relative frequency on the vertical axis)

Table 2.8: Frequency table of the Oxford birth times, with one observation per class.

time (hours)	frequency	relative frequency	cumulative frequency	cumulative relative frequency
0.0-1.5	1	1/95	1	1/95
1.5-2.0	1	1/95	2	2/95
2.0-2.1	1	1/95	3	3/95
2.1-2.5	2	2/95	5	5/95
...	...	...	...	...
16.0-16.5	1	1/95	94	94/95
16.5-19.0	1	1/95	95	1
total	95	1		

is often called the **empirical cumulative distribution function** or **empirical c.d.f**. The meaning will become clearer when we look at c.d.f.s in Section 4.1.

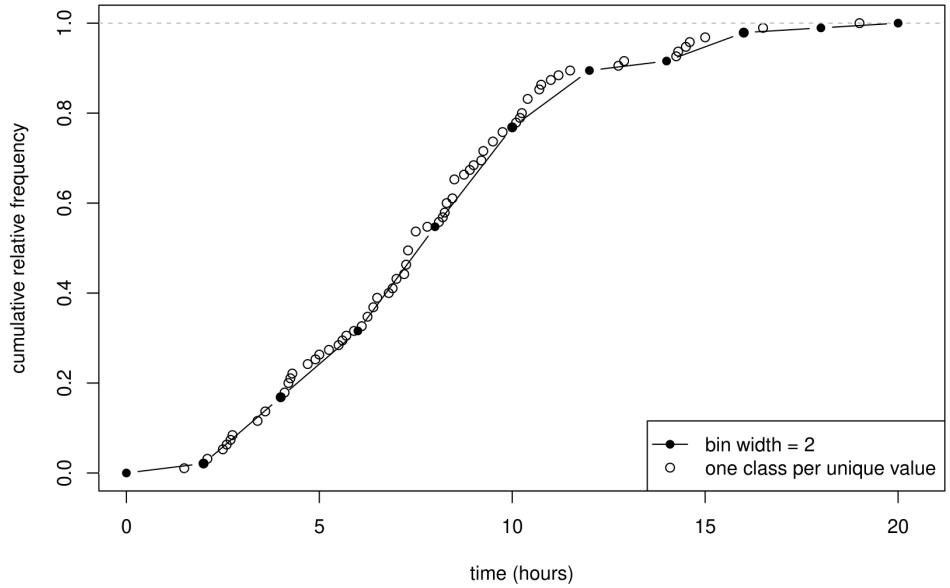

Figure 2.3: A cumulative relative frequency distribution of the Oxford birth times, with classes defined so that there is one data value in each class.

Sometimes data are given to us in the form of Table 2.7 and the individual data values are not available. For example, some birth data published by the Office of National Statistics data give mother's age in 5 year age bands. Data provided in this form are called **grouped data**. We will analyse data from tables in Chapter 7.

2.5 Graphs (1 variable)

These days it is very easy to plot a graph using a computer. However, **you** need to decide which type of graph is appropriate and the default graph produced by the computer may not be very good. Some general rules:

- **Always plot the data**. Often this will show clearly the important features of the data. Formal statistical methods may be unnecessary or simply confirm the visual impression given by the

plot. Also, plotting the data can reveal potential problems with the data, for example, outlying observations which do not fit in the with the general pattern of the data, or data which are clearly wrong.
- For datasets with more than one variable always plot the variables against each other. There may be observations which are not unusual when variables are considered separately but are clearly unusual when 2 variables are plotted. See Section 2.7.
- A good graph draws attention to important aspects of the data. Anything which distracts the viewer, for example, excessive shading, symbols, 3-dimensional effects, should be removed.

- Axis labels (remember the units!), the legend and caption should enable the viewer of the graph to understand the content of the graph.

2.5.1 Histograms

A **histogram** is a graphical display based on the relative frequency distribution of a set of data on a **continuous** variable. The variable of interest is plotted on the x-axis, which is divided into **bins** based on the classes of the frequency distribution. A rectangle is plotted for each bin. The height of a rectangle is calculated as

$$\text{height} = \frac{\text{relative frequency}}{\text{bin width}} = \frac{\text{frequency}}{n \times \text{bin width}}. \tag{2.2}$$

Therefore, for a given box in the histogram:

- the **area** represents the relative frequency of the observations in the interval;
- the **height** represents the relative frequency of the observations in the interval **per unit of measurement**, commonly known as the **density**.

The total area under a histogram is equal to 1. In fact a histogram is an estimate of a probability density function (see Section 4.2). The vertical axis is commonly labelled **density**.

It is common for people to plot frequencies (rather than relative frequencies per unit), giving what I will call a **frequency plot**, that is, **not** a true histogram.

If the bin widths are equal the shape of frequency plot is the same as the corresponding histogram. However, when drawing a frequency plot using unequal bin widths it is important to take into account the differing widths of the bins. For example, in the plot on the bottom left of Figure 2.4, the frequency in the box for 12-20 hours is 4 times too high because the longer width of this interval has not been taken into account. One solution is to divide the frequencies by the bin width to produce frequencies **per unit**, but then we may as well produce a true histogram, using equation (2.2).

A histogram can be useful to look at the shape of a distribution. However, especially for a small dataset, the shape of the histogram can depend greatly on the classes chosen to define the bins. Figure 2.4 contains histograms and frequency histograms of the all the times in Table 2.1. From the histograms we can easily see that the birth times data are slightly positively skewed.

2.5.2 Stem-and-leaf plots

A **stem-and-leaf plot** (or stem plot) is like an enhanced histogram. The **stem**, on the left, contain the times in whole hours. The **leaves**, on the right, contain the first digit after the decimal point. An advantage of a stem-and-leaf plot is that it gives the entire dataset (perhaps rounded) in sorted order. It is easy to calculate the five number summary of the data from a stem-and-leaf plot. Figure 2.5 shows a stem-and-leaf plot of the Oxford births data.

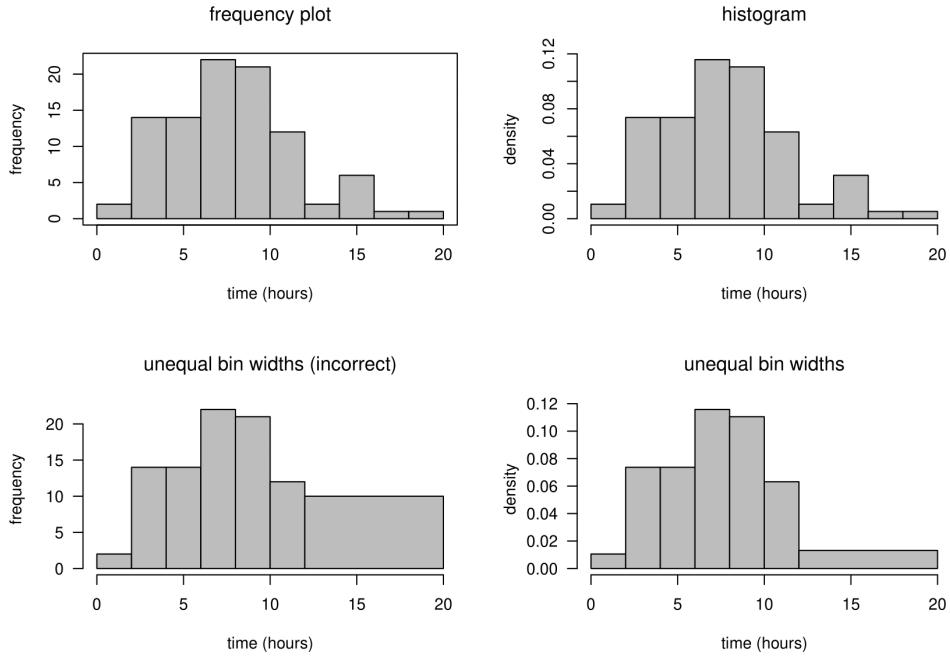

Figure 2.4: Frequency plots (left) and histograms (right) of the Oxford birth times.

1 | 5 = 1.5 hours

leaf unit = 0.1 hours

19 | 0 = 19.0 hours

1	5
2	0155678
3	44466
4	00012233779
5	035679
6	133445589
7	0023333355558
8	123335555589
9	02335588
10	0123444778
11	025
12	89
13	
14	3356
15	0
16	05
17	
18	
19	0

Figure 2.5: Stem-and-leaf plot of the Oxford birth times. The decimal point is at the vertical line |. The data are rounded to the nearest 0.1 before plotting.

2.5.3 Dotplots

Dotplots are simple plots in which each observation is represented by a dot. If there are repeated observations, that is, observations with the same value (perhaps after rounding), then their dots are stacked on top of each other. Figure 2.6 shows two dotplots. In the right hand plot the data were rounded to the nearest hour, producing a plot that looks a bit like a histogram with a bin width of 1 hour, with rectangles replaced by vertical lines of dots.

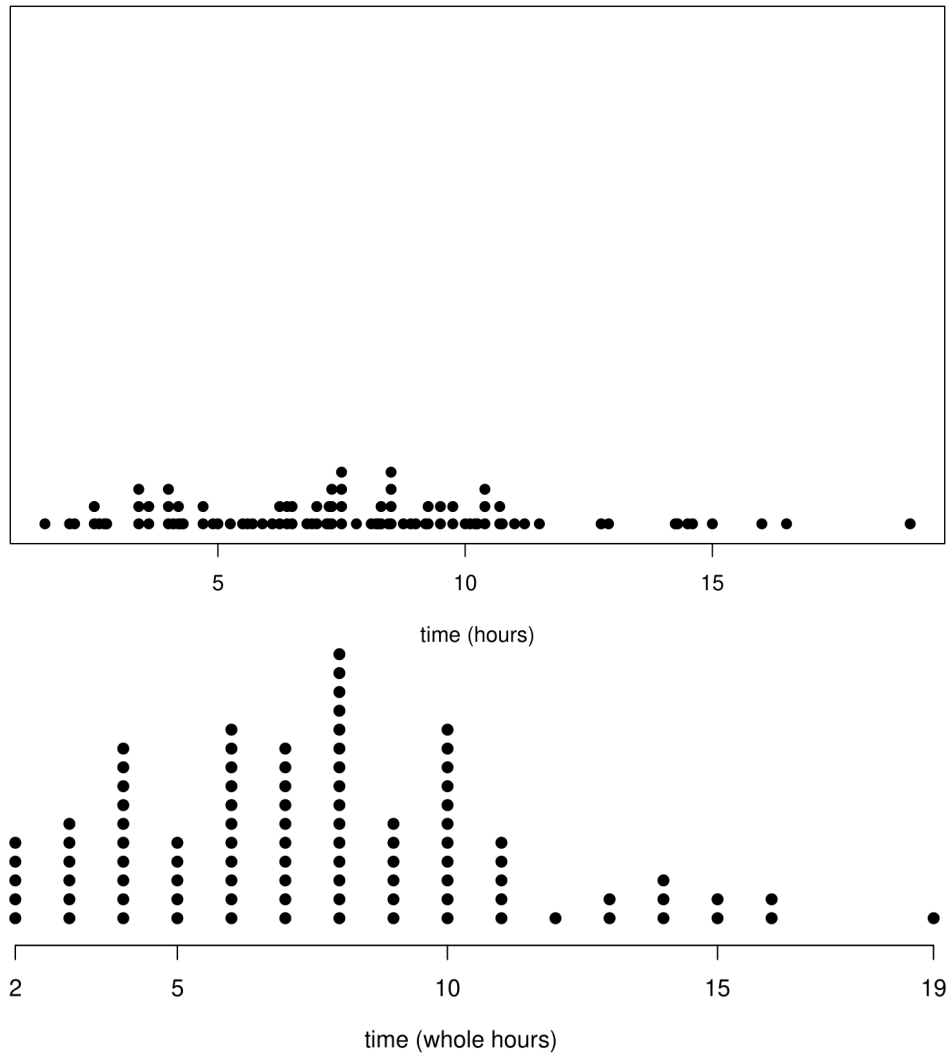

Figure 2.6: Dotplots of the Oxford birth times. Top: raw data (and lots of wasted white space). Bottom: data rounded to the nearest hour.

2.5.4 Boxplots

A **boxplot** (or box-and-whisker plot) is a graphical display containing the five-number summary. It also provides ways to assess the overall shape of the data. Figure 2.6 explains how a standard boxplot is constructed.

The 'box' shows where 50% of the data lie, that is, between the lower and uper quartiles. The 'whiskers' extend to the most extreme observations that are within 1.5 IQR of the ends of the box. Sometimes a different criterion is used to determine the ends of the whiskers. Any more extreme values are individually identified (with a dot here).

It can less easy easy to come to a conclusion concerning the nature of skewness using a boxplot than using a histogram. In this example the lengths of the whiskers and the presence of the value at 19 hours

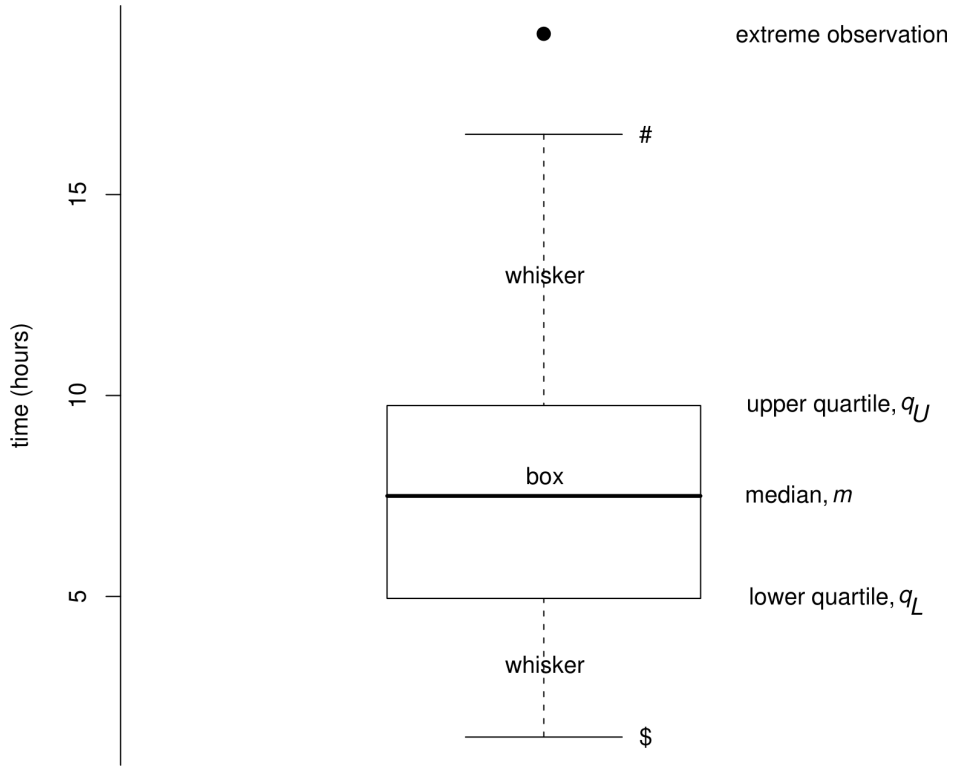

Figure 2.7: Boxplot of the Oxford birth times. The upper end (#) of the upper whisker is drawn at the largest observation within a distance $1.5(q_U - q_L)$ of q_U. The lower end ($) of the lower whisker is drawn at the smallest observation within a distance $1.5(q_U - q_L)$ of q_L.

Table 2.9: Frequencies of numbers of damaged O-rings for the space shuttle data.

number of damaged O-rings	0	1	2	3
frequency	16	5	1	1

suggest slight positive skewness. However, the relative positions of the samples quartiles suggest (very slight) negative skewness, which is the cause of the slightly negative value of sample quartiles skewness towards the end of Section 2.3.4.

Some alternatives are given in Figure 2.8. Which do you prefer?

2.5.5 Barplots

A **barplot** (or bar chart) has a similar appearance to a histogram but is used for numerical discrete data or categorical data. Therefore there are gaps between the bars in a barplot.

Example: numerical discrete data

Table 2.9 shows the frequencies of the number of damaged O-rings in the space shuttle example. Figure 2.9 shows barplots (or equivalent) of these data. Which do you prefer?

Example: categorical data

Table 2.10 shows the percentages of people in the UK with the 8 main blood groups O+, A+, B+, AB+, O−, A−, B− and AB−. See section 3.7.1 for more details about blood groups. These data are nominal.

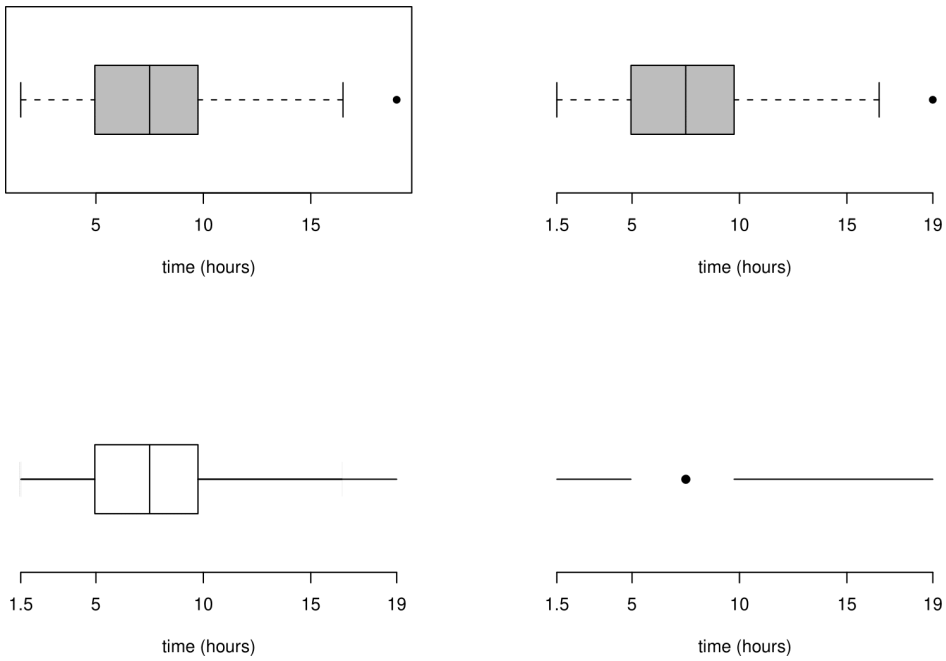

Figure 2.8: Alternative plots of the Oxford birth times based on the five-figure summary.

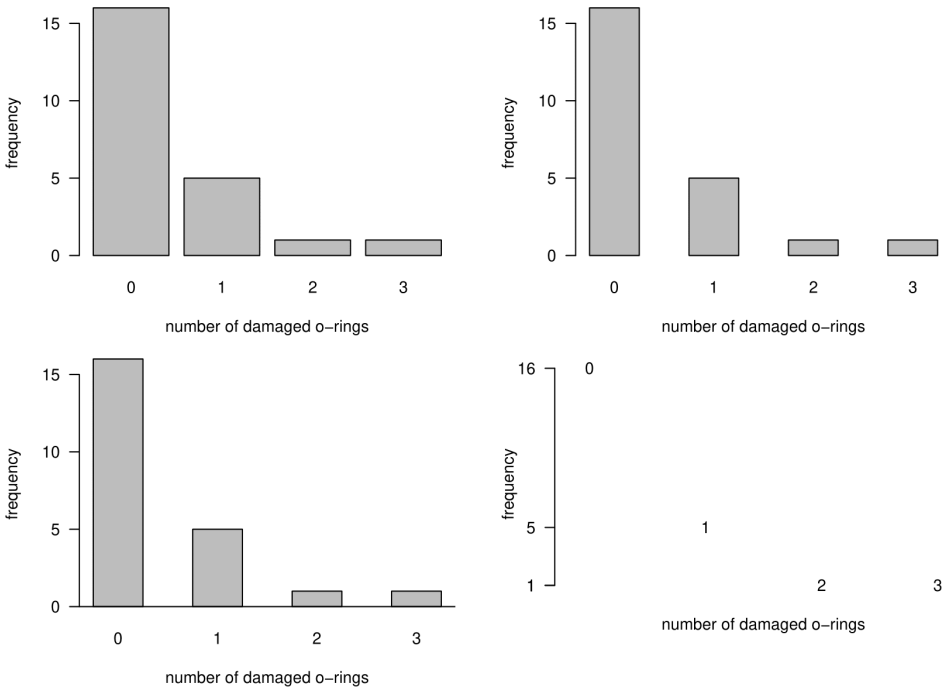

Figure 2.9: Barplots of numbers of damaged O-rings on space shuttle flights.

Table 2.10: Percentages of people in the UK with the 8 main blood groups

blood type	Rh+	Rh-	total
O	37	7	44
A	35	7	42
B	8	2	10
AB	3	1	4
total	83	17	100

Figure 2.10 displays these percentages in the form of a barplot. 2.11 does this using a pie chart. Note that in the barplot we have sorted the categories, separately within the + and − blood groups, in decreasing order of frequency. Do you prefer the table, the barplot or the pie chart? (Please do **not** choose the pie chart!)

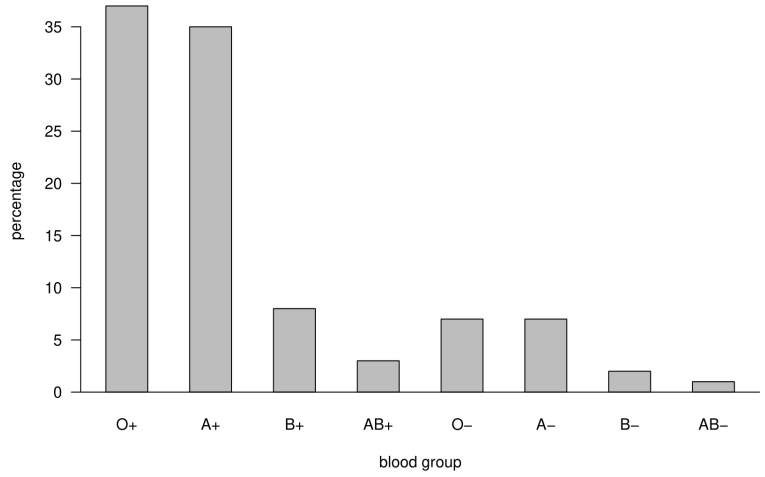

Figure 2.10: Barplot of the UK ABO blood group percentages.

2.5.6 Times series plots

The top plot of Figure 2.12 shows a **time series plot** (or time plot) of the weekly closing prices of the FTSE 100 share index from 2nd April 1984 to 13th August 2007. The bottom plot in this figure shows a different version of the same plot.

When observations are a time series, that is they are in time order, it is important to plot them against time to look for patterns. The sort of features that often turn up are upward or downward trends, or cyclical behaviour (alternative increases and decreases, often the result of seasonal behaviour), but you may see other aspects worth noting. Note that

- time should be plotted on the horizontal axis;
- the plot should be wider than it is high;
- joining the dots can help to make interesting patterns easier to see.

Figure 2.13 shows a time series plot of another set of data. Can you guess what these data might be?

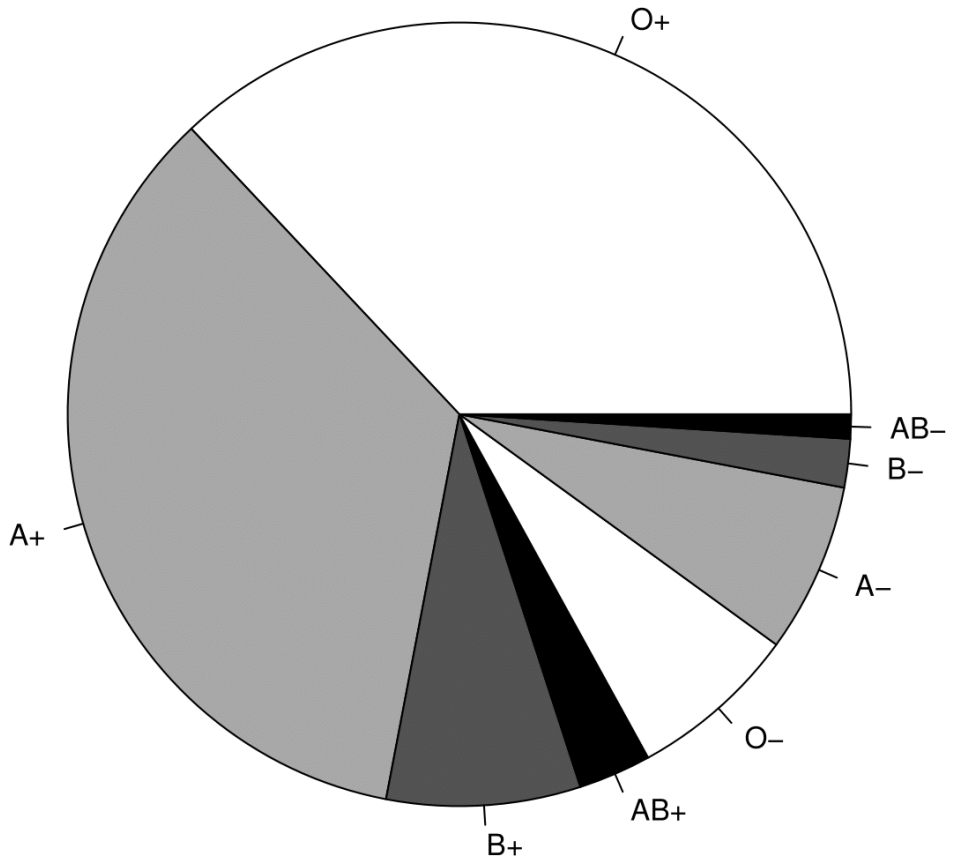

Figure 2.11: Pie chart (right) of the UK ABO blood group percentages.

2.6 2000 US Presidential Election

Smith, R. L. (2002) A Statistical Assessment of Buchanan's Vote in Palm Beach County. Statistcal Science, **17**(4), 441–457.

In the 2000 U.S. Presidential election George W. Bush, the Republican candidate, narrowly beat Al Gore, the Democrat candidate. The result in the state of Florida was particularly close: Al Gore lost by only 537 votes out of 6 million votes cast. If Al Gore had won in Florida he would have become the U.S. President. After the election many allegations of voting irregularities were made and it was a month before Al Gore conceded defeat.

One of the results which caused most surprise was in Palm Beach County, Florida. Pat Buchanan, the Reform Party candidate, got an unexpectedly large 3,407 votes. Based on results in Florida as a whole only 1,300 votes would be expected. Also, given that Palm Beach is largely a Democratic County, a right-wing candidate such as Buchanan would expect even fewer votes.

In the days following the election it was suggested that the type of ballot paper, a so-called Butterfly Ballot 2.14 used in Palm Beach had confused voters and lead to votes being cast for Buchanan by mistake. People found the Buchanan vote in Palm Beach surprising and there is a plausible explanation for how it occurred.

Smith (2002) uses election results, and other data (on race, age, education and income), from Florida to answer the following questions:

1. Is Buchanan's vote of 3,407 very clearly out of line with the pattern of results form the rest of Florida? In Statistics we call such data values **outliers**.
2. What level of vote for Buchanan would have been realistic in Palm Beach County?

Figure 2.15 suggests that the answer to question 1. is "Yes".

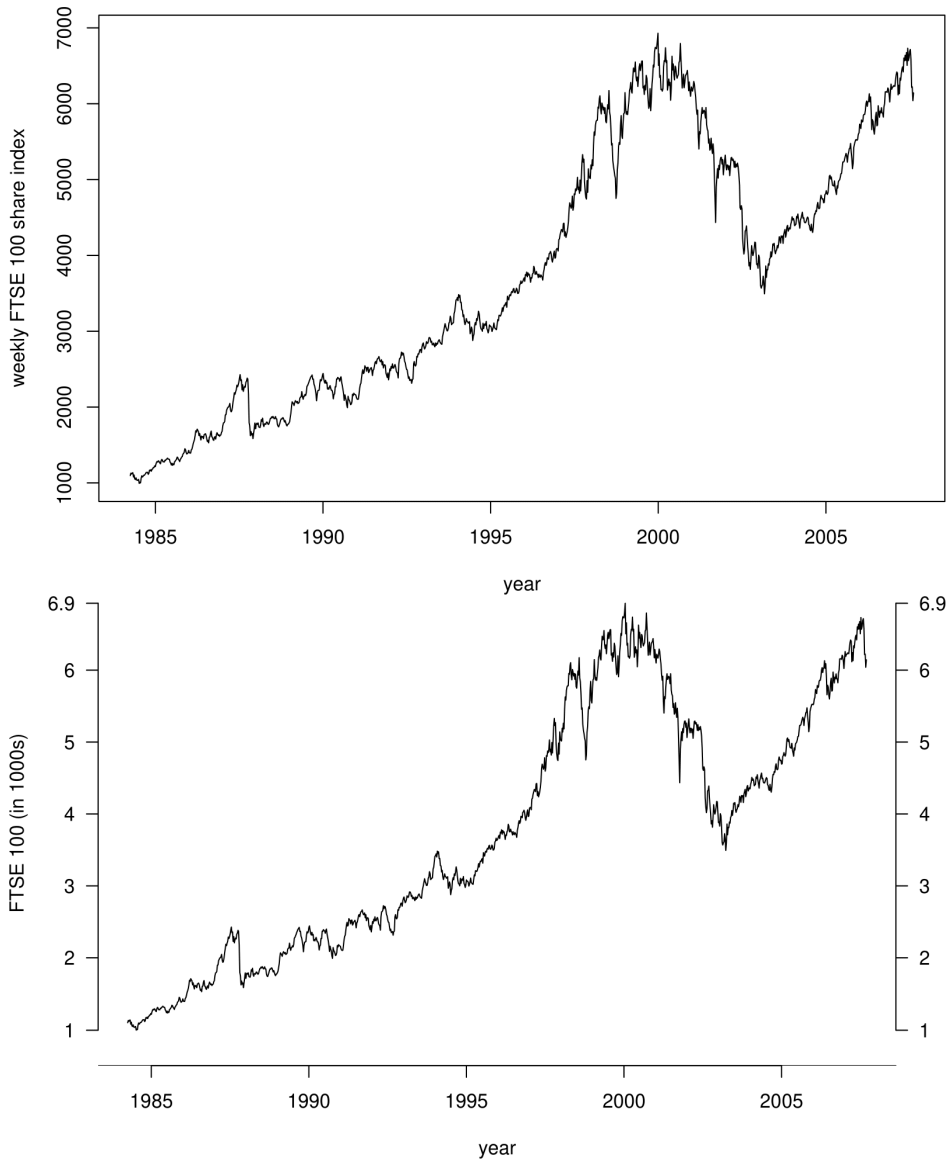

Figure 2.12: Time series plots of the FTSE 100 weekly closing values, 1984–2007. Top: default plot. Bottom: modified version, with two vertical axes and the index measured in 1000s.

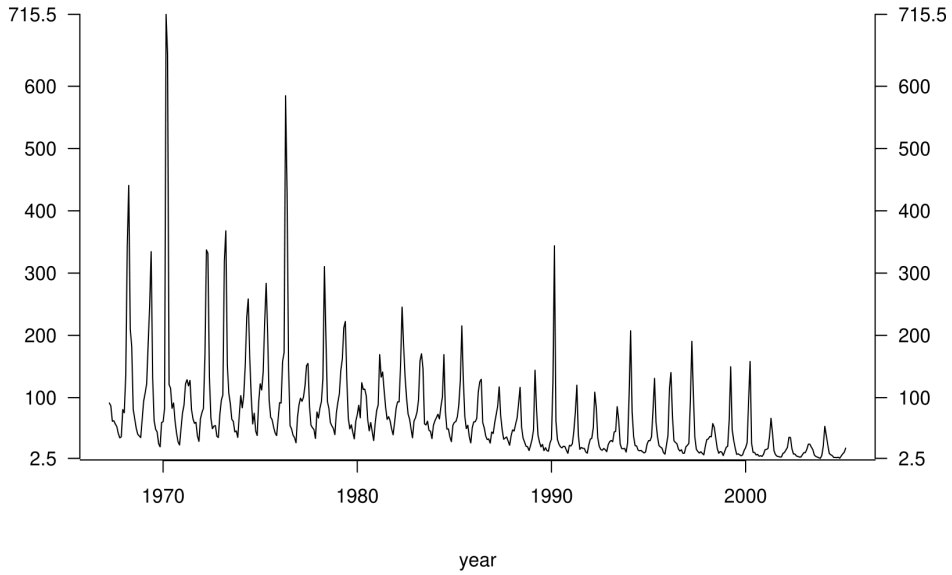

Figure 2.13: A time series plot of ?.

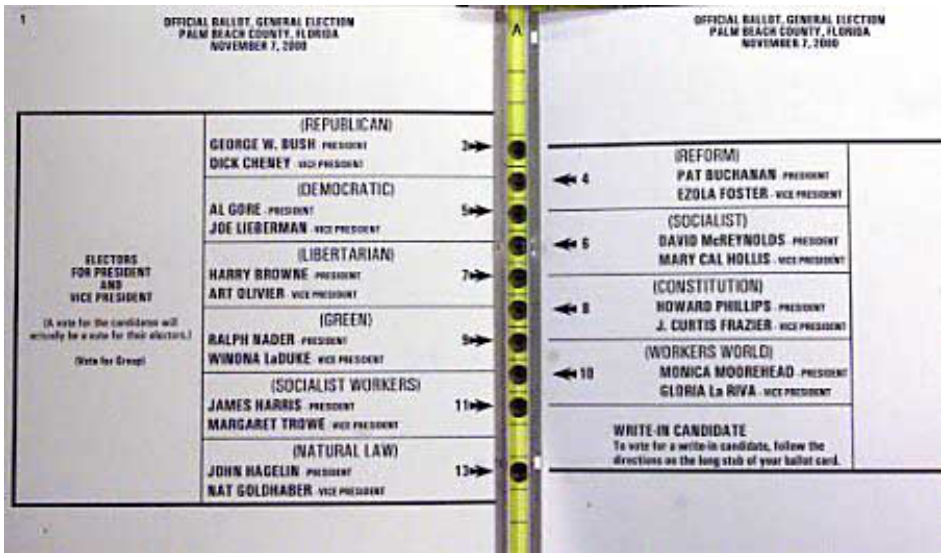

Figure 2.14: The Butterfly Ballot used in Palm Beach county.

On several of these plots Palm Beach stands out as a clear outlier. In these cases Buchanan gets many more votes than the pattern of the other points would suggest. We also see that the percentage of the vote that Buchanan gets tends to

- decrease with population size;
- decrease with the percentage of of Hispanics;
- decrease with the percentage of voters aged 65 or over;
- decrease with high school and college graduation rate;
- decrease with mean personal income;
- decrease with the percentage of Gore votes;
- increase with the percentage of Bush votes.

Smith (2002) answers questions 1. and 2. more formally by building a linear regression model. This model quantifies how the percentage of Buchanan vote Y, the **response variable**, depends on the other variables,the **explanatory variables** $x_1, \ldots, x_{12}$. The general idea is to

- build the model using all the data for Florida, apart from the data from Palm Beach, using only the explanatory variables that have a significant effect;
- predict the value of the Buchanan's vote in Palm Beach using the model.

We will study simple linear regression models (with only one explanatory variable) towards the end of STAT0002 (Chapter 8) and in STAT0003. The basic idea is to assume that a response variable has a linear (straight line) relationship with explanatory variables. The relationship will not be exact, so the model includes a **random error** term.

Smith (2002) finds that transformations are required in order that the assumptions of the model are satisfied approximately. In particular he finds that using the response variable $\sqrt{Y}$ is better than using Y itself (and better than other possible transformations). He also uses a $\log_{10}$ transformation on some of the explanatory variables (for example Total Population), that is, he uses $\log_{10}(x)$ rather than x. Figure 2.16 is a new version of figure 2.15 in which the square root of the percentage of the vote obtained by Buchanan is plotted against the (possibly log-transformed) explanatory variables.

Smith (2002) uses transformations of the original data in order to satisfy more closely the assumptions of the linear regression model:

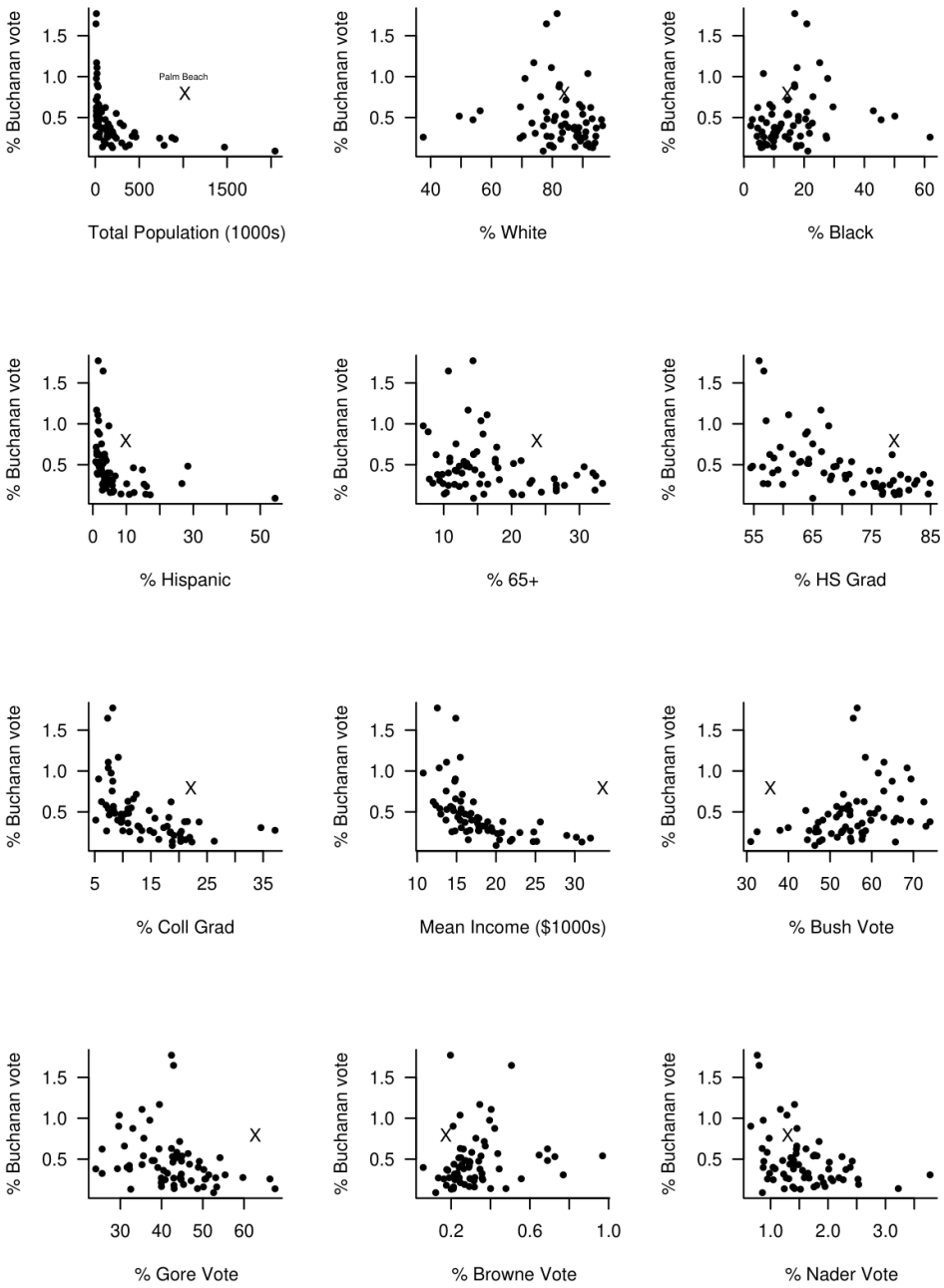

Figure 2.15: Percentage of Buchanan votes against explanatory variables. Palm Beach County is marked with a cross.

- response $= \sqrt{Y}$, instead of Y;
- for some explanatory variables, use $\log_{10}(x)$ instead of x.

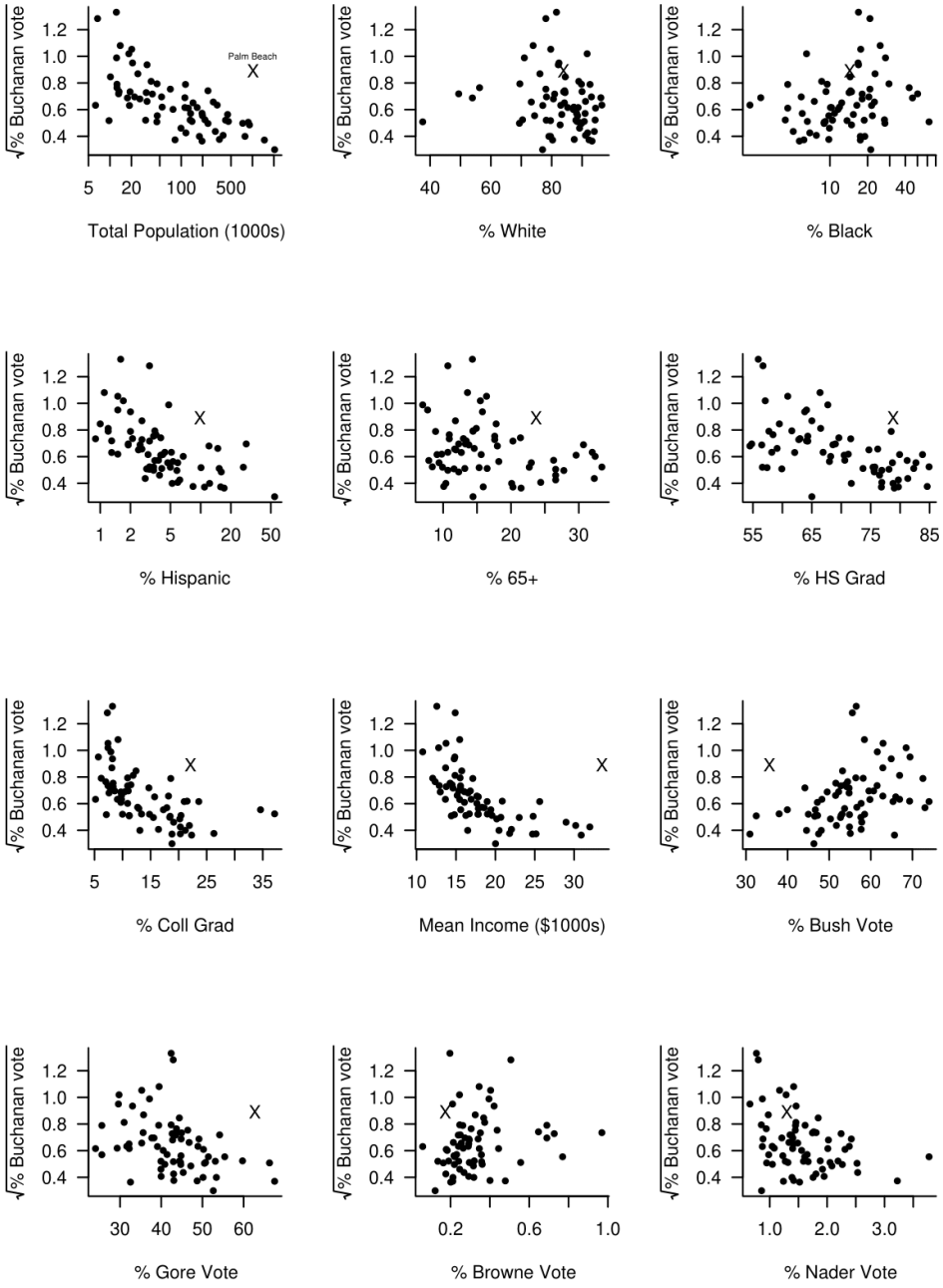

Figure 2.16: The square root of the percentage of Buchanan votes against explanatory variables. Palm Beach County is marked with a cross. Note the log scale on the x-axis of some plots.

Log scales on axes

Suppose that we produce a scatter plot where the data on the x-axis are 0.1, 1, 10, 100 and 1000. If we wish to plot $\log_{10}(x)$ on the axis instead of x we have two choices:

(a) Calculate $\log_{10}(x)$ and plot these values on the axis.
(b) Plot the values of x but on a $\log_{10}$ scale. On a $\log_{10}$ scale the values 0.1, 1, 10, 100 and 1000 are equally spaced. For example, from the basic rules of logs we have

$$\log_{10}(10r) = \log_{10}(10) + \log_{10}(r) = 1 + \log_{10}(r).$$

Therefore, on a $\log_{10}$ scale the values $10r$ and r are 1 unit apart. In other words adding 1 unit on a $\log_{10}(x)$ scale corresponds to **multiplying** by 10 on the original x-scale.

Both a. and b. will give exactly the same pattern of plot. The advantage of b. is that the original values are on the plot rather than the $\log_{10}(x)$ values. Figure 2.17 illustrates this.

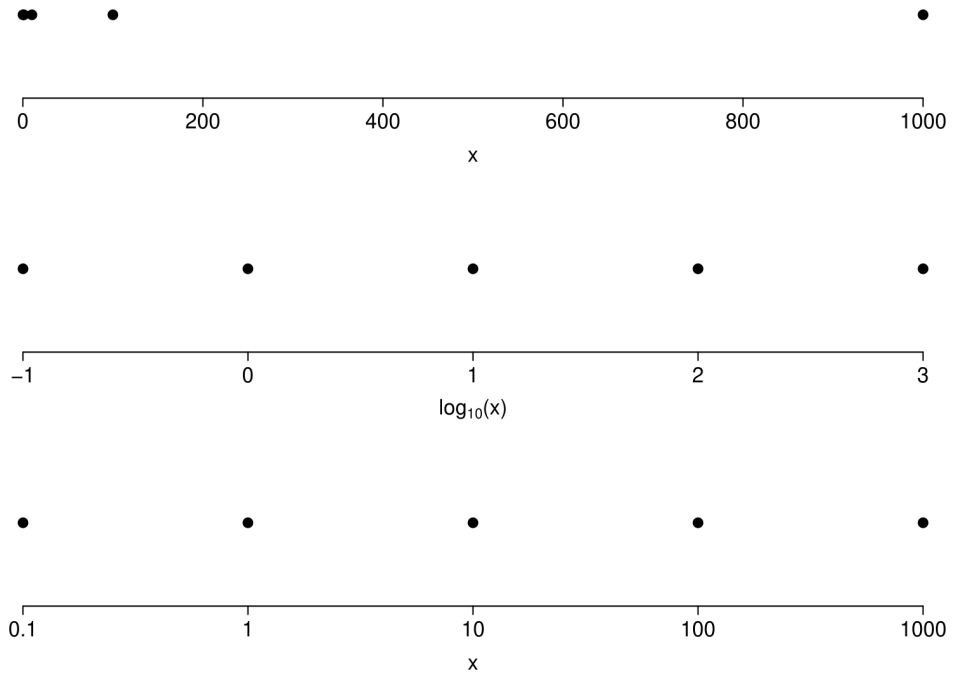

Figure 2.17: Plots to illustrate log-transformation of axes. Top: values of x plotted. Middle: values of $\log_{10}(x)$ plotted. Bottom: values of x plotted on a log-scale.

Other notes on logs:

- we have used logs to base 10 for simplicity but the base doesn't matter;
- logs are often helpful when the raw data are ratios (e.g. x/y) or products (e.g. xy). For example, exchange rates and price indices are ratios. If $x = y$ then $\log(x/y) = 0$; if $x = k\,y$ then $\log(x/y) = \log k$; if $y = k\,x$ then $\log(x/y) = \log(1/k) = -\log k$; which is a nice symmetry.

Imagine that the model has only one explanatory variable, Total Population. You can imagine fitting this linear regression model as drawing a line of best fit through the points on the graph in the top left hand corner of figure 2.16. With more than one explanatory variable it is more complicated than this but the basic idea is the same.

After removing the Buchanan vote in Palm Beach (which we have decided is an outlier) Smith (2002) finds that the model fits the data well.

The model predicts the Buchanan vote in Palm Beach to be 371, much lower than the official result of 3,407. This number (371) represents the 'best guess' at the Buchanan vote given the other data. To show just how unlikely was the vote of 3,407, Smith (2002) calculates a 95% prediction interval of (219,534) for the Buchanan vote at Palm Beach. If the model is true, this interval has a probability of 95% of containing the true value of the Buchanan vote.

Smith's analysis suggests that the true Buchanan vote should be approximately 3,000 votes lower than the official result. Given the design of the Butterfly Ballot it seems likely that most of these votes were intended for Al Gore. This would have given Gore the presidency instead of Bush.

2.7 Graphs (2 variables)

When we have 2 continuous variables it is common to examine the relationship between them using a **scatter plot**.

2.7.1 Scatter plots

We have already seen some scatter plots in the 2000 US Presidential Election example. We reproduce two of these plots in Figures 2.18 and 2.19. A scatter plot is used to examine the relationship between two variables. We need the data to occur in pairs. In Figures 2.18 and 2.19 each county has a pair of observations: the percentage of votes for Buchanan and the value of the explanatory variable.

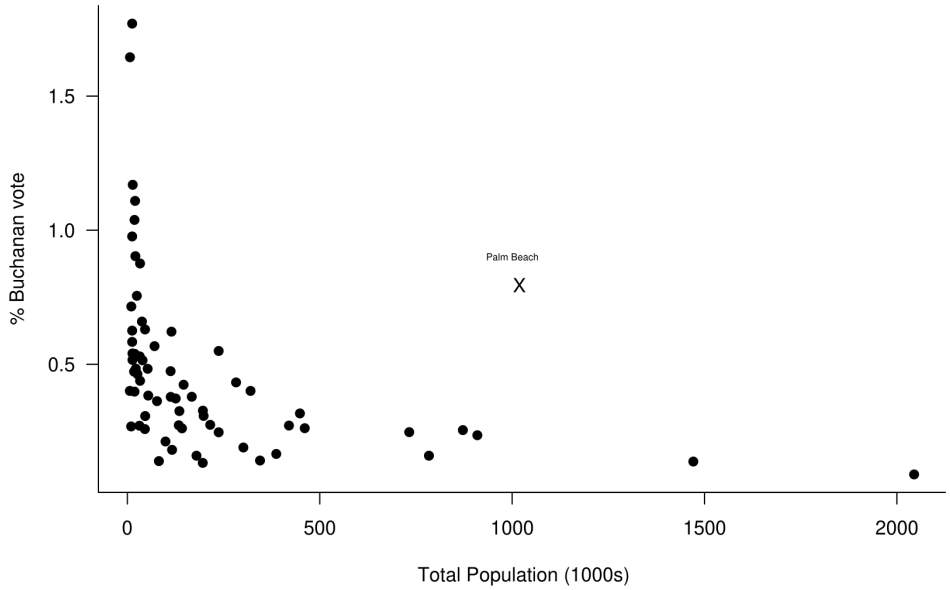

Figure 2.18: Scatter plot of the percentage of the vote obtained by Buchanan against the total population from the 2000 US Presidential Election data.

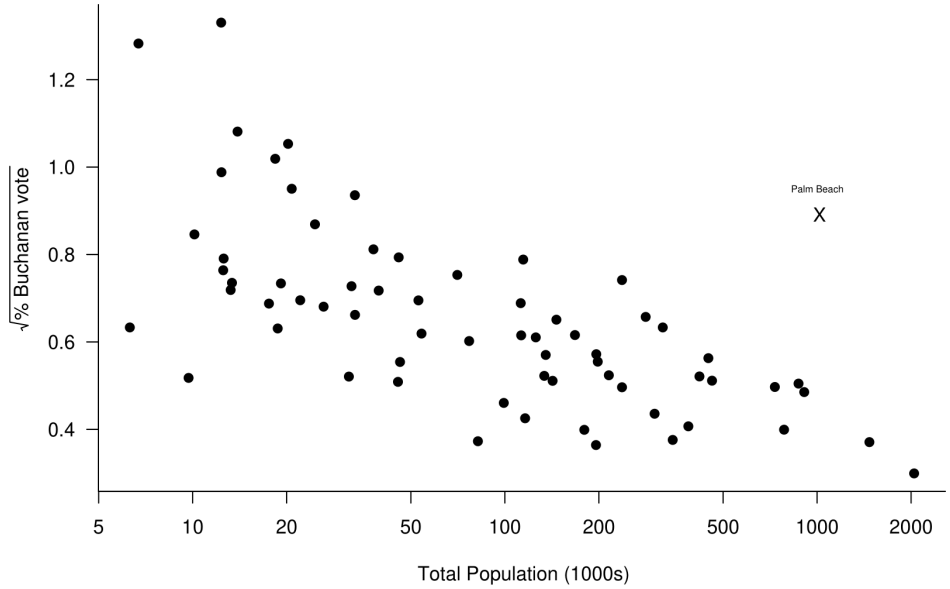

Figure 2.19: Scatter plot of the square root of the percentage of the vote obtained by Buchanan against the log of the total population from the 2000 US Presidential Election data. The plot suggests that these variables are aproximately linearly related.

Notice that we have plotted % Buchanan vote (on the vertical y-axis) against total population (on the horizontal x-axis). This is because it makes sense that % Buchanan vote depends on total population, that is, the size of population influences the vote, not the other way round.

Rules for deciding which variable to plot on the y-axis and which on the x-axis are:

- If the direction of dependence is clear, so that variable Y depends on variable X. For example, X=river depth influencing Y=flow rate.
- If one variable, X, is fixed by an experimenter and then the value of another variable, Y is observed. For example, X=dosage of drug and Y=reduction in blood pressure.
- If we wish to predict one variable, Y, using another, X. For example, X=share value today and Y=share value tomorrow.

It is clear in both these plots that the vote in Palm Beach is an outlier. However, if we had produced separate plots of % Buchanan vote and total population Palm Beach would not appear as an outlier.

2.8 Transformation of data

Some simple statistical methods are based on assumptions about the statistical properties of the data to which they are applied. For example, there are methods that work well provided that a variable of interest is approximately symmetrically distributed. If a variable has a distribution that is strongly skewed then the method will not have the properties that are expected and the results may be misleading. In linear regression (see Chapter 8) the mean of one variable is represented as being related linearly to the value of another variable. If the reality is that this relationship is far from being linear then results may be very misleading.

If we wish to make use of simple assumptions like symmetry of distribution and/or linearity of relationship, but it is clear that the raw data do not support these assumption, then a legitimate approach is to consider whether the assumptions are satisfied better after we transform the data. We illustrate this idea in Sections 2.8.1 and 2.8.2.

2.8.1 Transformation to approximate symmetry

The data in Table 2.11 resulted from an experiment (Simpson et al. (1975)) to see whether spraying silver nitrate into a cloud (known as **seeding** the cloud) could make it produce more rainfall. 52 clouds were studied. 26 of the clouds were chosen at random and seeded with silver nitrate. The amounts of rainfall, in acre-feet, produced by each cloud is recorded. (An acre-foot is a unit of volume equal to 43,560 feet^3 or, approximately, 1233.5m^3.)

Figure 2.21 shows separate boxplots of the rainfall amounts from seeded and unseeded clouds.

It is clear from the shape of these plots that the data are positively skewed. Also, the sample means are much greater than their corresponding sample medians. Measurements of (positive) environmental quantities are often positive skew. In addition, the rainfall values from the seeded clouds have a both higher location and a higher spread than the values from the unseeded clouds. After a log transformation (see 2.21), the data are closer to being approximately symmetric. The sample means are closer to their corresponding sample medians. In addition the log transformation makes the variances of the rainfall values in the two groups more nearly equal.

We have used a log transformation make positive skew data more symmetric. Other transformations which can be useful for this purpose are: y^c, where $c < 1$, for example, $\sqrt{y}$, $1/y$. These transformations stretch out the lower tail. In contrast, y^c, where $c > 1$, e.g. y^2, y^3, may be used to transform negative skew data to approximate symmetry. These transformations stretch out the upper tail. It may seem that the log transformation is of an entirely different form to the other transformations, that is, y^c for some

Table 2.11: The rainfall, in acre-feet, from 52 clouds, 26 of which were chosen at random to be seeded with silver nitrate.

unseeded		seeded	
1202.6	2745.6	200.7	200.7
830.1	1697.8	198.6	198.6
372.4	1656.0	129.6	129.6
345.5	978.0	119.0	119.0
321.2	703.4	118.3	118.3
244.3	489.1	115.3	115.3
163.0	430.0	92.4	92.4
147.8	334.1	40.6	40.6
95.0	302.8	32.7	32.7
87.0	274.7	31.4	31.4
81.2	274.7	17.5	17.5
68.5	255.0	7.7	7.7
47.3	242.5	4.1	4.1

$c \neq 0$. However, we will see that a log transformation can be obtained by considering the behaviour of the equivalent transformation $(y^c - 1)/c$ as c approaches zero.

It is possible to transform using y^c for **any** real value of c, but it is better to stick to simple powers, such as the ones above, as it is more likely that these will have a sensible interpretation. The further c is from 1 the more difference the transformation makes.

These rainfall data are positive so there is no problem using a transformation of the form y^c. However, if a dataset contains negative values then there are problems. If $y < 0$ then y^c can only be calculated in special cases where c is an integer. We also need to be able to invert the transformation, that is, to infer the value of y uniquely from the value of y^c. If y can be both negative and positive then this is only possible in the very special cases where c is an odd integer. If a dataset contains zeros then we cannot use a log transformation, or y^c for $c < 0$. The main point is that we need all the data to be positive in order to use the transformation y^c (or $(y^c - 1)/c$). If we wish to transform data with non-positive values it is common to add a suitable constant to all values, to produce positive data, before transformation.

The rainfall data range over several orders of magnitude, that is, from one to well over a thousand. Applying a log transformation is often useful when data range over several orders of magnitude.

An aside. If we particularly like stem-and-leaf plots then we could produce a back-to-back stem-and-leaf plot, as in the plots of the log-transformed rainfall totals in Figure 2.22.

2.8.2 Straightening scatter plots

Suppose we have drawn a scatter plot and the general form of the relationship between the variables y and x appears to be monotonic, y tends either to increase or decrease with the value of x, subject to some random scatter about this relationships. However, the relationship between the variables y and x is not even approximately a straight line, that is, it is non-linear, rather than linear.

There are two main reasons why we may want to **straighten out** a scatter plot, that is, make it closer to being linear:

- we may be find it easier to appeciate the relationship between variables when that relationship is linear compared to a case where the relationship is non-linear and more complicated;
- we may be hoping to be able to use a simple method of analysis that requires approximate linearity, for example, linear regression (see Chapter 8).

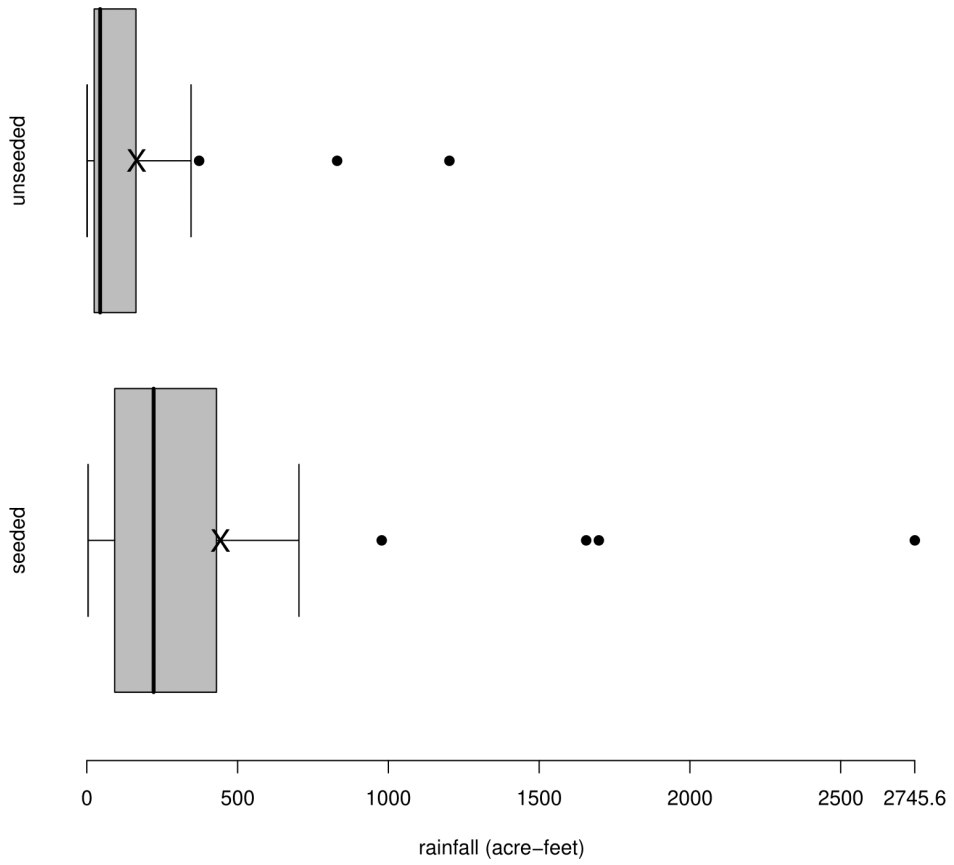

Figure 2.20: Boxplots of rainfall in acre-feet for seeded and unseeded clouds. Sample means are marked with a cross.

How can we straighten a scatter plot?

We could transform the y variable, that is, plot a function of y on the y-axis. As in Section 2.8.1, commonly-used transformations are

$$y^3\,, \quad y^2\,, \quad (y)\,, \quad \sqrt{y}\,, \quad \log y\,, \quad -\frac{1}{\sqrt{y}}\,, \quad -\frac{1}{y}\,, \quad -\frac{1}{y^2}\,, \quad -\frac{1}{y^3}\,,$$

Question: Why might we prefer to use the transformation $-1/y$, rather than $1/y$?

Instead of transforming the y-axis we could transform the x variable, or both the y and x variables. For example, in Figure 2.19 the use of a square root transformation on y and a log transformation on x produces a plot in which the relationship between the two variables is much closer being approximately linear than the variables plotted in Figure 2.18.

If we wish to use transformation to straighten a scatter plot then we have lots of choice about which transformations to try. These days it is easy to use trial-and-error, that is, to try lots of transformations and judge by eye which of the resulting plots we prefer. There are also automatic computational methods to do this. However, before the advent of modern computing, producing plots was more time-consuming and it was helpful to use the shape of the original plot to suggest which transformations might work. One way to do this is sometimes called **Tukey and Mosteller's bulging rule**.

Consider the curve plotted in Figure 2.23. Imagine a straight line drawn between the ends of this curve. In the middle of the curve the values of both y and x are smaller than the points on the imaginary line. We say that the curve is bulging down in both the y and x directions.

Similarly, in Figure 2.24 x is bulging down but now y is bulging up.

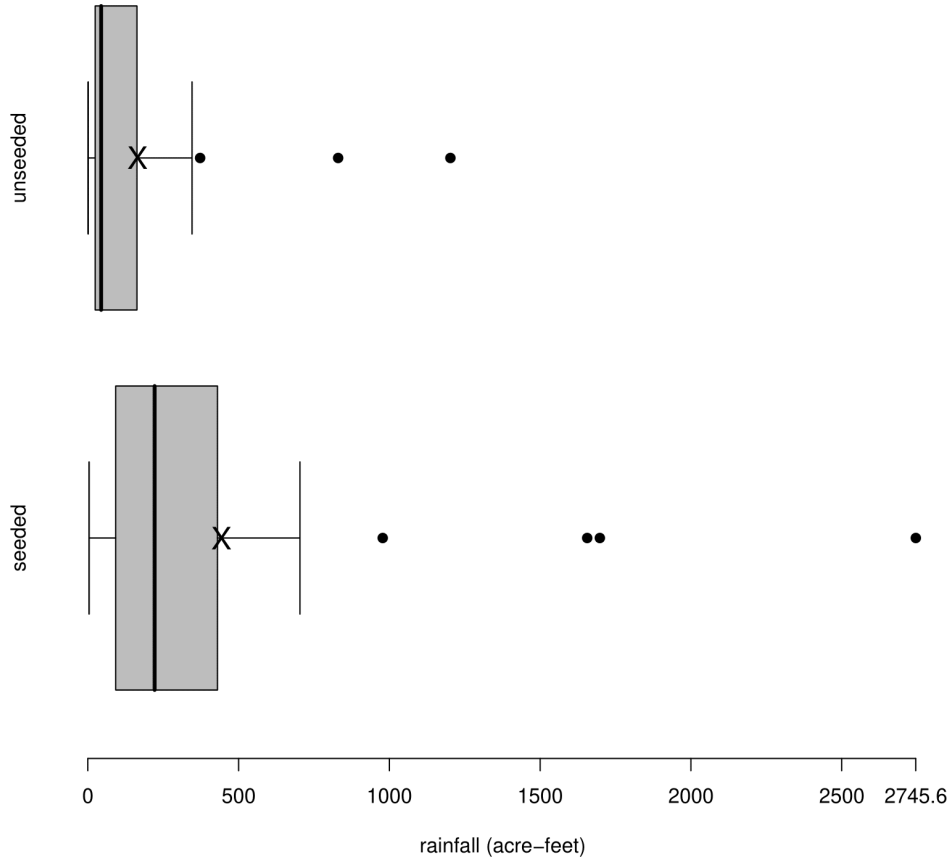

Figure 2.21: Boxplots of rainfall in acre-feet for seeded and unseeded clouds after a $\log_{10}$ transformation has been applied. Sample means are mared with a cross.

unseeded		seeded	
	0	0	0 | 0 = 0.0
7 | 0 =0.7	0	69	
444321	1	2	
99876655	1	556	
4220	2	01111334444	
9655	2	55678	
1 | 3 = 3.1	3	0224	3 | 0 = 3.0

Figure 2.22: Back to back stem-and-leaf plot of $\log_{10}$(rainfall) for the cloud seeding data. The decimal point is at the vertical line |. Leaf unit = 0.1 log(acre feet)

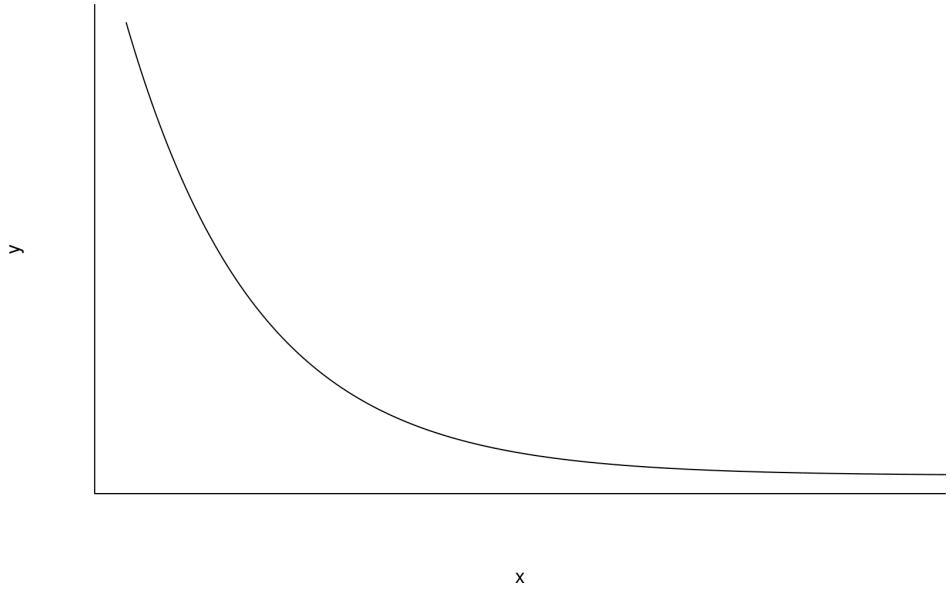

Figure 2.23: A curve in which both y and x are bulging down compared to an imaginary straight line drawn between the ends of the curve.

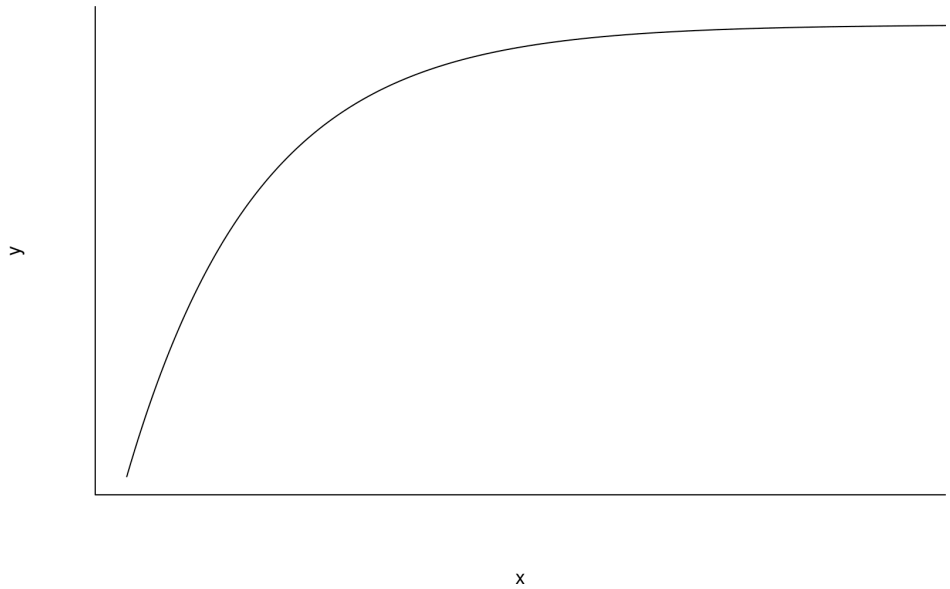

Figure 2.24: A curve in which both y and x are bulging down compared to an imaginary straight line drawn between the ends of the curve.

Suppose that we consider transforming only y. Tukey and Mosteller's bulging rule says that for scatter plots showing relationships like that depicted in Figure 2.23 we should try transformations like $\sqrt{y}$, $\log y$, $-1/\sqrt{y}$, $-1/y$, $-1/y^2$, ..., that is, y^c, for $c < 1$. For cases like Figure 2.23 we should try transformations like y^2, y^3, ..., that is, y^c, for $c > 1$.

Consider Figure 2.25 as an example. The relationship between the variables is similar to the curve in Figure 2.23. Therefore, we should try transforming y using using a transformation like $\sqrt{y}$, $\log y$, ...

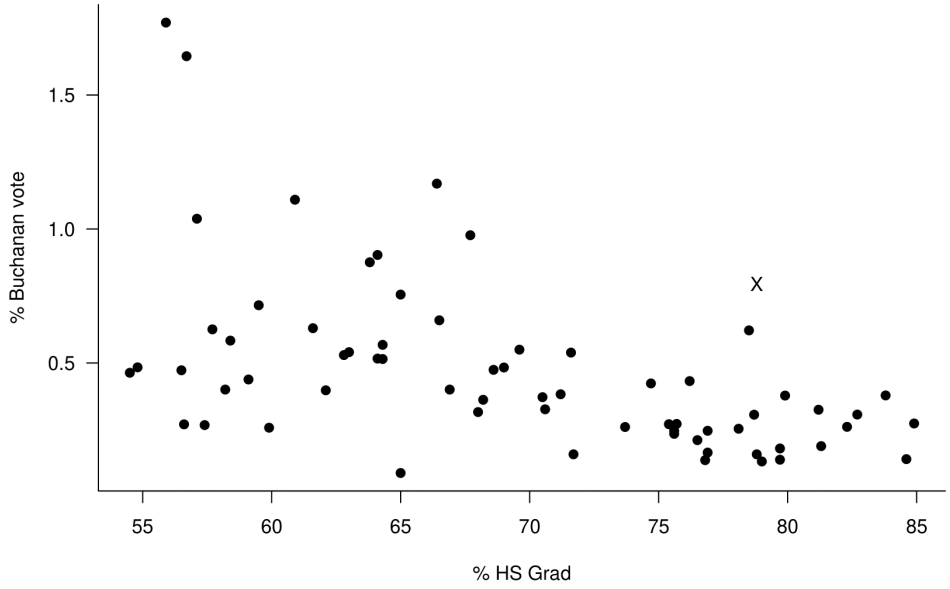

Figure 2.25: Scatter plot of the percentage of the vote obtained by Buchanan against the percentage of the population who graduate from high school the 2000 US Presidential Election data.

The curvature of the relationship shown in 2.25 is not strong, so it makes sense that in Figure 2.26 approximately linearity of relationship is acheived using the relatively weak transformation $\sqrt{y}$.

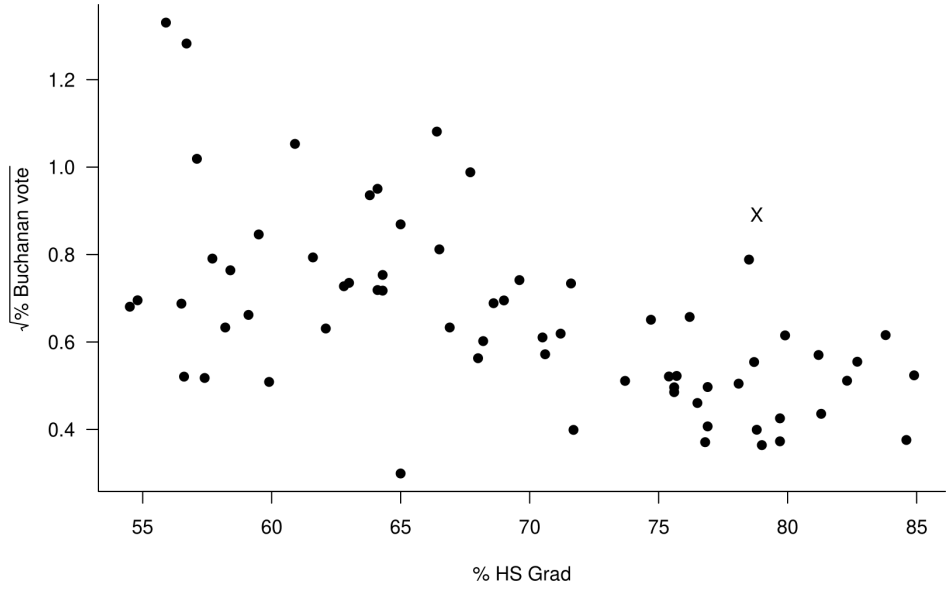

Figure 2.26: Scatter plot of the square root of the percentage of the vote obtained by Buchanan against the percentage of the population who graduate from high school the 2000 US Presidential Election data.

Figure 2.27 shows how Tukey and Mosteller's bulging rule works in the four different bulging cases, considering the possibilities of transforming y only, x only or both y and x. To use this figure first pick the curve that is relevant to the scatter plot in question. The expressions given at the ends of this curve are examples of the kind of transformations that you could try. The general forms of the indicated

transformations are given in the caption to the figure.

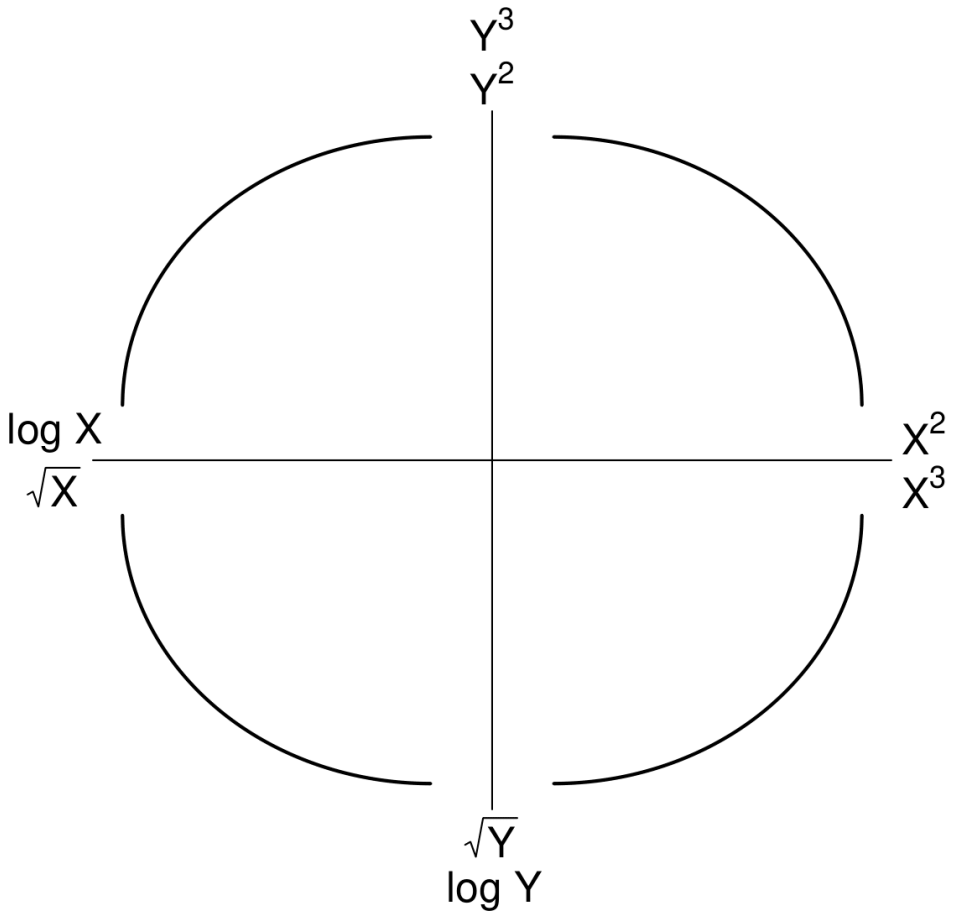

Figure 2.27: Summary of transformations, of the form Y^{c_y} and/or X^{c_x}, to try. Bottom left: $c_y < 1, c_x < 1$. Top left: $c_y > 1, c_x < 1$.Top right: $c_y > 1, c_x > 1$.Bottom right: $c_y < 1, c_x > 1$.

Linearity is not the only consideration

Although linearity can be important other things can be important too. Suppose that we draw a 'line-of-best-fit' on a scatter plot which looks approximately linear. Figure 2.28 is a copy of Figure 2.26 with such a line superimposed. In Chapter 8 we will see that in a simple linear regression model it is assumed that the amount of (vertical) scatter in the y direction of points about a line of best fit is the same for all values of the explanatory variable x.

In Figure 2.28 there is perhaps a greater spread of points about the line for small values of % HS Grad than for large values of % HS Grad.

In Section 8.4 we consider how to use transformation of y and/or x to satisfy better the assumptions of a linear regression model.

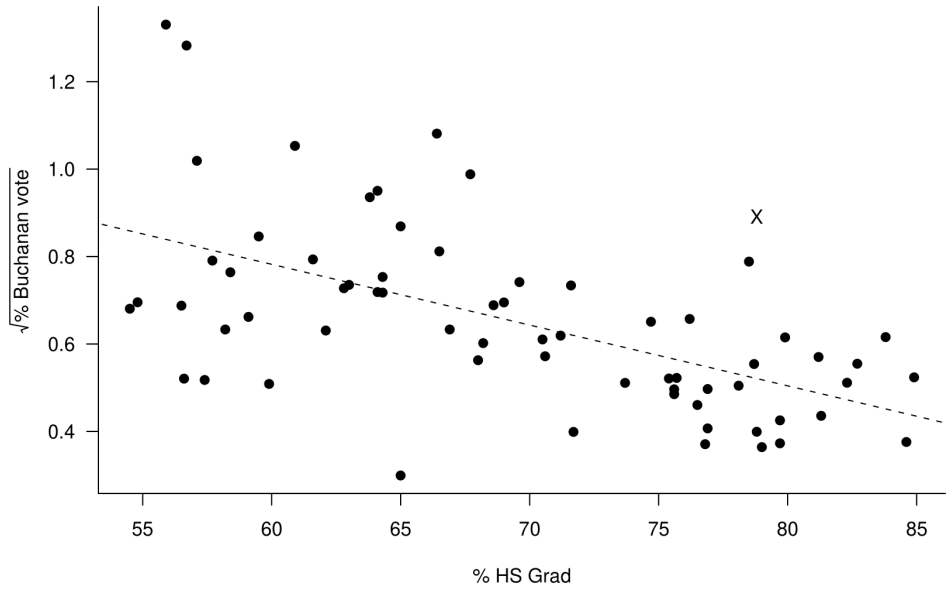

Figure 2.28: Scatter plot of the square root of the percentage of the vote obtained by Buchanan against the percentage of the population who graduate from high school the 2000 US Presidential Election data. A (dashed) line-of-best-fit is superimposed.

Chapter 3

Probability

Most people have heard the word **probability** used in connection with a random experiment, that is, an experiment whose outcome cannot be predicted with certainty, such as tossing a coin, tossing dice, dealing cards etc. We start by considering a criminal case in which fundamental ideas surrounding the use of probability were hugely important. Then we study the concept of probability using the traditional simple example of tossing a coin.

3.1 Misleading statistical evidence in cot death trials

In recent years there have been three high-profile criminal cases in which a mother has been put on trial for the murder of a her babies. In each case the medical evidence against the woman was weak and the prosecution relied heavily on statistical arguments to make their case. However, these arguments were not made by a statistician, but by a medical expert witness: Professor Sir Roy Meadows. However, there were two problems with Professor Meadows' evidence: firstly, it contained serious statistical errors; and secondly, it was presented in a way which is likely to be misinterpreted by a jury. To illustrate the error we consider the case of Sally Clark.

Sally Clark's first child died unexpectedly in 1996 at the age of 3 months. Sally was the only person in the house at the time. There was evidence of a respiratory infection and the death was recorded as natural; a case of **Sudden Infant Death Syndrome (SIDS)**, or **cot death**. In 1998 Sally's second child died in similar circumstances at the age of 2 months. Sally was then charged with the murder of both babies. There was some medical evidence to suggest that the second baby could have been smothered, although this could be explained by an attempt at resuscitation.

It appeared that the decision to charge Sally was based partly on the reasoning that cot death is quite rare so having two cot deaths in the same family must be very unlikely indeed. This is the basis of Professor Meadows' assertion that: "One cot death is a tragedy, two cot deaths is suspicious and, until the contrary is proved, three cot deaths is murder.". At her trial in 1999 Sally Clark was found guilty of murder and sentenced to life imprisonment.

Professor Meadows' statistical evidence

At Sally Clark's trial in 1999 Professor Meadows claimed that, in a family like Sally's (affluent, non-smoking with a mother aged over 26), the chance of two cot deaths is 1 in 73 million, that is, a probability of $1/73{,}000{,}000 \approx 0.000000014$. Professor Meadows had calculated this value based on a study which had estimated the probability of one cot death in a family like Sally's to be 1 in 8543, that is, 1 cot death occurs for every 8543 of such families.

Professor Meadows had then performed the calculation

$$\frac{1}{8543} \times \frac{1}{8543} = \frac{1}{72,982,849} \approx \frac{1}{73,000,000}.$$

There are problems with this evidence, both with this calculation and with the idea that this apparently small number provides evidence of guilt.

Can you identify these problems?

3.2 Relative frequency definition of probability

Example: tossing a coin

If you toss a coin, the outcome (the side on top when the coin falls to the ground) is either a Head (H) or a Tail (T). Suppose that you toss the coin a large number of times. Unless you are very skillful the outcome of each toss depends on chance. Therefore, if you toss the coin repeatedly, the exact sequence of Hs and Ts is not predictable with certainty in advance. This is usually the case with any experiment. Even if we try very hard to repeat an experiment under exactly the same conditions, there is a certain amount of variability in the results which we cannot explain, but we must accept. The experiment is a **random experiment**.

Nevertheless, if the coin is fair (equally balanced), and it is tossed fairly, we might expect the long run **proportion**, or **relative frequency**, of Hs to settle down to $1/2$. However, the only way to find out whether this is true is to toss a coin repeatedly, forever, and calculate the proportion of tosses on which H is the outcome. It is not possible, in practice, for any experiment to be repeated forever.

However, a South African statistician Jon Kerrich managed to toss a coin 10,000 times while imprisoned in Denmark during World War II. At the end of his effort he had recorded 5067 Heads and 4933 Tails. Figure 3.1 shows how the proportion of heads Kerrich threw changed as the number of tosses increased.

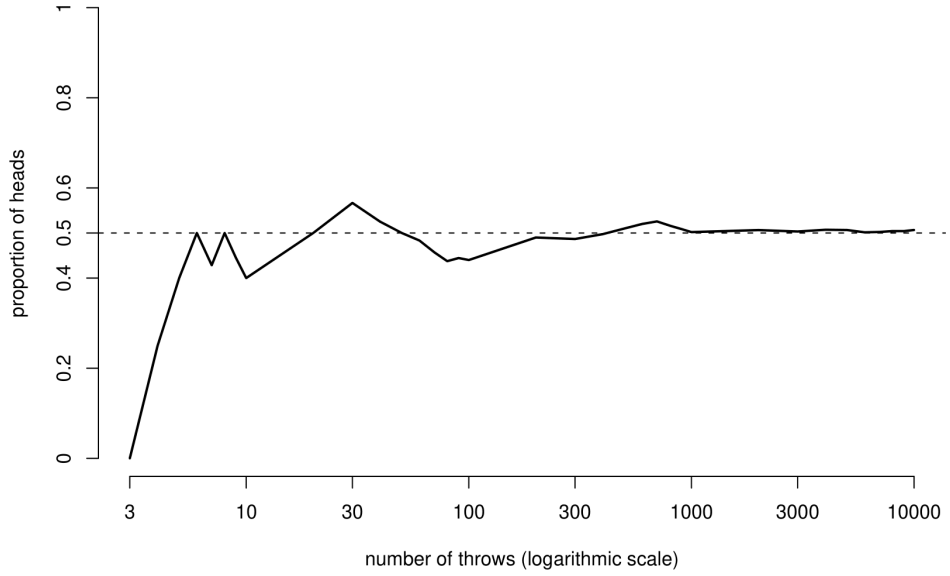

Figure 3.1: The proportion of heads in a sequence of 10,000 coin tosses. @coin.

Initially the proportion of heads fluctuates greatly but begins to settle down as the number of tosses increases. After 10,000 tosses the relative frequency of Head is 5067/10,000=0.5067. We might suppose that if Kerrich were able to continue his experiment forever, the proportion of heads would tend to a limiting value which would be very near, if not exactly, $1/2$. This hypothetical limiting value is the probability of heads and is denoted by $P(H)$.

Looking at this slightly more formally

The coin-tossing example motivates the **relative frequency** or **frequentist** definition of the probability of an event; namely

> the relative frequency with which the event occurs in the long run;

or, in other words,

> the proportion of times that the event would occur in an infinite number of identical repeated experiments.

Suppose that we toss a coin n times. If the coin is fair, and is tossed fairly, then it is reasonable to suppose that

$$\text{the relative frequency of } H \;=\; \frac{\text{number of times } H \text{ occurs}}{n},$$

tends to $1/2$ as n gets larger. We say that the event H has probability $1/2$, or $P(H) = 1/2$.

More generally, consider some event E based on the outcomes of an experiment. Suppose that the experiment can, in principle, be repeated, under exactly the same conditions, forever. Let $n(E)$ denote the number of times that the event E would occur in n experiments.

We suppose that

$$\text{the relative frequency of } E \;=\; \frac{n(E)}{n} \;\longrightarrow\; P(E)\,, \text{ as } n \longrightarrow \infty.$$

So, the probability $P(E)$ of the event E is defined as the limiting value of $n(E)/n$ as $n \to \infty$. That is,

$$P(E) \;=\; \lim_{n\to\infty} \frac{n(E)}{n}, \tag{3.1}$$

supposing that this limit exists. (Note: I have written 'limit' and not 'lim' because this is not a limit in the usual mathematical sense.)

In order to satisfy ourselves that the probability of an event **exists**, we do not need to repeat an experiment an infinite number of times, or even be able to. All we need to do is **imagine** the experiment being performed repeatedly. An event with probability 0.75, say, would be expected to occur 75 times out of 100 in the long run.

An alternative approach, considered in STAT0003, makes a simple set of basic assumptions about probability, called the **axioms of probability**. Using these axioms it can be proved that the limiting relative frequency in equation (3.1) does exist and that it is equal to $P(E)$. In STAT0002 we will not consider these axioms formally. However, they are so basic and intuitive that you will find that we take them for granted.

An aside. There is another definition of probability, the **subjective** definition, which is the degree of belief someone has in the occurrence of an event, based on their knowledge and any evidence they have already seen. For example, you might reasonably believe that the probability that a coin comes up Heads when it is tossed is $1/2$ because the coin looks symmetrical. If you are **certain** that $P(H) = 1/2$ then no amount of evidence from actually tossing the coin will change your mind. However, if you just think that $P(H) = 1/2$ is more likely that other values of $P(H)$ then observing many more Hs than Ts in a long sequence of tosses may lead you to believe that $P(H) > 1/2$. We will not consider this definition of probability again in this course. However, it forms the basis of the **Bayesian** approach to Statistics. You may study this in a more advanced courses, for example, STAT0008 Statistical Inference.

Example: tossing a coin (continued)

We now look at the coin-tossing example in a slightly different way. If we toss a coin forever we generate an infinite **population** of **outcomes**: {H,H,T,H,...}, say. Think about choosing, or **sampling**, one of these outcomes at random from this population. The probability that this outcome is H is the proportion of Hs in the population. If we **assume** that the coin is fair then the infinite population contains 50% Hs and 50% Ts. In that case $P(H) = 1/2$, and $P(T) = 1/2$.

Example: Graduate Admissioins at Berkeley

Table 3.1 contains data relating to graduate admissions in 1973 at the University of California, Berkeley, USA in the six largest (in terms of numbers of admissions) departments of the university. These data are discussed in Bickel et al. (1975). We use the following notation: A for an accepted applicant, R for a rejected applicant. Table 3.2 summarises the notation used for the frequencies in Table 3.1. For example, $n(A)$ denotes the number of applicants who were accepted. We will look at this example in more detail later.

Table 3.1: Numbers of graduate applicants by outcome.

A	*R*	total
1755	2771	4526

Table 3.2: Notation for Table 3.1.

A	*R*	total
n(A)	*n(R)*	*n*

The population of interest now is the population of graduate applicants to the six largest departments of the Berkeley. If we choose an applicant at random from this population the probability $P(A)$ that they are accepted is given by

$$P(A) = \frac{n(A)}{n} = \frac{1,755}{4,526} = 0.388.$$

Similarly, the probability $P(R)$ that a randomly chosen applicant is rejected is given by

$$P(R) = \frac{n(R)}{n} = \frac{2,771}{4,526} = 0.612.$$

In both the coin-tossing and Berkeley admissions examples we have imagined choosing an individual from a population in such a way that all individuals are equally likely to be chosen. The probability of the individual having a particular property, for example, that an applicant is accepted, is given by the proportion of individuals in the population which have this property.

In the coin-tossing example the population is hypothetical, generated by thinking about repeating an experiment an infinite number of times. In the Berkeley admissions example the population is an actual population of people.

Notation: sample space, outcomes and events

The set of all possible outcomes of a random experiment is called the **sample space** S of the experiment. We may denote a single **outcome** of an experiment by s. An **event** E is a collection of outcomes, possible just one outcome.

It is very important to define the sample space S carefully. In the coin-tossing example we have $S = \{H, T\}$. If the coin is unbiased then the probabilities of the outcomes in S are given by $P(H) = 1/2$ and $P(T) = 1/2$.

3.3 Basic properties of probability

Since we have defined probability as a proportion, basic properties of proportions must also hold for a probability. Consider an event E. Then the following must hold

- $0 \leq P(E) \leq 1$;
- if E is impossible then $P(E) = 0$;
- if E is certain then $P(E) = 1$;
- $P(S) = 1$. This is true because the outcome must, by definition, be in the sample space.

3.4 Conditional probability

In Section 3.1 the statistical evidence presented to the court was based on the estimate that "the probability of one cot death **in a family like Sally Clark's** is 1 in 8543". What does this mean?

The study from which this statistic was taken estimated the **overall** probability of cot death to be 1 in 1303. That is, cot death occurs in approximately 1 in every 1303 families.

However, the study also found that the probability of cot death depended on various characteristics such as, income, smoking status and the age of the mother. For example, the probability of cot death was found to be much greater in families containing one or more smokers than in non-smoking families.

The study estimated the probability of cot death for each possible combination of these characteristics. For the combination which is relevant to Sally Clark, whose family was affluent, non-smoking and she was aged over 26, the probability of cot death was estimated to be smaller: 1 in 8543.

This (1 in 8543) is a **conditional** probability. We have **conditioned** on the event that the family in question is affluent, non-smoking and the mother is aged over 26. The overall probability of cot death (1 in 1303) is often called an **unconditional** probability.

In this example it is perhaps easiest to think of the conditioning as selecting a specific sub-population of families from the complete population of families. Another way to think about this is in terms of the sample space. We have reduced the original sample space - the outcomes (cot death or no cot death) of all families with children - to a subset of this sample space - the outcomes of all affluent, non-smoking families where the mother is over 26.

Notation

When we are working with conditional probabilities we need to use a neat notation rather than write out long sentences like the ones above.

Let C be the event that a family has one cot death. Let F_1 be the event that the family in question is affluent, non-smoking, and the mother is over 26. Instead of writing "the probability of one cot death in a family conditional on the type of family is 1 in 8543"" we write

$$P(C \mid F_1) = \frac{1}{8543}.$$

The '|' sign means **"conditional on"**' or, more simply, **"given"**. Therefore, for $P(C \mid F_1)$ we might say ""the probability of event C conditional on event F_1", or "the probability of event C given event F_1".

The (unconditional) probability of one cot death is given by

$$P(C) = \frac{1}{1303}.$$

In section 3.1 I did not use the | sign in my notation (because we hadn't seen it then), but I did make the conditioning clear by saying **for a family like Sally Clark's**.]

In fact **all** probabilities are conditional probabilities, because a probability is conditioned on the sample space S. When we define S we rule out anything that is not in S. So instead of $P(C)$ we could write $P(C \mid S)$. We do not tend to do this because it takes more time and it tends to make things more difficult to read. However, we should always try to bear in mind the sample space when we think about a probability.

We return to the Berkeley admissions example to illustrate conditional probability, independence and the rules of probability.

Example: Graduate Admissions at Berkeley (continued)

Table 3.3 contains more information on data relating to graduate admissions in 1973 at Berkeley in the six largest (in terms of numbers of admissions) departments of the university.

We use the following notation: M for a male applicant, F for a female applicant, A for an accepted applicant, R for a rejected applicant. This is an example of a 2-way contingency table (see Chapter 7).

Table 3.3: Numbers of graduate applicants by sex and outcome.

	A	*R*	total
M	1198	1493	2691
F	557	1278	1835
total	1755	2771	4526

Table 3.4 summarises the notation used for the frequencies in Table 3.3. For example, $n(M, A)$ denotes the number of applicants who were both male and accepted. Of course, $n(M, A) = n(A, M)$.

Table 3.4: Notation for Table 3.3.

	A	*R*	total
M	*n(M, A)*	*n(M, R)*	*n(M)*
F	*n(F, A)*	*n(F, R)*	*n(F)*
total	*n(A)*	*n(R)*	*n*

If we divide each of the numbers in Table 3.3 by the total number of applications ($n = 4526$) then we obtain the proportions of applicants in each of the four categories. These proportions are given (to 3 decimal places) in Table 3.5.

Table 3.6 summarises the notation used for the probabilities in Table 3.5.

For example, the

- probability $P(M)$ that a randomly chosen applicant is male is 0.595; and

Table 3.5: Proportions based on Table 3.3.

	A	*R*	total
M	0.265	0.330	0.595
F	0.123	0.282	0.405
total	0.388	0.612	1.000

Table 3.6: Notation for Table 3.5.

	A	*R*	total
M	*P(M, A)*	*P(M, R)*	*P(M)*
F	*P(F, A)*	*P(F, R)*	*P(F)*
total	*P(A)*	*P(R)*	1

- probability $P(M, A)$, or $P(M \text{ and } A)$, or $P(M \cap A)$, or even $P(MA)$, that a randomly chosen applicant is both male and accepted is 0.265.

In the second bullet point four different forms of notation are used to denote the same probability. In these notes I may (deliberately, of course) use more than one form of notation, to get you used to the fact that different texts may use different notation. Of course, $P(M, A) = P(A, M)$ and so on.

The following explains how these probabilities are calculated.

$$P(M) = \frac{n(M)}{n} = \frac{2,691}{4,526} = 0.595.$$

The sample space is $\{M, F\}$.

$$P(M, A) = \frac{n(M, A)}{n} = \frac{1,198}{4,526} = 0.265.$$

The sample space is $\{(M, A), (M, R), (F, A), (F, R)\}$.

Suppose that we wish to investigate whether there appears to be any sexual discrimination in the graduate admissions process at Berkeley in 1973. To do this we might compare

- the probability of acceptance for males, that is, $P(A \mid M)$; and
- the probability of acceptance for females, that is, $P(A \mid F)$.

These are **conditional** probabilities. Firstly, we calculate $P(A \mid M)$. Look at Table 3.7. We are considering only male applicants (M) so we have shaded the M row. Since we are conditioning on M, female applicants are not relevant. We are only concerned with the $n(M) = 2691$ male applicants in the M row.

Of these, $n(M, A) = 1198$ are accepted and $n(M, R) = 1493$ are rejected.

The information that the applicant is male (that is, event M has occurred) has reduced the sample space to $\{(M, A), (M, R)\}$, or, since we **know** that M has occurred, the sample space is effectively $\{A, R\}$. The conditional probability of A given M is the probability that event A occurs when we consider only those occasions on which event M occurs.

Table 3.7: Conditioning on applicant being male (frequencies).

	A	*R*	total
M	1198	1493	2691
F	557	1278	1835
total	1755	2771	4526

Therefore,

$$P(A \mid M) = \frac{n(A, M)}{n(M)} = \frac{1198}{2691} = 0.445, \tag{3.2}$$

that is, the proportion of male applicants who are accepted is 0.445.

Now look at Table 3.8.

Table 3.8: Conditioning on applicant being male (probabilities).

	A	*R*	total
M	0.265	0.330	0.595
F	0.123	0.282	0.405
total	0.388	0.612	1.000

An equivalent way to calculate $P(A \mid M)$ is

$$P(A \mid M) = \frac{P(A, M)}{P(M)} = \frac{1198/4526}{2691/4526} = 0.445. \tag{3.3}$$

Instead of using the frequencies in the shaded M row, we have used the probabilities. We get exactly the same answer because the probabilities are simply the frequencies divided by 4526.

Exercise. Show that $P(A \mid F) = 0.304$.

Exercise. Find $P(R \mid M)$ and $P(R \mid F)$. What do you notice about $P(A \mid M)$ and $P(R \mid M)$?

The calculation of $P(A \mid M)$ in equation (3.2) based on the frequencies in Table 3.7 should make sense to you. From the equivalent calculation in equation (3.3) we can see that the following definition of conditional probability makes sense.

Definition. If B is an event with $P(B) > 0$, then for each event A, the **conditional probability of** A **given** B is

$$P(A \mid B) = \frac{P(A, B)}{P(B)}. \tag{3.4}$$

Remarks:

- It is necessary that $P(B) > 0$ for $P(A \mid B)$ to be defined. It does not make sense to condition on an event that is impossible.
- Equation (3.4) implies that $P(A, B) = P(A \mid B)\, P(B)$.
- For any event A, $P(A \mid B)$ is a probability in which the sample space has been reduced to outcomes containing the event B. Therefore, **all** properties of probabilities also hold for conditional probabilities, e.g. $0 \leq P(A \mid B) \leq 1$.

Dependence and Independence

We return to the Berkeley example. We have found that

$$P(A \mid M) = 0.445 \qquad \text{and} \qquad P(A \mid F) = 0.304.$$

From Table 3.5 we find that

$$P(A) = 0.388.$$

Therefore, the probability of acceptance depends on the sex of the applicant:

$$P(A \mid M) > P(A) \qquad \text{and} \qquad P(A \mid F) < P(A).$$

The probability $P(A \mid M)$ of being accepted for a randomly selected male is greater than the unconditional probability $P(A)$. Therefore, the occurrence of event M has affected the probability that event A occurs. This means that the events A and M are **dependent**. Similarly, $P(A \mid F) < P(A)$ means that the events A and F are **dependent**.

We consider independent and dependent events in more detail in Section 3.7.

The fact that $P(A \mid M) > P(A \mid F)$ might suggest to us that there is gender bias in the admissions process. However, as we shall see in Section 7.2, this may be a very misleading conclusion to draw from these data. There is an innocent explanation of why $P(A \mid M) > P(A \mid F)$.

Exercise. Can you think what this innocent explanation it might be?

3.5 Addition rule of probability

We continue with the Berkeley admissions data. Suppose that we wish to calculate the probability that an applicant chosen at random from the 4526 applicants is either male (event M) or accepted (event A), or both male and accepted. We denote this probability $P(M \text{ or } A)$ or $P(M \cup A)$.

Look at Table 3.9. The cells that satisfy either M or A, or both, have been shaded grey. To calculate $P(M \text{ or } A)$ we simply sum the numbers of applicants for which either M or A, or both, is satisfied and then divide by $n = 4526$, that is,

$$\begin{aligned} P(M \text{ or } A) &= \frac{n(M \text{ or } A)}{n} \\ &= \frac{n(M,A) + n(M,R) + n(F,A)}{n} \\ &= \frac{1198 + 1493 + 557}{4526} = \frac{3248}{4526} = 0.718. \end{aligned}$$

Table 3.9: Cells with M or A are shaded (frequencies).

	A	*R*	total
M	1198	1493	2691
F	557	1278	1835
total	1755	2771	4526

An equivalent way to calculate $P(M \text{ or } A)$ is to sum probabilities in Table 3.10, that is,

$$P(M \text{ or } A) = P(M \text{ and } A) + P(M \text{ and } R) + P(F \text{ and } A) \quad (3.5)$$
$$= 0.265 + 0.330 + 0.123 = 0.718. \quad (3.6)$$

Table 3.10: Cells with M or A are shaded (probabilities).

	A	*R*	total
M	0.265	0.330	0.595
F	0.123	0.282	0.405
total	0.388	0.612	1.000

Exercise. Can you see why these two calculations are equivalent?

Exercise. Can you see a (slightly) quicker way to calculate $P(M \text{ or } A)$?

Now consider a slightly different way to show that $n(M \text{ or } A) = 3248$:

$$n(M \text{ or } A) = n(M) + n(A) - n(M, A) = 2691 + 1755 - 1198 = 3248.$$

Can you see from Table 3.9 why this works?

Similarly,

$$P(M \text{ or } A) = P(M) + P(A) - P(M \text{ and } A) = 0.595 + 0.388 - 0.265 = 0.718.$$

From this example we can see that following rule makes sense.

Definition. For any two events A and B

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \quad (3.7)$$

3.5.1 Mutually exclusive events

Two events A and B are **mutually exclusive** (or **disjoint**) if they cannot occur together. For example, in the Berkeley example the events A and R are mutually exclusive: it is not possible for an applicant to be both accepted and rejected.

If two events A and B are mutually exclusive then $P(A \text{ and } B) = 0$. Substituting this into equation (3.7) we find that, **if events A and B are mutually exclusive**

$$P(A \text{ or } B) = P(A) + P(B). \quad (3.8)$$

You can only use this equation in the **special case** where events A and B are mutually exclusive. Otherwise, you must use the general rule in equation (3.7).

The complement of an event A

The **complement** of an event A is the event that A does not occur. This can be denoted not$A$ or $\bar{A}$ or A^c. The events A and not$A$ are mutually exclusive and $S = \{A, \text{not}A\}$. We have already seen that $P(S) = 1$. Therefore,

$$P(S) = P(A \text{ or not}A) = P(A) + P(\text{not}A) = 1.$$

and therefore

$$P(\text{not}A) = 1 - P(A).$$

3.6 Multiplication rule of probability

We continue with the Berkeley admissions data. We have already calculated that $P(M, A) = 0.265$. Now we calculate this in a different way.

Think about the process of applying for a place at university. First an applicant makes an application. Then the university decides whether to accept or reject. To have the event (M, A) we first need a male applicant to apply and then for the university to accept them.

The calculation we performed above was

$$P(M, A) = \frac{n(M, A)}{n}.$$

Assuming that $n(M) > 0$, that is, $P(M) > 0$, we can rewrite this as

$$P(M, A) = \frac{n(M)}{n} \times \frac{n(M, A)}{n(M)} = \frac{n(M)}{n} \times \frac{n(A, M)}{n(M)},$$

or

$$P(M, A) = P(M) \times P(A \mid M).$$

Firstly, we calculate the proportion of applicants who are male, and then the proportion of those male applicants who are accepted. Multiplying these proportions gives the overall proportion of applicants who are both male and accepted.

This also follows directly on rearrangement of the definition,

$$P(A \mid M) = \frac{P(M, A)}{P(M)}.$$

Definition. Consider two events A and B with $P(B) > 0$. Rearranging the definition of conditional probability (3.4) gives

$$P(A, B) = P(B)\, P(A \mid B). \tag{3.9}$$

This can be generalised to the case of n events to give

$$P(A_1, A_2, \ldots, A_{n-1}, A_n) \;=\; P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1, A_2) \cdots P(A_n \mid A_{n-1}, \ldots, A_1) \tag{3.10}$$

provided that all the conditional probabilities are defined. A sufficient (but not necessary) condition for this is $P(A_1, A_2, \ldots, A_{n-1}, A_n) > 0$. For example, $P(A_1, A_2, A_3)$ is the probability that events A_1, A_2 and A_3 all occur.

3.7 Independence of events

Definition. Two events A and B are **independent** if

$$P(A, B) = P(A)\, P(B). \tag{3.11}$$

Otherwise A and B are dependent events.

Remarks:

1. If $P(A) > 0$ and $P(B) > 0$, then independence of A and B implies

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(A)\, P(B)}{P(B)} = P(A)$$

and similarly $P(B \mid A) = P(B)$.

2. The definition applies for events that have zero probability. For example, suppose that $P(A) > 0$ and $P(B) = 0$. Then A and B are independent because
$$P(A, B) = 0 = P(A)\,P(B).$$
3. The definition is symmetric in A and B. If A is independent of B, then B is independent of A.
4. Notation: the notation $A \perp\!\!\!\perp B$ can be used for "A and B are independent".
5. The definition of independence can be extended to more than two events. $A_1, A_2, \ldots, A_n$ are (mutually) independent if, for $r \in \{2, \ldots, n\}$, for any subset $\{C_1, \ldots, C_r\}$ of $\{A_1, \ldots, A_n\}$ we have
$$P(C_1, C_2, \cdots C_r) = P(C_1)P(C_2)\cdots P(C_r).$$
For example, if $n = 3$ then we need
$$P(A_1, A_2) = P(A_1)P(A_2),\ P(A_1, A_3) = P(A_1)P(A_3),\ P(A_2, A_3) = P(A_2)P(A_3)$$
that is, (A_1, A_2, A_3) are pairwise independent, and
$$P(A_1, A_2, A_3) = P(A_1)P(A_2)P(A_3).$$

3.7.1 An example of independence

The 2 most important systems for classifying human blood are the ABO system, with blood types A, B, AB and O, and the Rhesus system, with blood types Rh+ and Rh-. In the ABO system an A (and/or B) indicates the presence of antigen A (and/or B) molecules on the red blood cells. These two systems are often combined to form 8 blood types A+, B+, AB+, O+, A–, B–, AB– and O–. Knowledge of blood type is important when a patient needs a blood transfusion.Giving blood of the wrong type can cause harm: giving Rh+ blood to someone who is Rh– will make that person ill, as will giving blood with a A or B antigens to someone without those antigens. An AB+ person can receive blood from anyone, but an O– person can only receive O- blood.

The proportions of blood types varies between countries. In the UK the percentages for the ABO system are estimated to be equal to those in Table 3.11. Table 3.12 gives the percentages for the Rhesus system.

Table 3.11: Distribution of ABO blood groups in the UK.

ABO group	percentage
O	44
A	42
B	10
AB	4

Table 3.12: Distribution of Rhesus blood groups in the UK.

Rhesus group	percentage
Rh+	83
Rh-	17

These tables tell us that, for the UK, $P(\text{O}) = 0.44, P(\text{A}) = 0.42, P(\text{B}) = 0.1, P(\text{AB}) = 0.04$, $P(\text{Rh}+) = 0.83$ and $P(\text{Rh}-) = 0.17$.

Your blood type is genetically inherited from your parents. Since the genetic code responsible for inheritance of ABO blood group and Rhesus blood group are on different chromosomes, ABO and Rhesus blood types are inherited independently of each other. That is, for any person, the blood type in the ABO system is independent of their blood type in the Rhesus system.

Assuming that ABO blood type is independent of Rhesus blood type gives the probabilities in Table 3.13. Some of these probabilities have been omitted.

Table 3.13: Distribution of Rhesus blood groups in the UK.

	O	A	B	AB	total
Rh+		0.349	0.083		0.830
Rh-	0.075	0.071	0.017		0.170
total	0.440	0.420	0.100	0.040	1.000

Exercise. Using equation (3.11), or otherwise, calculate the values that are missing from Table 3.13.

3.8 Law of total probability

We continue with the Berkeley admissions data. We have already calculated that $P(A) = 0.388$. Now we calculate this in a different way.

From table 3.2 we can see that

$$n(A) = n(A, M) + n(A, F).$$

From table 3.6 we can see that

$$P(A) = P(A, M) + P(A, F). \tag{3.12}$$

This also follows from equation (3.8) since the events (A, M) and (A, F) are mutually exclusive. That is,

$$\begin{aligned} P(A) &= P((A, M) \text{ or } (A, F)), && (3.13)\\ &= P(A, M) + P(A, F) - P((A, M), (A, F)), && (3.14)\\ &= P(A, M) + P(A, F). && (3.15) \end{aligned}$$

Applying the multiplication rule (3.9) to (3.12) gives

$$\begin{aligned} P(A) &= P(A, M) + P(A\,F), && (3.16)\\ &= P(A \mid M)\,P(M) + P(A \mid F)\,P(F) && (3.17)\\ &= 0.445 \times 0.595 + 0.304 \times 0.405, && (3.18)\\ &= 0.388. && (3.19) \end{aligned}$$

This is an example of the **law of total probability**. The probability of event A is expressed as a weighted average of the conditional probability of event A given that M has occurred, and the conditional probability of event A given that F has occurred. Each conditional probability is given a weight equal to the probability of the event on which it is conditioned.

Note that

- M and F are mutually exclusive events.

- $P(M \text{ or } F) = 1$, that is, together M and F make up the entire sample space of the possible values of sex. This means that M and F are **exhaustive** events.

Definition. **The law of total probability**. Let events $B_1, \ldots, B_n$ be

- possible, i.e. $P(B_i) > 0$, for $i = 1, \ldots, n$,
- (pairwise) mutually exclusive, i.e. $P(B_i, B_j) = 0$ for $i \neq j$, and
- exhaustive, that is, $P(B_1 \text{ or } B_2 \text{ or } \cdots \text{ or } B_n) = 1$.

Then, for any event A

$$\begin{aligned} P(A) &= P(A \mid B_1)\, P(B_1) + \cdots + P(A \mid B_n)\, P(B_n), && (3.20) \\ &= \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i). && (3.21) \\ & && (3.22) \end{aligned}$$

Some books refer to the law of total probability as the **partition theorem**. This is because events $B_1, \ldots, B_n$ that are both mutually exclusive and exhasutive are said to **partition** the sample space, that is, they split the sample space into n disjoint parts.

3.9 Bayes' theorem

We continue with the Berkeley admissions data. We have already calculated that $P(A \mid M) = 0.445$. Now we calculate $P(M \mid A)$. It is important that you understand the difference between these two probabilities.

$P(A \mid M)$ is the probability of acceptance given that the applicant is male; in other words, the probability that a male applicant is accepted. We calculated that $P(A \mid M) = 0.445$ using the M row of Table 3.7 (or Table 3.8), that is, by conditioning on the event M.

$P(M \ldots A)$ is the probability that the applicant is male given that they are accepted; in other words, the probability that an accepted applicant is male.

Exercise. Use Table 3.14 or Table 3.15 to show that $P(M \mid A) = 0.683$.

Table 3.14: Conditioning on applicant being accepted (frequencies).

	A	*R*	total
M	1198	1493	2691
F	557	1278	1835
total	1755	2771	4526

Now we calculate $P(M \mid A)$ in a different way. If you used table 3.15 to calculate $P(M \mid A)$ you used the equation

$$P(M \mid A) = \frac{P(M, A)}{P(A)}. \quad (3.23)$$

The multiplication rule in Section 3.6 gives

$$P(M, A) = P(A \mid M)\, P(M). \quad (3.24)$$

Table 3.15: Conditioning on applicant being accepted (probabilites).

	A	*R*	total
M	0.265	0.330	0.595
F	0.123	0.282	0.405
total	0.388	0.612	1.000

Substituting (3.23) into (3.24) gives

$$P(M \mid A) = \frac{P(A \mid M)\, P(M)}{P(A)}. \tag{3.25}$$

Equation (3.25) is an example of Bayes' theorem, which was first derived in a paper presented to the Royal Society in 1763 by Richard Price on behalf of the late Reverend Thomas Bayes.

In this example we can either calculate $P(A)$ using the law of total probability:

$$P(A) = P(A \mid M)\, P(M) + P(A \mid F)\, P(F),$$

or calculate $P(A)$ directly from the table.

Bayes' theorem. Let $B_1, \ldots, B_n$ be mutually exclusive, exhaustive events, with $P(B_i) > 0$ for all i. Let A be an event with $P(A) > 0$. Then

$$\begin{aligned} P(B_i \mid A) &= \frac{P(A \mid B_i)\, P(B_i)}{P(A)}, && (3.26)\\ &= \frac{P(A \mid B_i)\, P(B_i)}{P(A \mid B_1)\, P(B_1) + \cdots + P(A \mid B_n)\, P(B_n)}, && (3.27)\\ &= \frac{P(A \mid B_i)\, P(B_i)}{\sum_{i=1}^{n} P(A \mid B_i)\, P(B_i)}. && (3.28) \end{aligned}$$

The proof of Bayes' theorem is a straightforward extension of the case with $n = 2$ considered in the Berkeley admissions example above.

Conditioning on more than one event

$P(A \mid B)$ is the conditional probability that event A occurs given that event B has occurred. We can extend this idea to condition on more than one event.

For example, $P(A \mid B, C)$, or $P(A \mid B \text{ and } C)$, or $P(A \mid B \cap C)$ is the conditional probability that event A occurs given that **both** events B and C have occurred. The general principle is that we have conditioned on all events that are placed on the right hand side of the conditional | symbol. All the results that we have seen can be extended to probabilities conditioned on more than one event.

For example, for $P(B, C) > 0$,

$$P(A \mid B, C) = \frac{P(A, B \mid C)}{P(B \mid C)} \qquad \text{(definition of conditional probability)},$$

and if, in addition, $P(A, C) > 0$

$$P(A \mid B, C) = \frac{P(B \mid A, C)\, P(A \mid C)}{P(B \mid C)} \qquad \text{(Bayes' theorem)}.$$

In each of these equations, if you ignore the event C then you will see familiar equations. The general idea is that definition of conditional probability and Bayes' theorem continue to hold if we condition all probabilities on the event C, provided that all the conditional probabilities involved are valid.

Alternatively, noting that we could reverse the roles of B and C,

$$P(A \mid C, B) = \frac{P(A, C \mid B)}{P(C \mid B)} \qquad \text{(definition of conditional probability)},$$

$$P(A \mid C, B) = \frac{P(C \mid A, B)\, P(A \mid B)}{P(C \mid B)} \qquad \text{(Bayes' theorem)}.$$

These four expressions give different ways to express the probability $P(A \mid C, B)$. You will be able to find other ways to express this probability. For example,

$$P(A \mid B, C) = \frac{P(A, B, C)}{P(B, C)}.$$

Exercise. Why this true?

Misleading statistical evidence in cot death trials (continued)

We will return to this example and use Bayes' theorem to calculate the probability that Sally Clark was innocent given (only) the statistical evidence presented at her trial. We will make some assumptions that we know are unrealistic, but the general approach that we take will illustrate the importance of using sound probabilistic reasoning.

3.10 DNA identification evidence

DNA evidence is increasingly being used to catch and prosecute suspects of crimes. The following example is based on a real criminal case.

In 1996 Denis John Adams was put on trial for rape. Apart from the fact that he lived in the area local to where the crime was committed, the only evidence against him was that his DNA matched a sample of DNA obtained from the victim. In fact, all other evidence was in favour of Adams. The victim did not pick him out an identity parade; the victim said he did not look like her attacker, who she said was in his early 20s (Adams was 37); Adams had an alibi.

At Adam's trial the Prosecution said that the **match probablity**, the probability that Adam's DNA would match the DNA evidence if he was innocent, is 1 in 200 million. The Defence disagreed with this, saying that 1 in 20 million or even 1 in 2 million was correct.

At the trial it was stated that there were approximately 150,000 males in the local area between 18 and 60 years old who, before any evidence was collected, could have been suspects.

Questions

- Do you think the evidence against Adams is very strong?
- If you were on the jury would you have voted 'guilty'?
- Would you want to do any calculations first? If so, what would you calculate?

Chapter 4

Random variables

Example. We return to the space shuttle example.

Consider what happens to the O-rings on a particular test flight, at a particular temperature. A given O-ring either is damaged (shows signs of thermal distress) or it is not damaged. Let D denote the event that an O-ring is damaged and $\bar{D}$ the event that it is not damaged. If we consider all 6 O-rings, there are many possible outcomes in the sample space, $2^6 = 64$, in fact:

$$S = \{DDDDDD\}, \{DDDDD\bar{D}\}, \ldots, \{D\bar{D}\bar{D}\bar{D}\bar{D}\bar{D}\}, \{\bar{D}\bar{D}\bar{D}\bar{D}\bar{D}\bar{D}\}.$$

Suppose that we are not interested in which particular O-rings were damaged, just the total number N of damaged O-rings. The possible values for N are 0,1,2,3,4,5,6.

Each outcome in S gives a value for N in $\{0,1,2,3,4,5,6\}$:

$\{DDDDDD\}$ gives $N = 6$,

$\{DDDDD\bar{D}\}$ gives $N = 5$,

$\{DDDD\bar{D}D\}$ gives $N = 5$,

$\vdots$

$\{\bar{D}\bar{D}\bar{D}\bar{D}\bar{D}\bar{D}\}$ gives $N = 0$.

By defining N to be the total number of damaged O-rings, we have moved from considering outcomes to considering a variable with a numerical value. N is a real-valued function on the sample space S, that is, N maps each outcome in S to a real number. N is a rule that assigns a real number to every outcome s in S. Since the outcomes in S are random the variable N is also random, and we can assign probabilities to its possible values, that is, $P(N = 0), P(N = 1)$ and so on.

N is a **random variable**. In fact, if we assume that O-rings are damaged independently of each other and each O-ring has the same probability p of being damaged, N is a random variable with a special name. It is a binomial random variable with parameters 6 and p. We will consider binomial random variables in more detail in Section 5.2.

Notation. We denote random variables by upper case letters, for example, N, X, Y, Z. Once we have observed the value of a random variable it is no longer random: it is equal to a particular value. To make this clear we denote sample values of r.v.s. by lower case letters, for example, n, x, y, z and write $N = n, X = x$ and so on. Thus, $P(X = x)$ is the probability that the random variable X has the value x.

4.1 Discrete random variables

Definition. A discrete random variable is a random variable that can take only a finite, or countably infinite, number of values.

An example of a countably infinite set of values is {0,1,2,3,...}. The random variable N in the space shuttle example takes a finite number of values: 0,1,2,3,4,5,6. Therefore N is a discrete random variable.

Definition. Let X be a discrete random variable. The **probability mass function (p.m.f.)** $p_X(x)$, or simply $p(x)$, of X is

$$p_X(x) = P(X = x), \qquad \text{for } x \text{ in the support of } X.$$

The p.m.f. of X tells us the probability with which X takes any particular value x. The **support** of X is the set of values that it is possible for X to take. It is very important to write this down every time you write down a p.m.f.. A discrete random variable is completely specified by its probability mass function.

Properties of p.m.f.s

Let X take values $x_1, x_2, \ldots$. Then

1. $p_X(x_i) \geq 0$, for all i,
2. $\sum_i p_X(x_i) = 1$.

Note: 1. is true because the $p_X(x_i)$s are probabilities; 2. is true because summing over the x_is is equivalent to summing over the sample space of outcomes.

Definition. The cumulative distribution function (c.d.f.) of a random variable X is

$$F_X(x) = P(X \leq x), \qquad \text{for } -\infty < x < \infty.$$

Relationship between the c.d.f. and p.m.f. of a discrete random variable. For a discrete random variable:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i).$$

Therefore, assuming for the moment that the random variable takes only integer values,

$$P(X = x) = P(X \leq x) - P(X \leq x - 1) = F_X(x) - F_X(x - 1)$$

for any integer x

4.2 Continuous random variables

Example. We return to the Oxford birth times example.

The top plot in Figure 4.1 shows a histogram of the 95 birth times. The variable of interest in this example is a time. Time is a continuous variable: in principle, the times in this dataset could take any positive real value, uncountably many values. In practice, these times have been recorded discretely, in units of 1/10 of an hour or 1/4 of an hour.

Suppose that we continue to collect data on birth duration from this hospital, and, as new observations arrive, we add them to the top histogram in Figure 4.1. We imagine that the times are recorded continuously. As the number of observations n increases we decrease the bin width of the histogram. As n increases to infinity the bin width shrinks to zero and the histogram tends to a smooth continuous curve.

This is shown in the bottom 3 plots in Figure 4.1. The extra data are not real. They are data I have simulated, using a computer, to have a distribution with a similar shape to the histogram of the real data.

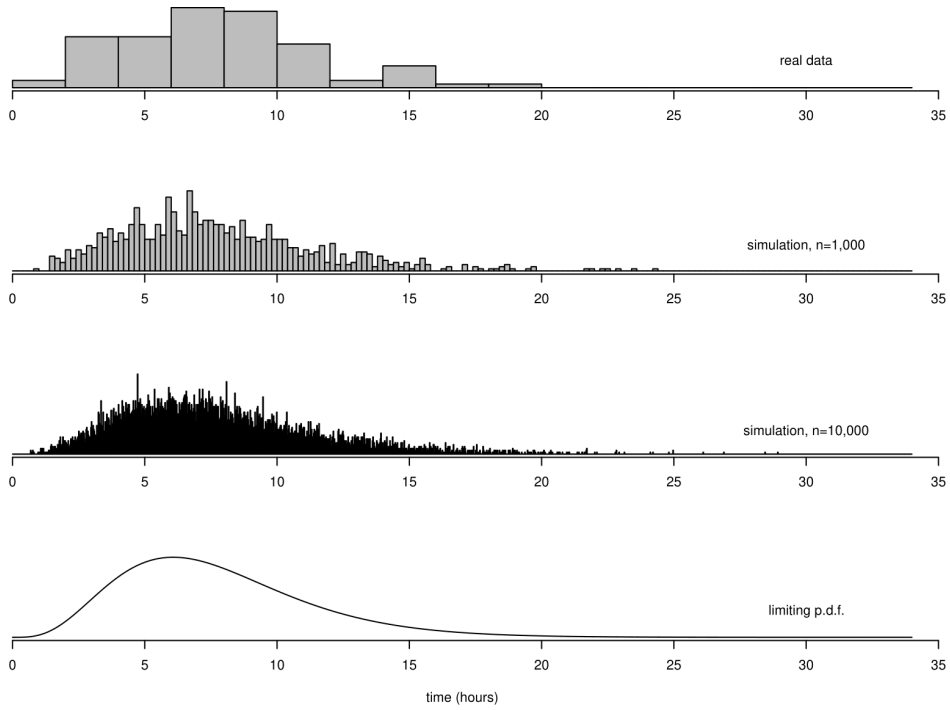

Figure 4.1: Top: histogram of the Oxford birth durations. Second from top: histogram of 1,000 values simulated from a distribution fitted to the data. Second from bottom: similarly for 10,000 simulated values. Bottom: p.d.f. of the distribution fitted to the Oxford birth times data.

Let T denote the time, in hours, that a woman arriving at the hospital takes to give birth. The smooth continuous curve at the bottom of Figure 4.1 is called the **probability density function (p.d.f.)** $f_T(t)$ of the random variable T. Since the total area of the rectangles in a histogram is equal to 1, the area $\int_{-\infty}^{\infty} f_T(t)\,\mathrm{d}t$ under the p.d.f. $f_T(t)$ is equal to 1.

Definition. A **probability density function (p.d.f.)** is a function $f_X(x)$, or simply $f(x)$, such that

1. $f_X(x) \geq 0$, for $-\infty < x < \infty$;
2. $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1.$

Therefore, p.d.f.s are always non-negative and integrate to 1. The support of a continuous random variable is the set of values for which the p.d.f. is positive. Suppose that we wish to find $P(4 < T \leq 12)$. To find the proportion of times between 4 and 12 using a histogram, we sum the areas of all bins between 4 and 12, that is, we find the area shaded in the histogram in Figure 4.2. To do this using the p.d.f. we do effectively the same thing: we find the area under the p.d.f. $f_T(t)$ between 4 and 12. Since $f_T(t)$ is a smooth continuous curve, (that is, the bin widths are zero) we integrate $f_T(t)$ between 4 and 12.

Therefore

$$P(4 < T \leq 12) = \int_4^{12} f_T(t)\,\mathrm{d}t = F_T(12) - F_T(4).$$

More generally,

$$P(a < T \leq b) = \int_a^b f_T(t)\,\mathrm{d}t = F_T(b) - F_T(a).$$

Definition. A random variable X is a **continuous random variable** if there exists a p.d.f. $f_X(x)$ such that

$$P(a < X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x,$$

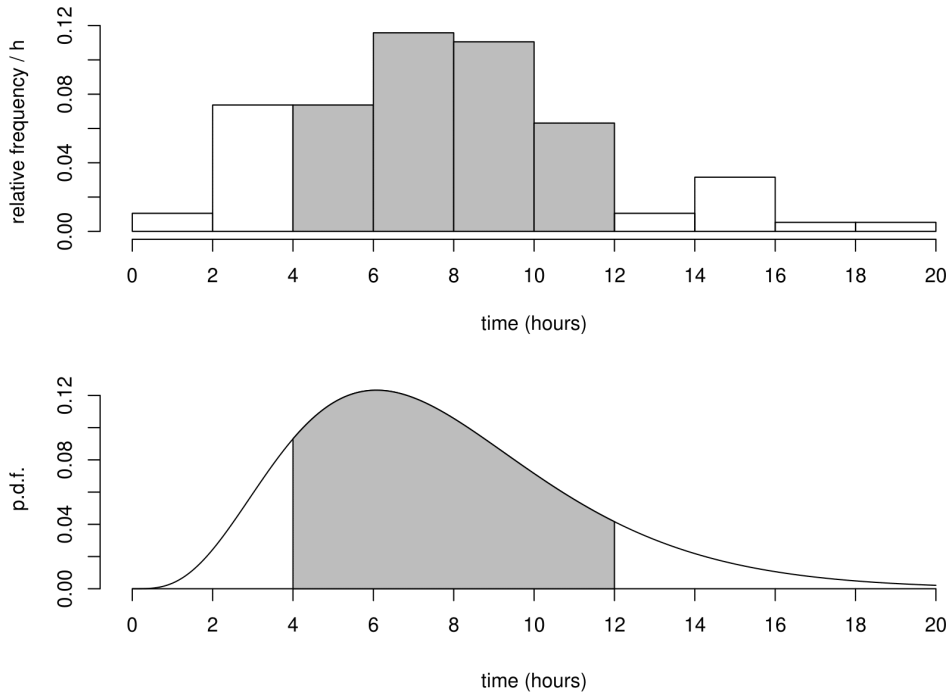

Figure 4.2: Top: histogram of the Oxford birth durations. Bottom: p.d.f. of the distribution fitted to the Oxford birth duration data.

for all a and b such that $a < b$.

Figure 4.3 illustrates the properties of a p.d.f..

Notes

- It is very important to appreciate that $f_X(x)$ is **not** a probability: it does **not** give $P(X = x)$. In fact $P(X = x) = 0$: the probability that a continuous random variable X takes the value x is zero.
- Indeed, it is possible for a p.d.f. to be greater than 1. Consider a continuous random variable X with p.d.f.

$$f_X(x) = \begin{cases} 2\,(1-x) & 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

  For this random variable $f_X(x) > 1$ for any $x \in [0, 1/2)$.
- Since $P(X = x) = 0$

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b).$$

- $f_X(x)$ is a probability **density**. The probability that X lies in a very small interval of length δ near x is approximately $f_X(x)\delta$. For the p.d.f. at the bottom of figure 4.1, $f_T(6) > f_T(12)$, indicating that a randomly chosen woman is more likely to spend approximately 6 hours giving birth than approximately 12 hours.

Relationship between the c.d.f. and p.d.f. of a continuous random variable. For a continuous random variable

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u.$$

Therefore,

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x).$$

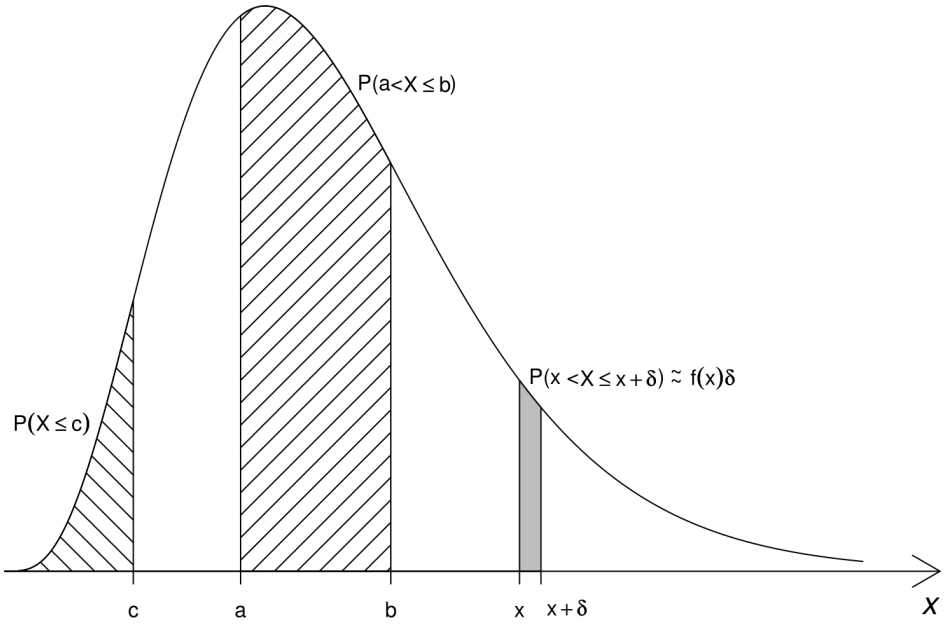

Figure 4.3: Properties of a p.d.f.. The areas that correspond to the probability that a random variable takes a value in a given interval are shaded.

4.3 Expectation

The expectation of a random variable is a measure of the location of its distribution.

4.3.1 Expectation of a discrete random variable

Example. We return to the space shuttle example.

Again we consider test flights conducted at a particular temperature, say 53°F. Suppose that NASA are able to conduct a very large number n of test flights at 53°F, producing a sample $x_1, \ldots, x_n$ of numbers of damaged O-rings.

Let $n(x)$ be the number of test flights on which x of the 6 O-rings were damaged. We can write the sample mean $\bar{x}$ of $x_1, \ldots, x_n$ as

$$\begin{aligned} \bar{x} &= \frac{0 \times n(0) + 1 \times n(1) + \cdots + 6 \times n(6)}{n}, \\ &= \sum_{x=0}^{6} x \, \frac{n(x)}{n}. \end{aligned}$$

As the sample size n increases to infinity, the sample proportion $n(x)/n$ tends to $P(X = x)$, for $x = 0, 1, \ldots, 6$. Therefore, in the limit as $n \to \infty$, $\bar{x}$ tends to

$$\sum_{x=0}^{6} x \, P(X = x). \tag{4.1}$$

This is known as the mean of the probability distribution of X. It is a measure of the location of the distribution.

The quantity in equation (4.1) is the value of the sample mean $\bar{x}$ that we would expect to get from a very large sample. Therefore it is often called the **expectation** or **expected value** of the random variable X and it is denoted $\mathrm{E}(X)$.

Definition. The **expectation** (or **expected value** or **mean**) $\mathrm{E}(X)$ of a discrete random variable X is given by

$$\mathrm{E}(X) = \sum_x x\, P(X = x). \tag{4.2}$$

This is a weighted average of the values that X can take, each value being weighted by $P(X = x)$.

Note:

- We often write μ or μ_X for $\mathrm{E}(X)$.
- Units: the units of $\mathrm{E}(X)$ are the same as those of X. For example, if X is measured in hours then $\mathrm{E}(X)$ is measured in hours.
- $\mathrm{E}(X)$ exists only if $\sum_x |x|\, P(X = x) < \infty$. If the number of values X can take is finite then $\mathrm{E}(X)$ will always exist.

4.3.2 Expectation of a continuous random variable

We can define the expectation of a continuous random variable in a similar way to a discrete random variable, replacing summation with integration.

Definition. The expectation $\mathrm{E}(X)$ of a continuous random variable X is given by

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x. \tag{4.3}$$

Note:

- Like the discrete case, this is a weighted average of the values that X can take, but now each value is weighted by the p.d.f. $f_X(x)$.
- The range of integration in equation (4.3) is over the whole real line but, in practice, integration will be over the range of possible values of X.
- $\mathrm{E}(X)$ exists only if $\int_{-\infty}^{\infty} |x|\, f_X(x)\, \mathrm{d}x < \infty$.

4.3.3 Properties of $\mathrm{E}(X)$

If a and b are constants then

$$\mathrm{E}(a\,X + b) = a\,\mathrm{E}(X) + b.$$

This makes sense. If we multiply all observations by a their mean will also be multiplied by a. If we add b to all observations their mean will be increased by b, that is, the distribution of X shifts up by b.

- If $X \geq 0$ then $\mathrm{E}(X) \geq 0$.
- If X is a constant c, that is, $P(X = c) = 1$ then $\mathrm{E}(X) = c$.
- It can be shown that

$$\mathrm{E}(X_1 + X_2 + \cdots + X_n) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n).$$

4.3.4 The expectation of $g(X)$

Suppose that $Y = g(X)$ is a function of of X, such as $aX + b$, X^2 or $\log X$. Then Y is also a random variable. If we find the p.m.f (if Y is discrete) or p.d.f. (if Y is continuous) of Y then we can find the the expectation of Y using equation (4.2) or (4.3) as appropriate.

$$\mathrm{E}(Y) = \mathrm{E}[g(X)] = \begin{cases} \sum_x g(x)\, P(X = x) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)\, f_X(x)\, \mathrm{d}x & \text{if } X \text{ is continuous.} \end{cases} \tag{4.4}$$

Note, it is usually the case that

$$\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$$

although there are exceptions.

4.4 Variance

The variance of a random variable is a measure of the spread of its distribution.

4.4.1 Variance of a discrete random variable

Example. We return the space shuttle example.

As before we let $n(x)$ be the number of test flights on which x of the 6 O-rings were damaged. We saw in Section 2.3.2 that a measure of the spread of a sample $x_1, \ldots, x_n$ is the sample variance s_X^2 which, in this example, can be written as

$$\begin{aligned} s_X^2 &= \frac{1}{n-1}\left\{(0-\bar{x})^2\, n(0) + (1-\bar{x})^2\, n(1) + \cdots + (6-\bar{x})^2\, n(6)\right\}, \\ &= \sum_{x=0}^{6} (x-\bar{x})^2\, \frac{n(x)}{n-1}. \end{aligned}$$

As the sample size n increases to infinity, $\frac{n(x)}{n-1}$ tends to $P(X = x)$, for $x = 0, 1, \ldots, 6$ and $\bar{x}$ tends to μ=$\mathrm{E}(X)$.

Therefore, as $n \to \infty$, s_X^2 tends to

$$\sum_{x=0}^{6} (x-\mu)^2\, P(X = x). \tag{4.5}$$

This is known as the variance of the probability distribution of X. It is a measure of the spread of the distribution. The quantity in equation (4.5) is the value of the sample variance s_X^2 that we would expect to get from a very large sample.

Definition. The variance $\mathrm{var}(X)$ of a discrete random variable X with mean $\mathrm{E}(X) = \mu$ is given by

$$\mathrm{var}(X) = \sum_x (x-\mu)^2\, P(X = x). \tag{4.6}$$

This is a weighted average of the squared differences between the values that X can take and its mean μ, each value being weighted by $P(X = x)$.

Note:

- $\mathrm{var}(X)$ exists only if μ exists and $\sum_x (x-\mu)^2\, P(X = x) < \infty$. If the number of values X can take is finite then $\mathrm{var}(X)$ will always exist.
- We often write σ^2 or σ_X^2 for $\mathrm{var}(X)$.

4.4.2 Variance of a continuous random variable

We can define the variance of a continuous random variable in a similar way to a discrete random variable, replacing summation with integration.

Definition. The variance $\mathrm{var}(X)$ of a continuous random variable X with mean $\mathrm{E}(X) = \mu$ is given by

$$\mathrm{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,\mathrm{d}x. \tag{4.7}$$

Note: $\mathrm{var}(X)$ exists only if μ exists and $\int_{-\infty}^{\infty}(x-\mu)^2\, f_X(x)\,\mathrm{d}x < \infty$.

4.4.3 Variance and standard deviation

Definition. Let X be a random variable with $\mathrm{E}(X) = \mu$. The variance $\mathrm{var}(X)$ is given by

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

This follows from equations (4.6) and (4.7) and the expression in equation (4.4) for the expectation of a function $g(X)$ of a random variable X.

There is an alternative way to calculate $\mathrm{var}(X)$:

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

Exercise. Prove this.

Definition. The standard deviation $\mathrm{sd}(X)$ of X is given by $\mathrm{sd}(X)=+\sqrt{\mathrm{var}(X)}$.

Notes on $\mathrm{var}(X)$:

- $\mathrm{var}(X) \geq 0$. A variance cannot be negative.
- Units: the units of $\mathrm{var}(X)$ are the square of those of X. For example, if X is measured in hours then $\mathrm{var}(X)$ is measured in hours^2 (and $\mathrm{sd}(X)$ is measured in hours).

4.4.4 Properties of $\mathrm{var}(X)$

- If a and b are constants then
$$\mathrm{var}(a\,X + b) = a^2\,\mathrm{var}(X).$$
This makes sense. If we multiply all observations by a their variance, which is measured square units, will be multiplied by a^2. If we add b to all observations their variance will be unchanged because the distribution simply shifts up by b and its spread is unaffected.
- If X is a constant c, that is, $P(X = c) = 1$ then $\mathrm{var}(X) = 0$: the distribution of X has zero spread.
- It can also be shown that **if the random variables $X_1, X_2, \ldots X_n$ are independent** then

$$\mathrm{var}(X_1 + X_2 + \cdots + X_n) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \cdots + \mathrm{var}(X_n). \tag{4.8}$$

Note. Independence is sufficient for this result to hold but it is not necessary. Taking $n = 2$ as an example, in generality we have

$$\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2),$$

where $\text{cov}(X_1, X_2)$ is the **covariance** between the random variables X_1 and X_2. Covariance is a measure of the strength of **linear** association. If X_1 and X_2 are independent (have no association of any kind) then $\text{cov}(X_1, X_2) = 0$, because they have no linear association. However, it is possible for X_1 and X_2 to be dependent but $\text{cov}(X_1, X_2) = 0$, because, although they have some kind of association, they have no **linear** association. Thus, independence is a stronger requirement than zero covariance.

Returing to general n we have

$$\text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n) + 2\sum_{i<j}\sum \text{cov}(X_i, X_j).$$

If $\text{cov}(X_i, X_j) = 0$ for all $i < j$ then equation (4.8) holds. We will study covariance, and its standardised form **correlation**, in Chapter 9.

4.5 Other measures of location

4.5.1 The median of a random variable

Recall that the sample median of a set of observations is the middle observation when the observations are arranged in order of size. We define the median of a random variable X as the value, median(X), such that

$$P(X < \text{median}(X)) \leq \frac{1}{2} \leq P(X \leq \text{median}(X)).$$

In other words, $\text{median}(X)$ is the value where a plot of the c.d.f. $F_X(x) = P(X \leq x)$ crosses $1/2$.

For a continuous random variable X we have

$$F_X(\text{median}(X)) = P(X \leq \text{median}(X)) = \frac{1}{2}.$$

and the median will divide the distribution into two parts, each with probability $1/2$:

$$P(X < \text{median}(X)) = P(X > \text{median}(X)) = \frac{1}{2}.$$

This will not necessarily hold for a discrete distribution. For example, suppose that

$$P(X = 0) = \frac{1}{6}, \qquad P(X = 1) = \frac{1}{2}, \qquad P(X = 2) = \frac{1}{3}.$$

Then

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{6} & \text{for } 0 \leq x < 1, \\ \frac{2}{3} & \text{for } 1 \leq x < 2, \\ 1 & \text{for } x \geq 2, \end{cases}$$

Therefore, $\text{median}(X) = 1$. However, $P(X < 1) = \frac{1}{6}$ and $P(X > 1) = \frac{1}{3}$.

4.5.2 The mode of a random variable

Recall that the sample mode of categorical or discrete data is the value (or values) which occurs most often. We define the mode, mode(X), of a random variable as follows.

For a discrete random variable X, the mode is the value which has the highest probability of occurring: $P(X = \text{mode}(X))$ will be larger than for any other value X can have. In other words, $\text{mode}(X)$ is the value at which the p.m.f. is maximised.

For a continuous random variable X, the mode is the value at which the p.d.f. is maximised. **If the maximum occurs at a turning point of** $f_X(x)$ then it can be found by solving the equation

$$\frac{\mathrm{d}}{\mathrm{d}x} f_X(x) = 0,$$

and checking that you have indeed found a maximum.

4.6 Quantiles

To keep things simple we consider a **continuous** random variable X. The $100p\%$ quantile of X is defined to be the value x_p such that

$$F_X(x_p) = P(X \leq x_p) = p.$$

Thus, $x_{1/4}$ is the lower quartile of X, $x_{1/2}$ is the median of X and $x_{3/4}$ is the upper quartile of X. The inter-quartile range is $x_{3/4} - x_{1/4}$, which is a measure of spread.

4.7 Measures of shape

The **moment coefficient of skewness** of a random variable X with mean μ and standard deviation σ is given by

$$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\sigma^3},$$

provided that $\mathrm{E}[(X-\mu)^3]$ exists.

The **quartile skewness** of a random variable X with c.d.f $F_X(x)$ is given by

$$\frac{[F_X(3/4) - F_X(1/2)] - [F_X(1/2) - F_X(1/4)]}{F_X(3/4) - F_X(1/4)}.$$

Chapter 5

Simple distributions

In this section we use a dataset to introduce some commonly-used simple distributions. We will study the **discrete** distributions: Bernoulli, binomial, geometric and Poisson. We will also study the **continuous** distributions: uniform, exponential, normal.

5.1 The Bernoulli distribution

5.2 The binomial distribution

5.3 The geometric distribution

5.4 The Poisson distribution

5.5 The uniform distribution

5.6 The exponential distribution

5.7 The normal distribution

5.8 QQ plots

Chapter 6

Statistical Inference

Chapter 7

Contingency tables

7.1 2-way contingency tables

7.1.1 Comparing probabilities

Note that in Ross, S. (2010) A First Course in Probability the odds ratio of an event A is defined (incorrectly) as $P(A)/(1-P(A))$. This is a ratio of probabilities. The conventional use of the term **odds ratio** is for a ratio of odds.

7.2 3-way contingency tables

Chapter 8

Linear regression

Chapter 9

Correlation

9.1 Correlation: a measure of linear association

9.2 Covariance and correlation

9.3 Use and misuse of correlation

Chapter 10

A general strategy for statistical modelling

Bibliography

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. Science, 187(4175):398–404.

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-challenger prediction of failure. Journal of the American Statistical Association, 84(408):945–957.

Simpson, J., Olsen, A., and Eden, J. C. (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. Technometrics, 17(2):161–166.

Smith, R. L. (2002). A statistical assessment of Buchanan's vote in palm beach county. Statist. Sci., 17(4):441–457.