# MT1007: Statistical Theory

## Contents

1. Statistical theory .................................................. 2

2. Probability theory .................................................. 2
   2.1 A formal definition of probability .......................... 3
   2.2 Outcome counting .............................................. 4
      2.2.1 Example of outcome counting: the birthday problem 6
   2.3 Probabilities of compound events .......................... 7
      2.3.1 \( \Pr(A \text{ and } B) \) .................................. 7
      2.3.2 \( \Pr(A \text{ and/or } B) \) .................................. 8
      2.3.3 Complementary events .................................... 9
   2.4 Conditional Probability ....................................... 9
      2.4.1 Total probability rule .................................. 10
      2.4.2 Bayes' Rule .............................................. 10
      2.4.3 Independence ............................................ 11
   2.5 Generalizations ................................................. 11

3. Random Variables (r.v.'s) ........................................ 12
   3.1 Discrete random variables .................................... 13
      3.1.1 Expectation and variance of r.v.s ..................... 14
      3.1.2 Some useful distributions for modelling discrete r.v.s 16
      3.1.3 The Poisson distribution ............................... 18
   3.2 Dealing with several r.v.s: joint distributions ....... 19
      3.2.1 Independence of r.v.s .................................. 20
   3.3 \( \Pr[X \leq x] \): the cumulative distribution function (c.d.f.) 20
   3.4 Continuous random variables ................................ 21
      3.4.1 \( \Pr[X = x] = 0 \) if \( X \) is a continuous r.v. ...... 22
      3.4.2 Probability density functions (p.d.f.) ............... 22
      3.4.3 Joint distributions, independence and expectation .... 23
      3.4.4 Seven useful facts about expectations and variances 25
   3.5 The normal distribution ....................................... 28
   3.6 c.d.f.s, p.d.f.s and p.m.f.s in \( \mathbb{R} \) ..................... 29

4. Estimation and Hypothesis testing ............................... 30
   4.1 Point Estimation ............................................... 31
   4.2 The Method of Moments (MOM) ................................ 31
   4.3 Least Squares .................................................. 33
   4.4 Maximum Likelihood estimation ............................. 38
   4.5 Bias and variance in estimation ............................ 39
      4.5.1 Bias .................................................. 39
      4.5.2 Estimator variance ..................................... 40
   4.6 Confidence Intervals .......................................... 41
      4.6.1 How to obtain CI's in practice: an example ........... 42
      4.6.2 Calculating CI's in general: the ideal recipe ......... 43
      4.6.3 Confidence intervals when the variance is unknown .... 43
      4.6.4 Discussion and approximations .......................... 46
   4.7 Hypothesis testing ............................................. 46
      4.7.1 \( p \)-values: measuring the evidence against a hypothesis 46
      4.7.2 Specific example: testing the mean of a normal population 47
      4.7.3 More generality: what to do when variance is unknown 48

---

To view the document, please refer to the above table of contents. The page number indicates where each section begins.
1 Statistical theory

In the climate change case study the concepts of probability, random variables, statistical modelling, estimation and hypothesis testing were introduced in a somewhat intuitive manner, but to be general enough to be widely applicable the theory for statistical modelling needs to be put on a firmer mathematical basis.

The purpose of this segment of the course is to present, compactly, the basic mathematical theory needed for a proper understanding of statistical modelling and an appreciation of what statistics is about. This core of ideas will be presented in a moderately abstract mathematical manner. The reason for doing this is to provide a theory which is general enough to cover a very wide range of practical statistical investigations. Sometimes this may mean that it looks like things are being made more complicated than they need to be - usually they are not, the aim is to produce the simplest theory that will be capable of covering all the situations in which we might want to use it. In this section of the course the examples will tend to be very simple, and perhaps a little boring: the reason for this is that they are there to illustrate the theory as simply and clearly as possible. If you get a thorough familiarity with the basic concepts of mathematical statistics in the context of very simple examples it will be easier to use those concepts in the context of real problems, later.

2 Probability theory

Probability provides our basic model for dealing with chance events, and in this section we’ll develop a fairly precise mathematical description of probability, that will allow the derivation of rules for combining probabilities.

To start with consider a couple of working definitions of probability:

Definition. In a process which can be carried out in \( n \) equally likely ways, \( r \) of which lead to a particular result, the probability of the result is \( r/n \).

alternatively:

Definition. Consider a process which can give rise to a number of different results including a particular result, \( A \), at random. Suppose that the process is repeated independently \( n \) times and \( A \) occurs in \( r_n \) repeats. The probability of \( A \) is \( \lim_{n \to \infty} r_n/n \).

We will usually shorten “the probability that \( A \) occurs” to “Pr(\( A \))”.
2.1 A formal definition of probability

Now let’s make things more precise, and define some formal mathematical language for describing chance processes. This will allow the derivation of some very general results about probabilities, from some very simple first principles.

**Definition.** A *trial* is any process which, when repeated, generates a set of results or observations.

**Definition.** An *outcome* is the result of carrying out a trial.

**Definition.** An *event* is a set which consists of one or more of the possible outcomes of a trial.

To make this rather abstract set of definitions more concrete, consider a couple of simple examples.

**Example 1:** Consider a list of employees of a small supermarket: Susan, Leo, Joy, Thomas, Elsie and Katharine. Picking an employee at random from a list is a trial. The picked employee being Susan is an outcome (and a slightly trivial example of an event). The picked employee being a woman is an event (but not an outcome).

**Example 2:** Consider tossing a coin 3 times. The act of tossing the coin 3 times is a trial. The possible outcomes of the trial are HHH, HHT, HTH, HTT, THH, THT, TTH, TTT. The event “there were exactly 2 heads” is the set \{HHT, HTH, THH\}.

Notice one important point about events. An event occurs if the outcome of a trial lies in the set defining the event. The event is not the simultaneous occurrence of all the outcomes in the set.

To develop probability theory further, consider trials that give rise to equally likely possible outcomes. A wide range of practical probability problems can be reduced to this “equally likely outcome” model (and still more can be approximated using it). Given this “equally likely outcome” assumption, set theory provides a straightforward way of deriving most of the results in probability that we will need.

We will refer to the set of all possible outcomes of a trial as the sample space∗, \(S\), of the trial. Performing a trial consists of picking one of the elements of the sample space at random, in such a way that all points are equally likely to be picked. It’s useful to illustrate outcomes and events within \(S\) using Venn diagrams.

The supermarket employee example above might look like this, where \(W\) is the event “the employee is a woman”

---

*It’s known as the “sample space” rather than the “sample set” because in more advanced theory we may want to introduce some notion of distance into it.*
Draw the sample space for the coin tossing example yourself in this space, and show the event $H = \text{"the first coin toss gives a head"}$. 

Now let $n(X)$ denote the number of elements in a (sub)set $X$ . . .

**Definition.** If a trial has a set, $S$, of equally likely outcomes, then the probability of an event $E$ where $E \subseteq S$ is $\Pr(E) = n(E)/n(S)$.

For example: in the randomly selected employee case: $\Pr(W) = 4/6 = 2/3$; in the 3 coin tosses example $\Pr(H) = 4/8 = 0.5$.

**Exercise:** Find the probability of getting at least 2 heads in 3 coin tosses:

Several things follow immediately from the definition of probability:

1. For any event $E$: $0 \leq \Pr(E) \leq 1$.
2. $\Pr(S) = 1$.
3. $\Pr(\emptyset) = 0$ where $\emptyset$ is the empty set.
4. If $E^c$ is the set of all points in $S$ that are not in $E$ then it’s not hard to see that $\Pr(E^c) = 1 - \Pr(E)$. i.e. the probability that an event happens is one minus the probability that it does not happen (This is covered more carefully in section 2.3.3). Find the probability of obtaining less than 2 heads in 3 coin tosses:

### 2.2 Outcome counting

Directly counting outcomes in order to establish probabilities can get a little tedious if $S$ is large, and even quite simple trials can give rise to very large sample spaces. As an illustration, consider the trial “tossing a fair coin 10 times” . . . this has $2^{10} = 1024$ possible outcomes: HHHHHHHHHH, HHHHHHHHHT, HHHHHHHHTH, etc. Direct counting of all the outcomes in an event of interest would be rather boring (consider trying to work out the probability that there are 5 heads by working out all the outcomes in this event). Fortunately there are 4 simple results from the theory of permutations and combinations that simplify matters considerably. The first result is almost trivial:

**The Multiplication Principle:** if there are $n$ ways of carrying out $A$ and $m$ ways of carrying out $B$ then there are $nm$ ways of carrying out “$A$ and $B$”.

For example: if I possess 2 pairs of footwear (sandals and brogues) and 3 pairs of socks (black, white and a terribly amusing pair featuring a red-nosed raindeer and a bottle of whiskey) then (in addition to
being a little sad) there are \(2 \times 3 = 6\) ways in which I could choose to dress my feet: \{ sandals and white socks, brogues and white socks, sandals and black socks, etc \}.

The remaining results require a couple of definitions. Consider a set of objects.

**Definition.** A permutation is an ordering of all the objects.

**Definition.** A combination is a subset of the set of objects, in which the order of the objects is immaterial.

**Examples:** Consider the set of objects \{A,B,C,D\}. i) \{A,B,C,D\}, \{B,D,C,A\} and \{B,A,C,D\} are all different permutations of the set, but are all the same combination. ii) \{A,B\}, \{B,C\}, \{A,C\}, \{A,B,D\} are all the possible combinations of size 3 from the set.

**Theorem 1.** There are \(n!\) permutations of \(n\) objects.

This result is easily seen by considering, for example, the assembly of a list of 4 objects:

\[
\begin{align*}
&\text{A B C D} \\
&\text{A B C D} \\
&\text{A B C D} \\
&\text{A B C D}
\end{align*}
\]

There are 4 ways of filling position 1...

which leaves 3 ways of filling position 2...

which leaves 2 ways of filling position 3...

which leaves only 1 way of filling position 4.

So, by the multiplication principle there are \(4 \times 3 \times 2 \times 1 = 4!\) ways of arranging 4 objects in a list.

This argument clearly generalises to any finite number of objects.

i.e. in general there are \(n\) ways of filling the first slot in an ordered list, which uses up one object and leaves \(n - 1\) ways of filling the 2nd slot, which uses a second object so that there are \(n - 2\) ways of filling the third slot, etc. etc. until there is just one way left to fill the final slot. So by the multiplication principle there are \(n!\) ways of filling the list altogether.

**Theorem 2.** There are \(n!/(n-r)!\) different permutations of \(r\) objects chosen from \(n\).

To see this consider filling a list of length \(r = 4\) from a set of \(n = 6\) objects:

\[
\begin{align*}
&\text{A B C D} \\
&\text{A B C D} \\
&\text{A B C D} \\
&\text{A B C D}
\end{align*}
\]

There are 4 ways of filling position 1...

which leaves 3 ways of filling position 2...

which leaves 2 ways of filling position 3...

which leaves only 1 way of filling position 4.

So, by the multiplication principle there are \(6 \times 5 \times 4 \times 3 = 6!/4\) ways of arranging 4 objects selected from a set of 6 objects.

This argument clearly generalises to any finite number of objects and any length of list less than that number.

i.e. in general there are \(n\) ways of filling the first slot which leaves \(n - 1\) ways of filling the 2nd slot which in turn leaves \(n - 2\) ways of filling the 3rd slot, and so on until there are \(n - r + 1\) ways of filling the \(r\)th slot. That is there are \(n \times (n - 1) \times (n - 2) \times \ldots \times (n - r + 1) = n!/(n-r)!\) ways of arranging \(r\)

\[^1\text{Don’t forget } n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 = \text{ “n factorial”, and by convention 0! = 1.}\]
objects chosen from n. Note that it is common to write \(^nP_r\) for \(n!/(n-r)!\) (see calculator buttons, for example).

The final result also deals with choosing \(r\) objects from \(n\), but covers the case in which we don’t care about the order in which the objects are selected.

**Theorem 3.** The number of combinations of \(r\) things selected from \(n\) is:

\[
\frac{n!}{r!(n-r)!} \quad \left[\text{usually written } \binom{n}{r}\text{ or } ^nP_r\right]
\]

You should notice that \(\binom{n}{r} = \frac{n P_r}{r!}\): this is the key to understanding the result. \(^nP_r\) counts each different ordering of the same \(r\) objects selected from \(n\) separately. Since there are \(r!\) ways of arranging \(r\) objects this means that each combination of size \(r\) is counted \(r!\) times in \(^nP_r\). So the number of combinations of \(r\) objects from \(n\) must be \(\frac{n P_r}{r!}\).

**Example:** To see the relationship between \(\binom{n}{r}\) and \(^nP_r\) consider the example of selecting 3 objects from \{A, B, C, D\}. The following table illustrates what is being counted for the two cases:

<table>
<thead>
<tr>
<th>All permutations counted in (^nP_r = 4 P_3)</th>
<th>Corresponding combination counted by (\binom{n}{r} = \binom{4}{3})</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC  ACB  BAC  BCA  CAB  CBA</td>
<td>{A,B,C}</td>
</tr>
<tr>
<td>DBC  DCB  BDC  BCD  CDB  CBD</td>
<td>{B,C,D}</td>
</tr>
<tr>
<td>ADC  ACD  DAC  DCA  CAD  CDA</td>
<td>{A,C,D}</td>
</tr>
<tr>
<td>ABD  ADB  BAD  BDA  DAB  DBA</td>
<td>{A,B,D}</td>
</tr>
</tbody>
</table>

| Total = 4!/1! = 24 = \(^4P_3\)          | Total = 4!/(3!1!) = 4 = \(\binom{4}{3}\) |

**Example:** In the UK national lottery, the players select 6 numbers out of 49 (with no repeat numbers). If a player manages to choose the same six numbers as are selected randomly by the lottery machine, then that player wins the jackpot. If one trial consists of the lottery machine generating 6 numbers, how many elements are there in the sample space of this trial? Hence work out the probability of a player winning the lottery jackpot with a single go.

### 2.2.1 Example of outcome counting: the birthday problem

The key point when using counting methods is to identify exactly what the “outcomes” are that are “equally likely”. A classic example is the “Birthday problem”, which is easy once you spot the correct outcomes to consider, and dreadfully difficult otherwise. Here it is:

**What is the probability that at least 2 people in a class of 20 have the same birthday?**

Let’s make the modelling assumption that birthdays are equally likely on any day of the year, and neglect leap years. Then the problem breaks down as follows:

1. The “trial” in this case is the selection of 20 birthdays at random, and the “outcomes” are lists of 20 numbers from 1 to 365 corresponding to the days of the year. Under the assumptions made, all possible examples of such lists make up the sample space, \(S\) and all are equally likely.

2. Given the sample space, it’s probably easiest to count the number of outcomes corresponding to the event that no two people in the class have the same birthday: call this event \(B\). All the remaining outcomes correspond to the event that at least two people share a birthday.
3. So we are interested in how many lists of length 20 we could construct in which no birthdays occur twice: exactly the sort of situation covered by theorem 2. So \( n(B) \) is just the number of different permutations of \( r \) objects chosen from \( n: \frac{365!}{345!} \).

4. Clearly there are \((365)^{20}\) ways of writing down an outcome, so \( n(S) = (365)^{20} \).

5. Hence, the required probability is just \( \frac{n(S) - n(B))}{n(S)} \approx 0.41 \).

### 2.3 Probabilities of compound events

The simple probability theory developed so far makes it very easy to work out the answers to questions such as: “what is the probability of event \( A \) and event \( B \) occurring?”, “What is the probability of event \( A \) or event \( B \) happening?”, and so on.

#### 2.3.1 \( \Pr(A \text{ and } B) \)

The probability that \( A \) and \( B \) occur is \( \Pr(A \cap B) = \frac{n(A \cap B)}{n(S)} \). In the following example \( A \cap B \) is shaded and it’s clear that \( \Pr(A \cap B) = \frac{2}{16} = \frac{1}{8} \):

![Diagram of sets A and B intersecting]

To make things more concrete consider randomly picking an employee again, and now suppose that Susan, Leo and Joy are qualified “first-aiders”.

![Diagram of set S with subsets A, B, F, and W]

Clearly \( \Pr[\text{“Employee is a woman and a first aider”}] = \Pr(W \cap F) = \frac{n(W \cap F)}{n(S)} = \frac{2}{6} = \frac{1}{3} \).

Sometimes two events can not co-occur...

**Definition.** Events \( A \) and \( B \) are **mutually exclusive** iff \( A \cap B = \emptyset \). i.e. if \( \Pr(A \cap B) = 0 \).

For example, let \( M \) be the event that a randomly selected employee is a man.
Clearly \( W \cap M = \emptyset \), that is \( \Pr(\text{Employee is a woman and a man}) = \Pr(W \cap M) = 0 \): “being a woman” and “being a man” are mutually exclusive events. Note also that individual outcomes are mutually exclusive.

### 2.3.2 \( \Pr(A \text{ and/or } B) \)

Another combined outcome of interest is the case where either or both of events \( A \) and \( B \) occur. Continuing with the randomly chosen employee example:

\[ \Pr(\text{Employee is a woman and/or a first-aider}) = \Pr(W \cup F) = \frac{n(W \cup F)}{n(S)} = \frac{5}{6}. \]

One useful thing to be able to do is to work out \( \Pr(A \cup B) \) from \( \Pr(A) \), \( \Pr(B) \) and \( \Pr(A \cap B) \). This is straightforward:

\[
\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)
\]

To see this result in action, consider again the probability that a randomly selected employee is a Woman and/or a First-aider, i.e. \( \Pr(W \cup F) \). We have that \( \Pr(W) = \frac{2}{3} \), \( \Pr(F) = \frac{1}{2} \) and \( \Pr(W \cap F) = \frac{1}{3} \). Hence applying the result just derived: \( \Pr(W \cup F) = \frac{2}{3} + \frac{1}{2} - \frac{1}{3} = \frac{5}{6} \), as expected.

Obviously, if \( A \) and \( B \) are mutually exclusive:

\[
\Pr(A \cup B) = \Pr(A) + \Pr(B)
\]

**Exercise:** The probability of my front bicycle tyre puncturing on the way home is 0.01 while the probability of my back tyre puncturing is 0.015. The probability of punctures in both tyres is 0.005. What is my probability of making it home without a puncture?
2.3.3 Complementary events

**Definition.** The complementary event $A^c$ to event $A$ is the set of all points (outcomes) in the sample space $S$ that are not in $A$.

So, $A \cup A^c = S$. Since $Pr(S) = 1$ and $A$ and $A^c$ are mutually exclusive:

$$Pr(A \cup A^c) = 1 \Rightarrow Pr(A) + Pr(A^c) = 1 \Rightarrow Pr(A^c) = 1 - Pr(A)$$

**e.g.** Considering our running example, the event “a randomly chosen employee is not a first aider” is $F^c$, and $Pr(F^c) = 1 - Pr(F) = 0.5$.

2.4 Conditional Probability

Often we would like to know the probability that some event occurs *given* that some other event in the same sample space has occurred. For example the probability that a randomly chosen school pupil is over 2m tall *given* that they are in the sixth form, will be rather different from the probability that a randomly chosen school pupil if over 2m tall, given no extra information about them.

Once again use of set theory makes conditional probabilities quite easy to handle. First consider an example of 2 events $A$ and $B$ in a sample space $S$:

![Diagram of sample space S with events A and B]

Now, in this example $Pr(A) = 6/16 = 3/8 = Pr(B)$, but suppose that $A$ is known to have occurred: what is the probability of $B$ given that $A$ has happened (usually written $Pr(B|A)$)? Given that $A$ is known to have happened, the sample space has effectively shrunk to $A$, while event $B$ has effectively shrunk to $A \cap B$:

![Sample space given that A has occurred]

so in this example, it’s clear that $Pr(B|A) = 2/6 = 1/3$, while in general $Pr(B|A) = n(A \cap B)/n(A)$.

As usual, it is helpful to have a form for $Pr(A|B)$ that doesn’t necessarily involve counting outcomes directly. To obtain this just divide top and bottom of the expression for $Pr(A|B)$ by $n(S)$:

$$Pr(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{n(A \cap B)/n(S)}{n(B)/n(S)} = \frac{Pr(A \cap B)}{Pr(B)}$$

---

*it’s important to note that $B|A$ does not represent a set*
2.4.1 Total probability rule

Consider the situation in which the sample space is partitioned into 3 mutually exclusive events $A_1$, $A_2$ and $A_3$ (so $S = A_1 \cup A_2 \cup A_3$) as illustrated here:

Clearly $B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B)$. Which means that:

$$\Pr(B) = \frac{n(A_1 \cap B) + n(A_2 \cap B) + n(A_3 \cap B)}{n(S)} = \frac{n(A_1 \cap B)}{n(S)} + \frac{n(A_2 \cap B)}{n(S)} + \frac{n(A_3 \cap B)}{n(S)}$$

now multiplying the ith term by $n(A_i)/n(S)$ gives an interesting result:

$$\Pr(B) = \frac{n(A_1 \cap B)}{n(A_1)} \frac{n(A_1)}{n(S)} + \frac{n(A_2 \cap B)}{n(A_2)} \frac{n(A_2)}{n(S)} + \frac{n(A_3 \cap B)}{n(A_3)} \frac{n(A_3)}{n(S)}$$

$$= \Pr[B|A_1] \Pr[A_1] + \Pr[B|A_2] \Pr[A_2] + \Pr[B|A_3] \Pr[A_3]$$

Obviously this generalizes. If $A_1, A_2, \ldots, A_n$ are mutually exclusive events which partition the whole of the sample space (so $A_1 \cup A_2 \cup \ldots \cup A_n = S$), and $B$ is another event in the sample space:

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B|A_i) \Pr(A_i)$$

**Example:** Suppose that the probability that a randomly selected person has HIV is 0.001. A test for HIV will give a positive result with probability 1 if the subject really has HIV and will give a positive result with probability 0.01 if the person does not have HIV. What is the probability that a randomly selected person will test positive?

2.4.2 Bayes’ Rule

From the definitions of conditional probabilities $\Pr(A|B)$ and $\Pr(B|A)$ it’s easy to see that:

$$\Pr(A \cap B) = \Pr(A|B) \Pr(B) = \Pr(B|A) \Pr(A)$$

Re-arranging yields the second most influential result in statistics:

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)}$$
Example: Continuing with the HIV example. What is the probability that a randomly selected person has HIV if they test positive?

2.4.3 Independence

If knowing that event $A$ has occurred does not change the probability of event $B$, then events $A$ and $B$ are independent and $Pr(A|B) = Pr(A)$. This means that:

$$Pr(A \cap B) = Pr(A|B)Pr(B) = Pr(A)Pr(B)$$

and this is used as the defining feature of independence.

Definition. Events $A$ and $B$ are independent iff $Pr(A \cap B) = Pr(A)Pr(B)$

The generalization to any number of events is a slightly complicated process involving consideration of all possible combinations of the events concerned, however the “only if” part of the definition generalizes easily:

If events $A_1, A_2, A_3, \ldots, A_n$ are mutually independent then

$$Pr(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n) = Pr(A_1)Pr(A_2)Pr(A_3)\ldotsPr(A_n).$$

Statistical independence is an important idea to understand, as it is a key assumption of many statistical models and procedures.

Example: In the random employee selection example, $Pr(W) = 2/3$, $Pr(F) = 1/2$ and $Pr(W \cap F) = 1/3$, so $Pr(W \cap F) = Pr(W)Pr(F)$, meaning that events $F$ and $W$ are independent. In other words probability of being a first-aider is independent of gender, in this case.

2.5 Generalizations

To simplify derivation of the theory presented above, we only considered finite countable sample spaces in which each possible outcome is equally likely. The same results turn out to hold for much more general sample spaces. For example, the assumption of equally likely outcomes can be relaxed (allowing us to deal with events whose probability of occurance is an irrational number, for example). Then each outcome, $o_i$ in the sample space has an associated probability $p_i$, and the definition of the probability of an event $E$ becomes:

$$Pr(E) = \sum_{o_i \in E} p_i$$

You can see why this must be in a couple of ways:

1. $E = o_{(1)} \cup o_{(2)} \cup o_{(3)} \cup \ldots \cup o_{(n(E))}$ where $o_{(1)}$, $o_{(2)}$, $o_{(3)}$, $\ldots$, $o_{(n(E))}$ are the outcomes in $E$, but the outcomes $o_{(i)}$ are themselves mutually exclusive events. Hence:

$$Pr(E) = Pr(o_{(1)}) + Pr(o_{(2)}) + Pr(o_{(3)}) + Pr(o_{(n(E))}) = \sum_{o_i \in E} p_i$$

2. The definition is a natural generalisation of our original definition of probability of an event. To see this, note that under the “equally likely outcome” assumption $p_i = 1/n(S) \forall i$.

$$\Rightarrow Pr(E) = \sum_{o_i \in E} \frac{1}{n(S)} = \frac{1}{n(S)} \sum_{o_i \in E} 1 = \frac{n(E)}{n(S)}$$

The important point to note is that under all generalizations of the sample space necessary to deal with statistical problems, the basic rules of probability that we have derived still hold.
3 Random Variables (r.v.’s)

Now consider the case in which each point (outcome) in the sample space has a real number associated with it, so that each time a point in the space is sampled a value is obtained. Let’s denote this quantity, which varies across the sample space, by $X$. The interesting “events” in the sample space will now be of the form:

“Variable $X$ takes particular value $x$”,

or more generally:

“Variable $X$ takes values between $a$ and $b$”.

Returning to the employee list example, the variable $X$ might be the age in years of a randomly selected employee . . .

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Susan, 40  Leo, 28</td>
</tr>
<tr>
<td>Joy, 20</td>
<td>Thomas, 20</td>
</tr>
<tr>
<td>Elsie, 51</td>
<td>Katharine, 35</td>
</tr>
</tbody>
</table>

So, for example, $\Pr[0 < X < 17] = 0$, $\Pr[X = 20] = 2/6 = 1/3$, $\Pr[X \geq 20] = 1$, $\Pr[X > 20] = 4/6 = 2/3$, etc.

**Definition.** A Random Variable (r.v.) associates a real number with each point in the sample space.

Usually in statistics we are interested in learning about a population using a random sample taken from that population (e.g. we would like to find out about the voting intentions of the country based on the intentions of a random sample of voters, or we would like to know how exposure to depleted uranium affects health of people in general, based on a random sample of people with varying levels of exposure to DU, etc.). In this context it’s helpful to have an equivalent working definition of a random variable:

**Definition.** A random variable is any quantity or attribute whose value varies from one unit of a (possibly hypothetical) population to another.

For example height of british people is a random variable. If you pick somebody randomly from the whole population, you don’t know their height in advance, although you’ll have some idea of the distribution of heights that you expect (i.e. high probability of being between 1.4m and 2m, with a substantial but lower probability of being less than 1.4m and a much lower probability of being over 2m).

When we need to make a clear distinction between a random variable and a particular observation of it, we will usually give the random variable a capital letter, $X$, for example, while a particular observation of the random variable will be given the corresponding small letter: $x$, in this case. So, $x$ stands for a particular number obtained by observing $X$, while $X$ itself stands for something which may take any of a range of values at random. In the employee list example, $X$ stands for “age of a randomly selected employee”, while if we wanted to have labels for the particular values obtained the first and second times that an employee was randomly selected, we might use $x_1$ and $x_2$. Suppose that 2 employees are selected randomly and independently and they turn out to be Joy and Elsie. In this case $x_1 = 20$ and $x_2 = 51$ (while $X$ is still “age of a randomly selected employee”).
Random variables are the main building blocks for statistical models, so it is worth developing some theory for handling them (we want to avoid having to go back to considering events in the sample space every time anything needs to be calculated).

The most important concept associated with a random variable is its distribution: that is the range of values that it can take and the probabilities associated with it taking any particular values. But to develop a mathematical description of distribution a distinction needs to be drawn between discrete and continuous r.v.s.

3.1 Discrete random variables

A discrete random variable is a random variable that can only take discrete values: usually (but not always) non-negative integers. A continuous random variable, on the other hand, can take any real value within some defined range. Although most of the main concepts are the same, there are some differences in the details of the theory for discrete and continuous r.v.s, so they will be dealt with in separate sections.

It’s often useful to represent the distribution of a discrete random variable using a table. For example, consider the random variable $X$ defined on the sample space $S$ (of equi-probable outcomes):

\[
\begin{array}{ccc}
0 & 1 & 2 \\
0.2 & 0.4 & 0.3 & 0.1
\end{array}
\]

...this is very compact relative to drawing and labelling the sample space.

Example: Let $Y =$ “number of heads in 3 tosses of a coin”. In this case it’s easy to illustrate the sample space:

\[
\begin{array}{ccc}
\text{HHH (3)} & \text{HHT (2)} & \text{HTH (2)} \\
\text{HHT (2)} & \text{THH (2)} & \text{THT (1)} \\
\text{HTT (1)} & \text{TTH (1)} & \text{TTT (0)}
\end{array}
\]
...and tabulate the distribution of $Y$:

Tabulation is fine for r.v.s that can only take a small number of possible values, but becomes unwieldy as the number of possible values increases (and is impossible if the number of possible values is infinite). Hence an even more compact way of describing the distribution is required:

**Definition.** The probability mass function (p.m.f) of a random variable $X$ is the function, $f(x)$ such that:

$$f(x) = \Pr[X = x].$$

i.e. if you feed some particular value $x$ into $f(x)$ then what you get back is the probability that $X$ takes that particular value $x$.

**Example:** if $Y =$ “number of heads in 3 coin tosses” then its p.m.f is:

$$f(y) = \begin{cases} \frac{1}{4} - \frac{1}{4}|y - 1.5| & \text{if } y = 0, 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}$$

(where “$|x|$” means “the magnitude of $x$”). You can check this - plug in any of the possible values of $Y$ and you get back the probability of its occurrence.

To get an idea of the shape of a distribution it’s a good idea to plot the p.m.f. For example:

![p.m.f of "number of heads in 3 coin tosses"](image)

**Exercise:** Let $X$ be a r.v. with p.m.f:

$$f(x) = \frac{12}{25} \frac{1}{x} \quad x = 1, 2, 3or4$$

Evaluate: Pr[$X = 2$], Pr[$X < 3$] and Pr[$1 \leq X \leq 4$].

### 3.1.1 Expectation and variance of r.v.s

The p.m.f. completely describes the random variable that it applies to, but it is often useful to have simpler, single value summaries that describe different properties of the distribution in a way that is easy to compare across random variables with different distributions. The most obvious such summary is the mean value of a random variable:
Definition. The **mean**, **average** or **expected value** of a random variable, $X$ which can take values $x_1, x_2, \ldots, x_n$ and has p.m.f. $f(x)$ is:

$$E(X) = \sum_{all \ x_i} f(x_i)x_i$$

(“$E(X)$” means “the expectation of $X$” or “the expected value of $X$”). It’s not hard to see that the expected value of a random variable is just the average over the whole sample space of the random variable (or equivalently the average over the whole population). To see this most clearly consider, for example, a sample space of equi-probable outcomes with an r.v., $X$, defined on it:

![Sample Space Diagram]

Clearly the mean of $X$ is simply:

$$\begin{align*}
  & \frac{x_1 + x_1 + x_2 + x_2 + x_2 + x_3 + x_3 + x_3 + x_4 + x_4 + x_4}{12} \\
  & = \frac{2}{12}x_1 + \frac{3}{12}x_2 + \frac{4}{12}x_3 + \frac{3}{12}x_4 \\
  & = \Pr(X = x_1)x_1 + \Pr(X = x_2)x_2 + \Pr(X = x_3)x_3 + \Pr(X = x_4)x_4 \\
  & = \sum_{i=1}^{4} f(x_i)x_i = E(X)
\end{align*}$$

**Exercise.** Evaluate the expected value of the number of heads in 3 coin tosses.

Since a function of a random variable is also a random variable, it has an expected value. If $X$ is an r.v. taking values $x_1, x_2, \ldots, x_n$ with p.m.f. $f(x)$ and $g(X)$ is any function of $X$:

$$E(g(X)) = \sum_{all \ i} f(x_i)g(x_i)$$

For example:

$$\begin{align*}
  & E(X^2) = \sum_{all \ i} f(x_i)x_i^2; \quad E[\log(X)] = \sum_{all \ i} f(x_i)\log(x_i), \ \text{etc.}
\end{align*}$$

One important special case of the above is the variance of a random variable, which measures how spread out the distribution of the r.v. is around its mean.

---

\*Despite the name the expected value of a random value is **not** the most probable value (that’s the mode).
**Definition.** The variance\(^\dagger\) of a random variable \(X\) with mean \(\mu = E(X)\) is

\[
\text{var}(X) = E[(x - \mu)^2].
\]

Clearly

\[
\text{var}(X) = \sum_{i} f(x_i)(x_i - \mu)^2.
\]

For calculation purposes note that \(\text{var}(X) = E(X^2) - [E(X)]^2\). This easily proven (don’t learn this, just understand it!):

\[
\sum_{i} f(x_i)(x_i - \mu)^2 = \sum_{i} f(x_i)(x_i^2 - 2\mu x_i + \mu^2)
\]

\[
= \sum_{i} f(x_i)x_i^2 - 2E(X) \sum_{i} f(x_i)x_i + [E(X)]^2 \sum_{i} f(x_i)
\]

\[
= E(X^2) - 2E(X)E(X) + [E(X)]^2 \times 1 = E(X^2) - [E(X)]^2
\]

**Exercise:** Evaluate the variance of the number of heads in 3 coin tosses:

Finally . . .

**Definition.** The **standard deviation** of r.v. \(X\) is \(\sqrt{\text{var}(X)}\).

### 3.1.2 Some useful distributions for modelling discrete r.v.s

3 standard distributions will be considered here, partly by way of illustration of the concepts just covered, and partly because they are of practical use for constructing models.

**The Bernoulli distribution**

The Bernoulli distribution is probably the simplest distribution that is of any practical use. The p.m.f. for a Bernoulli random variable is:

\[
f(x) = \begin{cases} 
  p & \text{if } x = 1 \\
  1 - p & \text{if } x = 0 \\
  0 & \text{otherwise}
\end{cases}
\]

so a Bernoulli r.v. takes the value 1 with probability \(p\) and the value 0 with probability \(1 - p\). It’s useful as a model of a trial that can “succeed” \((x = 1)\) with probability \(p\) or fail \((x = 0)\) with probability \(1 - p\). (e.g. Did the patient recover, or not? Did the student pass or not? Did we find mushrooms at this location or not? etc.)

**Notation:** “\(\sim\)” is used to mean “is distributed as”. e.g. “\(X \sim \text{Bern}(p)\)” means “\(X\) has a Bernoulli distribution with parameter \(p\).”

It’s useful to be able to express the mean and variance of most distributions in terms of any parameters of the distribution \((p\) in the Bernoulli case). For the Bernoulli distribution this is very easy to do. If \(X \sim \text{Bern}(p)\):

\[
E(X) = 0 \times (1 - p) + 1 \times p = p
\]

\(^\dagger\)Notice that the variance of a random variable is defined over the whole sample space and is a fixed (non-random) quantity - this contrasts with the sample variance which we met before, which varies from sample to sample, and is best viewed as an estimate of the variance defined here.
similarly:

\[ E(X^2) = 0^2 \times (1 - p) + 1^2 \times p = p \]

so,

\[ \text{var}(X) = \]

**The binomial distribution**

The Bernoulli distribution covers the case of a single “trial” that can succeed or fail: it’s fairly straightforward to generalize and get a much more useful distribution by considering the number of successes in several trials. Let \( X \) be the number of successes in \( n \) independent Bernoulli trials each with probability of success \( p \). \( X \) is said to be a binomial random variable, with p.m.f.

\[ f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n \]

(remember from section 2.2 that \( \binom{n}{x} = \frac{n!}{x!(n-x)!} \)). It’s easy to derive the form of this p.m.f. using the probability theory already developed. For example, consider a sequence of 10 independent Bern(\( p \)) trials with 4 successes:

0010110001

say. Since the Bernoulli trials are independent this sequence clearly has probability:

\[ (1-p) \times (1-p) \times p \times (1-p) \times p \times (1-p) \times (1-p) \times (1-p) \times p = p^4 (1-p)^6 \]

In fact any sequence of results of 10 similar trials containing 4 successes has probability \( p^4 (1-p)^6 \) of occurring. Furthermore, all possible “4 success” sequences are mutually exclusive. Hence to get the probability of observing 4 successes in 10 trials we need to add up the probabilities of each of the mutually exclusive ways that this can happen: i.e. add up the probabilities of each of the sequences. Since all the sequences of interest have the same probability, all we really need to do is to count up how many of them there are. At first sight this looks a little daunting, but the way to think about it is this: to construct a sequence you are really choosing 4 of the 10 locations in the list to receive 1’s (the rest get 0’s): each distinct combination of 4 sites chosen from 10 corresponds to one of the sequences of interest. So you need to know the number of combinations of size 4 that can be selected from 10: from section 2.2 this is \( \binom{10}{4} \). Hence the probability of 4 successes out of 10 trials is:

\[ \binom{10}{4} p^4 (1-p)^6 \]

The following diagram is another attempt to explain this . . .

![Diagram showing combinations of successes and failures](image)

It is very easy to see how this generalizes to the give the binomial p.m.f. Note also that if \( X \sim \text{bin}(n, p) \) then \( E(X) = np \) and \( \text{var}(X) = np(1-p) \). These results are quite easy to derive using the fact that \( X \) can be viewed as the sum of \( n \) independent Bern(\( p \)) r.v.s, but a couple of general results on sums of expectations and variances are needed to do so (see section 3.4.4).
Exercise: What is the probability of 7 heads in 10 tosses of a fair coin?

3.1.3 The Poisson distribution

Suppose that \( X \sim \text{bin}(n, p) \) where \( p \) is small while \( n \) is very large. In this case \( \Pr(X = x) \) scarcely depends on \( n \) at all and \( X \) approximately follows the Poisson distribution. A random variable \( Y \) has Poisson distribution with mean \( \lambda \) if its p.m.f. is:

\[
f(y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad y = 0, 1, 2, \ldots, \infty
\]

Proving that a Poisson r.v. is a limiting case of a binomial r.v. for which \( n \) is large and \( p \) is small, involves a limiting argument where \( n \) is allowed to tend to \( \infty \) while the mean of the binomial r.v. \( np \) is held constant. The proof is quite straightforward, but a bit long-winded, so it is omitted here. However, the following figure provides an example of the process on which the proof is based:

Moving from top left to bottom right, the first 5 figures each show a histogram of 10000 binomial random deviates from distributions with successively increased values for \( n \), but with \( p \) set so that \( np = 2 \) in each case (i.e. the expected value of each binomial deviate is 2). The final panel\(^\dagger\) shows the histogram of 10000 random deviates from a Poisson distribution with mean 2. As you can see as \( n \) increases while the mean is kept constant the binomial distribution tends to the Poisson distribution.

Note that if \( Y \sim \text{Poi}(\lambda) \) then \( E(Y) = \lambda \) and \( \text{var}(Y) = \lambda \). It is also possible to show that if \( Y_1 \sim \text{Poi}(\lambda_1) \) and \( Y_2 \sim \text{Poi}(\lambda_2) \) then \( Y_1 + Y_2 \sim \text{Poi}(\lambda_1 + \lambda_2) \).

Examples: The Poisson distribution is useful for modelling many sorts of data that can be characterized as “counts”: for example numbers of customers arriving at a queue in some specified time interval, number of fish eggs in a net hauled up from the bottom of the sea to the top.

Exercise: The number of calls arriving at a call centre every minute can be modelled as a Poisson random variable with mean 4 (per minute). What is the probability of no calls arriving in a given minute?

\(^\dagger\)Each binomial plot was generated with a command like \texttt{hist(rbinom(10000,100,0.02),main=“bin(100,0.02)“)}, whereas the Poisson plot was generated with the command \texttt{hist(rpois(10000,2),main=“poi(2)“)}. 
3.2 Dealing with several r.v.s: joint distributions

In practical modelling we are usually interested in modelling several data at once, not just a single observation. For this reason we need to have some theory for dealing with the joint distribution of multiple random variables. Returning again to the randomly selected employee example, you can imagine having two random variables defined on the sample space: employee age and employee weight: there may well be some relationship between these, in the sense that knowing one of them may improve the accuracy with which you can guess the other one. This section looks at such joint distributions mathematically.

**Definition.** If \(X\) and \(Y\) are discrete r.v.s then their **joint p.m.f.** \(f(x, y)\) is the function such that:

\[
\Pr[X = x \text{ and } Y = y] = f(x, y).
\]

For example: let \(X\) be number of fillings and \(Y\) be average number of times per day that teeth are cleaned for primary school children in some area. A table of the joint p.m.f. of \(X\) and \(Y\) might be:

<table>
<thead>
<tr>
<th>(x)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.02</td>
<td>0.15</td>
<td>0.45</td>
<td>0.04</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.07</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.02</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

This means, for example, that the probability that a randomly chosen child brushes their teeth once a day and has 2 fillings is 0.02. The joint p.m.f. is especially useful for working out the probabilities for more complicated events. For example, to work out \(\Pr[Y > 1 \text{ and } X < 2]\) we just need to add up the probabilities for all the combinations of \(X\) and \(Y\) for which this is true: i.e. \(\Pr[Y > 1 \text{ and } X < 2] = 0.45 + 0.04 + 0.1 + 0.01 = 0.6\). One important special case is the calculation of so called **marginal probabilities** for \(X\) and \(Y\). For example, suppose that we want to know \(\Pr[X = 2]\) - to find this we need to add up the probabilities for each combination of \(X\) and \(Y\) for which \(X = 2\): this gives \(\Pr[X = 2] = 0 + 0.02 + 0.1 + 0 = 0.12\). So, given the joint p.m.f. of \(X\) and \(Y\) it is easy to work out the p.m.f. of \(X\) or the p.m.f of \(Y\), which in this context are known as the **marginal p.m.f.s.** In general:

\[
\Pr[X = x] = \sum_{\text{all } y} f(x, y) \quad \text{and} \quad \Pr[Y = y] = \sum_{\text{all } x} f(x, y)
\]

Usually, to distinguish the marginal distributions from the joint distribution in this context, we write \(f_x(x) = \Pr[X = x]\) for the p.m.f. of \(X\) and \(f_y(y) = \Pr[Y = y]\) for the p.m.f. of \(Y\).

**Exercise:** Tabulate the marginal p.m.f.s of \(X\) and \(Y\) for the toothcare example:

Obviously the notion of a joint p.m.f. generalizes to any number of r.v.s: if \(X_1, X_2, X_3, \ldots, X_n\) are discrete r.v.s then their joint p.m.f. is \(f(x_1, x_2, x_3, \ldots, x_n) = \Pr(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } X_3 = x_3 \text{ and } \ldots \text{ and } X_n = x_n)\). The relationship of marginals to the joint p.m.f. is also similar. For example if \(X, Y\) and \(Z\) are r.v.s with joint p.m.f. \(f(x, y, z)\), then the marginal p.m.f. for \(X\) is just:

\[
f_x(x) = \sum_{\text{all possible } z, y \text{ pairs}} f(x, y, z).
\]
3.2.1 Independence of r.v.s

The notion of independence of r.v.s follows directly from the definition of independence of events:

**Definition.** Random variables $X_1, X_2, \ldots, X_n$ ($n \geq 2$) with joint p.m.f. $f(x_1, x_2, \ldots, x_n)$ and marginal p.m.f.s $f_{x_i}(x_i)$ $i = 1, \ldots, n$ are independent iff:

$$f(x_1, x_2, \ldots, x_n) = f_{x_1}(x_1)f_{x_2}(x_2)\ldots f_{x_n}(x_n)$$

i.e. if their joint p.m.f. is given by the product of their marginal p.m.f.s.

Intuitively, independence of two r.v.s means that observing the value of one of them tells you no more about the likely value of the other than you already knew.

**Exercise:** consider the two pairs of r.v.s $U,V$ and $X,Y$: only one of the pairs are independent r.v.s - which pair is it?

<table>
<thead>
<tr>
<th>v</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

---

3.3 $\Pr[X \leq x]$: the cumulative distribution function (c.d.f.)

As you will see later, one of the most common sorts of quantity that needs to be calculated in statistics is: “what is the probability that random variable $X$ is less than or equal to some particular value, $a$, say?” Obviously this is straightforward to work out from the p.m.f.

$$\Pr[X \leq a] = \sum_{x \leq a} f(x)$$

but this sort of probability is required so often that it’s useful to define a function that returns it directly:

**Definition.** The cumulative distribution function, $F(x)$, of a random variable, $X$, is the function such that:

$$\Pr[X \leq x] = F(x)$$

For example, here is the c.d.f. of the number of fillings for a randomly selected child from the example in the previous section.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$(-\infty, 0)$</th>
<th>$[0,1)$</th>
<th>$[1,2)$</th>
<th>$[2,3)$</th>
<th>$[3,4)$</th>
<th>$[4,\infty)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pr[X \leq x]$</td>
<td>0</td>
<td>0.66</td>
<td>0.84</td>
<td>0.96</td>
<td>0.97</td>
<td>1</td>
</tr>
</tbody>
</table>

§

It’s useful to have a picture of what a c.d.f. looks like, so here is a plot of the above c.d.f. . . .

---

\[\text{Note that } [1,2) \text{ means “the interval 1 to 2, including 1 but excluding 2”. Similarly (1,2) would mean “the interval 1 to 2 excluding 1 and 2”, and so on for the other possible combinations of bracket.}\]
The plotted •’s are there to emphasise that 0 is part of the line segment covering $0 \leq x < 1$, 1 is part of the line segment covering $1 \leq x < 2$, and so on: hence, for example, the plot correctly implies that $\Pr(X < 1) = 0.66$ while $\Pr(x \leq 1) = 0.84$.

### 3.4 Continuous random variables

So far we have covered theory for discrete random variables. These are useful for modelling discrete quantities like “number of children in a family”, “age in years”, “number of whales sighted in a survey” and so on, but we also need to be able to deal with quantities like “mean annual temperature”, “height of a person”, “amount of light energy received by a plant” etc., which are all continuous variables - the values can be any real number within the appropriate ranges.

So, continuous random variables are r.v.s that can take any real value within some specified range. Such r.v.s can obviously take an uncountably infinite set of values, so that the sample spaces associated with them must also have an uncountably infinite number of elements. This makes direct use of the sample space rather technical.

The easiest way to start describing the distributions of continuous r.v.s is via the c.d.f. Exactly as in the discrete case the c.d.f. $F(x)$ of a continuous r.v. $X$ is the function with the property $F(x) = \Pr[X \leq x]$. The only real difference between the c.d.f.’s of continuous and discrete r.v.s is that the c.d.f. of a continuous r.v. doesn’t usually have any discontinuities, being a smooth increasing curve, rather than a series of steps.

**Example:** The exponential distribution is often used as a model of the time between events that occur at random (e.g. arrivals at a shopping checkout, or calls at a telephone exchange). If $X$ has an exponential distribution then its c.d.f. is:

$$F(x) = \begin{cases} 
1 - e^{-\lambda x} & x \geq 0 \\
0 & x < 0
\end{cases}$$

where $\lambda$ is a parameter.

Note that like all c.d.f.s this one increases monotonically from 0 at $-\infty$ to 1 at $\infty$, although in practice the range of $x$ values for which $F(x)$ is anything other than 0 or 1 may be smaller than $(-\infty, \infty)$.

As in the discrete case all sorts of useful probabilities can be calculated directly from a c.d.f. for example:

$$\Pr[X \leq a] = F(a)$$
$$\Pr[X > a] = 1 - F(a)$$
$$\Pr[a < X \leq b] = \Pr[X \leq b] - \Pr[X \leq a] = F(b) - F(a)$$

---

*Note that what mathematicians mean by “countable” is really closer to “could-start-counting-able” - they are not really concerned with whether you would ever finish. Hence, for example, the integers are countable, despite the fact that there are an infinite number of them, since you can put them in order and start counting them. You can’t even start on the process of putting the reals in order and counting them…*
Exercise: While the time between departures of aeroplanes from an airport is easy to control, arrival intervals are inevitably quite random, so that the air-traffic control and holding system has to be carefully planned to cope. Suppose that the time between planes arriving at Heathrow follows an exponential distribution with $\lambda = 0.5$. What is the probability that two planes arrive separated by less than 1 minute?

3.4.1 \textbf{Pr}[X = x] = 0 if $X$ is a continuous r.v.\textbf{.}

In contrast to the discrete case and provided that the c.d.f. is continuous: \( \text{Pr}[X \leq a] = F(a) = \text{Pr}[X < a] \). Hence, \( \text{Pr}[X = a] = \text{Pr}[X \leq a] - \text{Pr}[X < a] = F(a) - F(a) = 0 \). This can seem like an alarming property of something that we are going to use to model the real world. e.g. if we are using a continuous r.v. to model the process of taking a measurement, then the model seems to be saying that any particular outcome of the measurement is impossible! In fact there is no problem if we think about the measurement process a little more carefully: in practice any real measurement can only be made to a finite number of significant figures - all values of the quantity being measured are effectively rounded - hence any particular measurement is really recording the fact that the quantity being measured was within some small interval - an event that usually has a non-zero probability under the model. e.g. if we measure heights of randomly selected people to the nearest cm a measurement of 169cm really means that the height was in the range 168.5cm to 169.5cm, not that it was 1.6900000000 cm. Now you might still object that the person had a real height, and this talk of measurement only fudges the issue. I don’t think that such an argument holds up - as we start piling on the decimal places the idea of somebody having a single height gets less and less easy to sustain. Even if you look at something inanimate, like a metal bar kept at a constant temperature, there are only so many significant figures that you can use to describe its length before you get down to the atomic level, and have to worry about exactly where the edge of an atom is: quantum mechanics ends up saying that you can not fix this exactly. Similar arguments apply to all other notionally continuous quantities.

In practice, you can almost always use continuous r.v.s without ever worrying about these issues.

3.4.2 \textbf{Probability density functions (p.d.f.)}\textbf{.}

Although the probability of a continuous r.v. taking any particular value is zero, it is still useful to be able to find out how probable it is that an r.v. will lie in some small region around any particular value. This is where the idea of a probability density function comes in. Assuming that the c.d.f. is continuous then:

\textbf{Definition.} The \textbf{probability density function} (p.d.f.) of a random variable $X$ with c.d.f. $F(x)$ is

\[ f(x) = \frac{dF}{dx} \]

\( f(x) \) is the probability per unit $x$ interval at $x$.

For example, the p.d.f. of an exponential r.v. with parameter $\lambda$ is:

\[ f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \]
Considering the definition of the p.d.f. given above, we get to the key property of a p.d.f. For any real numbers \( a \) and \( b \) such that \( b \geq a \):

\[
\int_a^b f(x) \, dx = \int_a^b \frac{dF}{dx} \, dx = F(b) - F(a) = \Pr[a < x \leq b]
\]

i.e. the probability that r.v. \( X \) lies between \( a \) and \( b \) is given by the area under \( f(x) \) between \( a \) and \( b \), as illustrated here for the case where \( X \) has an exponential distribution . . .

It’s important to re-emphasise that \( \Pr[X = a] \neq f(a) \) if \( f(x) \) is a p.d.f. In fact, as we have seen \( \Pr[X = a] = \int_a^a f(x) \, dx = 0 \). On the other hand if \( \epsilon \) is any small number then \( \Pr[a - 0.5\epsilon < X \leq a + 0.5\epsilon] \approx \epsilon f(a) \).

Note that from the definition of a p.d.f:

\[
\int_{-\infty}^{\infty} f(x) \, dx = 1 \quad \text{and} \quad f(x) > 0.
\]

**Exercise:** Consider a random variable \( Y \) with p.d.f.

\[
f(y) = \begin{cases} 
    c/y^2 & y > 1 \\
    0 & \text{otherwise}
\end{cases}
\]

What must \( c \) be?

### 3.4.3 Joint distributions, independence and expectation

Once you have grasped the difference between a p.m.f. and a p.d.f. the rest of the results that were covered for discrete r.v.s have quite natural equivalents for continuous r.v.s. The main difference is that integrals replace summations.

**Definition.** The joint p.d.f. of continuous r.v.s \( X \) and \( Y \) is the function \( f(x, y) \) such that:

\[
\Pr[a < X \leq b \text{ and } c < Y \leq d] = \int_a^b \int_c^d f(x, y) \, dy \, dx
\]

for any real numbers \( a, b, c \) and \( d \) such that \( a \leq b \) and \( c \leq d \).\(^\text{II}\)

\(^\text{II}\)Strictly this should be more general: the region of the \( x - y \) plane doesn’t have to be rectangular.
Similar definitions hold for p.d.f.s of more than 2 r.v.s

**Example:** Let $X$ and $Y$ be r.v.s with joint p.d.f.

$$f(x, y) = 4xy \quad 0 \leq x \leq 1, \ 0 \leq y \leq 1$$

Evaluate $\Pr[X < 0.5 \text{ and } Y < 0.5]$.

The relationship between a joint p.d.f. and the corresponding marginal p.d.f.s is much the same as in the discrete case for p.m.f.s. For example, if $X$ and $Y$ are continuous r.v.s with joint p.d.f. $f(x, y)$ and marginal p.d.f.s $f_x(x)$ and $f_y(y)$, then:

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{and similarly} \quad f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

The definition of independence of continuous r.v.s is just like its counterpart for discrete r.v.s:

**Definition.** r.v.s $X_1, X_2, \ldots, X_n$ are independent iff:

$$f(x_1, x_2, \ldots, x_n) = f_{x_1}(x_1) f_{x_2}(x_2) \ldots f_{x_n}(x_n)$$

where $f(x_1, x_2, \ldots, x_n)$ is the joint p.d.f. and $f_{x_i}(x_i)$ the marginal for $X_i$.

Expectation of a continuous r.v. is also similar to expectation in the discrete case, except that rather than summing $f(x) x$ over all possible $x$ values, we now integrate $f(x) x$ over all possible $x$ values. So, if $X$ is a r.v. with p.d.f. $f(x)$:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

To try and ensure that it’s clear why this is a sensible definition of the average of a continuous r.v. consider the following graphical example. Plot (a) shows a p.d.f., $f(x)$. In plot (b) this p.d.f. is approximated by the tops of $n$ strips of equal width $\epsilon$ where the centres of the strips are at $x_1, x_2, \ldots, x_n$, and the height of the $i^{th}$ strip is $f(x_i)$. Plot (c) illustrates the obvious fact that by increasing $n$ and correspondingly decreasing $\epsilon$ we get a better approximation to $f(x)$.

![Graphical Example](image)

Now, the approximate probability of $X$ being in interval $(x_i - \epsilon/2, x_i + \epsilon/2)$ is $f(x_i) \epsilon$. Which immediately suggests a way of approximating the average value of $X$:

$$E(X) \approx \sum_{i=1}^{n} \Pr[X \text{ is approximately } x_i] x_i = \sum_{i=1}^{n} f(x_i) \epsilon x_i$$

That is, we are approximating the value of every $X$ in the interval $(x_i - \epsilon/2, x_i + \epsilon/2)$ by $x_i$, and then using the definition of expectation of a discrete r.v. Obviously this approximation will get better and better as $n \to \infty$ and $\epsilon \to 0$.

---

**Note that the integration range here is $-\infty$ to $\infty$. This sometimes causes confusion, when random variables have zero probability of being outside some smaller range of values. The confusion arises because in this latter circumstance it’s common to be a little sloppy and only specify $f(x)$ over the range were it is non-zero, leaving it as implicit that it is zero everywhere else. All this is means is that you get to be a little careful when integrating p.d.f.s to make sure that you split the integral into its zero and non-zero parts.**
Now consider \( \sum_i f(x_i) x_i \epsilon \): as illustrated by (d) below, this amounts to summing up the areas of a set of rectangles of width \( \epsilon \) and height \( f(x_i) x_i \). (e) illustrates that as \( n \) increases and \( \epsilon \) decreases (the tops of) this set of rectangles get closer and closer to the curve \( f(x)x \), so that the sum of these rectangles gets ever closer to the area under the curve \( f(x)x \) illustrated in (f) . . .

In general if \( g(X) \) is a function of a continuous random function \( X \), with p.d.f. \( f(x) \) then:

\[
E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx
\]

Example: Suppose \( X \sim \exp(\lambda) \).

\[
E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{0} x \times 0 dx + \int_{0}^{\infty} x \lambda e^{-\lambda x} dx
\]

\[
= \left[-xe^{-\lambda x} + \int e^{-\lambda x} dx\right]_0^\infty
\]

\[
= \left[-xe^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \frac{1}{\lambda}
\]

Exercise: Let r.v. \( U \) have a uniform distribution between \( a \) an \( b \) so that it’s p.d.f is:

\[
f(u) = \frac{1}{b-a} \ \ a \leq u \leq b
\]

Find the mean of \( U \), \( E(U) \), in terms of \( a \) and \( b \):

### 3.4.4 Seven useful facts about expectations and variances

1. Consider r.v.s \( X \) and \( Y \) with joint p.d.f. \( f(x,y) \). Let \( g(X,Y) \) be any function of \( X \) and \( Y \):

\[
E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dx dy
\]

Example: suppose that \( g(X,Y) = X \) then:

\[
E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y) dx dy = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x,y) dy dx = \int_{-\infty}^{\infty} x f_x(x) dx \text{ (as expected)}
\]

Using this definition leads to an important result . . .

2. For any two r.v.s \( X \) and \( Y \):

\[
E(X + Y) = E(X) + E(Y)
\]

Proof: let \( f(x,y) \) be the joint p.d.f of \( X \) and \( Y \), then

\[
E(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x,y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) dx dy
\]

\[
= E(X) + E(Y)
\]
Notice that \( f(x, y) \) can be any joint p.d.f. and there are no assumptions of independence or any other special assumptions about the relationship between the r.v.s. The generalization to more than two r.v.s is obvious.

**Example:** If the height of husbands and wives in some population can be modelled as random variables \( H \) and \( W \) where the mean height of husbands is 1.78m and the mean height of wives is 1.62m, what is the mean combined height of a randomly chosen couple?

3. If \( X \) is a r.v. and \( a \) and \( b \) are (real) constants:

\[
E(aX + b) = aE(X) + b
\]

**Proof:** Let \( f(x) \) be the p.d.f. of \( x \).

\[
E(aX + b) = \int_{-\infty}^{\infty} (aX + b)f(x)dx = \int_{-\infty}^{\infty} axf(x)dx + \int_{-\infty}^{\infty} bf(x)dx = a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx = aE(X) + b
\]

**Example:** continuing the wife and husband example, what is the expected average height of a randomly chosen couple?

4. If \( X \) and \( Y \) are independent r.v.s then

\[
E(XY) = E(X)E(Y)
\]

the reverse implication does not hold.

**Proof:** let \( f(x, y) \), \( f_x(x) \) and \( f_y(y) \) be the joint and marginal p.d.f.s of \( X \) and \( Y \):

\[
E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dxdy
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x y f_x(x) f_y(y)dxdy \text{ by independence}
= \int_{-\infty}^{\infty} x f_x(x)dx \int_{-\infty}^{\infty} y f_y(y)dy
= E(X)E(Y)
\]

**Example:** You can use the above result to answer one of the following questions, identify which one, and answer it. (i) What is \( E(HW) \) in the husbands and wives example given above? (ii) Let \( R \) be rainfall in a day and \( T \) be mean daily temperature. At a particular site \( E(T) = 10^\circ C \), \( E(R) = 10 \text{ (mm) } \) and \( E(TR) = 100 \) — are \( R \) and \( T \) independent? (iii) Let \( L \) be proportion of left handers and \( A \) the number of traffic accidents per year per 10000 people for a set of small towns — if \( E(A) = 9 \) and \( E(L) = 0.3 \) what is \( E(AL) \)?

5. Recall that the definition of variance is \( \text{var}(X) = E[(X - \mu_x)^2] \) where \( \mu_x = E(X) \). \( \text{var}(X) \) measures how spread out the distribution of \( X \) is. Variance is a special case of **covariance**, which measures how related two random variables are and how spread out they are.
**Definition.** The **covariance** of two random variables $X$ and $Y$ with means $\mu_x$ and $\mu_y$ is:

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)] \quad (\equiv E(XY) - E(X)E(Y))$$

The term in brackets is almost always the one to use in calculations - the other version is more helpful for understanding what covariance is. It’s not hard to see why the two versions are equivalent:

$$E[(X - \mu_x)(Y - \mu_y)] = E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) = E(XY) - E(X)E(Y)$$

**Exercise:** If $X$ and $Y$ are independent, what is their covariance?

6. If $X$ and $Y$ are r.v.s then:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$

**Proof:**

$$\text{var}(X + Y) = E[(X + Y - \mu_x - \mu_y)^2]
= E[(X - \mu_x)^2 + (Y - \mu_y)^2 + 2(X - \mu_x)(Y - \mu_y)]
= E[(X - \mu_x)^2] + E[(Y - \mu_y)^2] + 2E[(X - \mu_x)(Y - \mu_y)]
= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$

**Exercise:** If $X$ and $Y$ are independent what is $\text{var}(X + Y)$?

7. If $X$ is a random variable and $a$ and $b$ are (real) constants:

$$\text{var}(aX + b) = a^2\text{var}(X)$$

**Proof:** writing $\mu_x$ for $E(X)$

$$\text{var}(aX + b) = E[(aX + b - a\mu_x - b)^2] = E[(aX - a\mu_x)^2]
= a^2E[(X - \mu_x)^2] = a^2\text{var}(X)$$

Note that 2 to 7 apply equally well to discrete r.v.s (although the proofs tend to be a bit messier).

### 3.5 The normal distribution

The “normal” or “Gaussian” distribution is the most important distribution in statistical theory, for reasons that we’ll discuss.

**Definition.** An r.v. $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$ if its p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty$$
$X \sim N(\mu, \sigma^2)$ means “random variable $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$.” The shape of the normal p.d.f. is:

![Normal Distribution Diagram]

Although its p.d.f. is a bit inconvenient, and there is no simple formula for its c.d.f., the normal distribution has some nice properties, and one astonishing one. Two nice properties are these:

1. If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ are two independent r.v.s then

   $$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

   i.e. sums of independent normal random variables are themselves normal random variables (with appropriate means and variances).

2. If $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. i.e. linear transformations of Normal random variables are also normal random variables.

You have already seen how to derive the means and variances for $X + Y$ and $aX + b$: what is new here is that we know the whole distribution of the new variables, not just their means and variances.

**Exercise:** If $X \sim N(\mu, \sigma^2)$, find the distribution of:

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

The normal distribution turns out to be quite a good model for all sorts of things. For example, heights of adult males or adult females, errors in physical measurements, logged differences between predicted and actual changes in animal populations, distance moved by a molecule of a tracer gas in still air after some set length of time, etc., etc. Part of the reason that the normal crops up so often is this:

**The Central Limit Theorem (CLT):** If $X_1, X_2, \ldots, X_n$ are independent identically distributed r.v.s from any distribution with finite mean, $E(X_i) = \mu$, and finite, non-zero variance $\text{var}(X_i) = \sigma^2$, then in the limit as $n \to \infty$:

$$\frac{1}{n}(X_1 + X_2 + \ldots + X_n) \sim N(\mu, \sigma^2/n)$$

In practical terms this is the same as saying that $X_1 + X_2 + \ldots + X_n$ tends to $N(n\mu, n\sigma^2)$ as $n$ increases. Proof of this result is beyond the scope of this course, but the following pictures provide an illustration produced by simulating. Each panel shows the results of repeatedly taking $n$ independent observations of identically distributed Poisson r.v.s, and finding the mean of each of these size $n$ sets. This process was repeated 50,000 times to get 50,000 size $n$ samples and hence 50,000 means for each sample size (each $n$). The bar charts show the frequency with which each value of the mean occurred for that sample size. Since the mean of 1 variable is just the variable itself, the $n = 1$ chart just shows the distribution of a Poisson random variable. The rest of the barcharts show how the distribution of the sample mean tends towards a normal distribution as $n$ increases.
The practical utility of the result is this: many statistical procedures involve summing or finding the average of sets of data, which are modelled as being observations of independent r.v.s. The CLT says that (almost) whatever the distribution of the r.v.s, their sum or their mean should approximately follow a normal distribution, provided only that there are enough of them. This is what makes the CLT the most important result in statistics - it enables us to find approximate distributions of quantities when finding exact distributions is difficult or impossible, and thereby allows us to calculate approximate probabilities in the common situation in which exact probabilities can not be obtained. Note that despite it’s amazing generality the CLT does not cover all r.v.s - there are some pathological distributions that don’t have finite variance (the $t_1$ distribution, for example) - sums of these don’t tend to the normal, but for practical work, such distributions are very rare indeed.

### 3.6 c.d.f.s, p.d.f.s and p.m.f.s in R

R has some useful functions built in for obtaining probabilities from the c.d.f. or p.m.f. of a random variable, and for evaluating the p.d.f. of r.v.s. The functions for evaluating a c.d.f. all start `p`, while the functions for the probability density/mass function start with the letter `d`. For example consider working out the probability that $Z \sim N(0,1)$ is less than -2 . . .

```r
> pnorm(-2,0,1) # 0 is the mean, 1 is s.d.
[1] 0.02275013
```

i.e. $Pr[Z < -2] = 0.02275013$. Similarly here is the command to find $Pr[X \leq 0.6]$ where $X \sim N(1,4)$:

```r
> pnorm(0.6,1,2)
[1] 0.4207403
```

If we want to know the height of the p.d.f. of $X$ at 0.6 which is **not** the probability that $X = 6$, the command is:

```r
> dnorm(0.6,1,2)
[1] 0.1955213
```

Here’s how to find $Pr[0.2 < Y \leq 0.6]$ where $Y \sim N(0,9)$, and $Pr[V > 0.5]$ where $V \sim N(1,1)$:

```r
> pnorm(0.6,0,3)-pnorm(0.2,0,3)
[1] 0.05268325
> 1-pnorm(0.5,1,1)
[1] 0.6914625
```

It’s also sometimes useful to find the value $a$ such that $Pr[X < a]$ is equal to some specified value. Functions starting `q` provide this facility. For example to work out the value $a$ such that $Pr[Z < a] = 0.975$ where $Z \sim N(0,1)$:
Similar functions are available for many distributions including Poisson \(( \text{dpois, ppois, qpois} \) and binomial \(( \text{dbinom etc.)}. For example to evaluate \( \Pr[Y = 2] \) where \( Y \sim \text{poi}(0.3) \):

\[
> \text{dpois}(2, 0.3) \\
[1] 0.03333682
\]

While to work out \( \Pr[Y \leq 2] \) (don’t forget that the difference between \( \leq \) and \(<\) is important for discrete r.v.s):

\[
> \text{ppois}(2, 0.3) \\
[1] 0.9964005
\]

Or less concisely:

\[
> \text{dpois}(0, 0.3) + \text{dpois}(1, 0.3) + \text{dpois}(2, 0.3) \\
[1] 0.9964005
\]

4  **Estimation and Hypothesis testing**

Suppose that you have some observations \( x_1, x_2, \ldots, x_n \) which can be modelled as observations of random variables \( X_1, X_2, \ldots, X_n \), where the properties of the \( X_i \)'s are described by a model, some of the parameters of which are unknown. This section is about what you can learn about these unknown parameters from the observations. This topic is of practical interest, because most data are of interest not in their own right, but for what they tell us about some larger population from which they come. In this circumstance the parameters of a statistical model can be thought of as characteristics of the population which we would like to infer things about using the sample.

For example, in the climate change case study, we sometimes viewed the global temperature record as being just one sample from the population of global temperature records that could have occurred if the sceptics model of global temperature is correct. The auto-regressive models that we built summarized how the population of global temperature records should behave under the sceptics model, and we estimated the parameters of these models using the single sample that we had (the real temperature record). Now consider an example for which the population is more concrete. If we want to estimate the average timber volume of trees in a large forestry plantation, then it might be reasonable to model timber volumes of trees as a normally distributed random variable with unknown mean and variance. To estimate these unknown parameters it would usually be impractical to measure every tree in the plantation, rather a random sample of trees would be taken, and the mean and variance parameters estimated from this sample. When dealing with practical modelling problems it is usually a good idea to ask yourself “what is the population that these data are a (random) sample from”.

The process of using models to learn about populations from samples is known as **statistical inference**. This section of the course covers the 3 main types of method used in statistical inference.

1. **Point estimation** is about estimating the values of model population parameters using samples of data. There are 3 main ways of doing this, and some theory for judging what constitutes a good estimate.

2. **Interval estimation** is about using samples of data to obtain intervals that have some specified probability of including a model population parameter of interest. These combine information about the value we think a parameter has, with some idea of how certain we are about that value.

3. **Hypothesis testing** provides a means for asking whether the data are consistent with a parameter (or set of parameters) taking a particular pre-specified value (or set of values). This turns out to be very useful for comparing alternative models of data.
4.1 Point Estimation

This section is about estimating model parameters using data. We will be interested in obtaining just a single “best estimate” for each parameter. The section is therefore called “point estimation” to distinguish it from the case in which we would like to estimate a range of “plausible” values for each parameter, which is known as “interval estimation”.

In simple situations there are often “obvious” ways of estimating parameters. For example, consider a set of data $x_1, x_2, \ldots, x_n$ which can be modelled as being independent observations of $X \sim N(\mu, \sigma^2)$, where $\mu$ and $\sigma^2$ are unknown parameters. Since $\mu$ is the mean of $X$ and $\sigma^2$ is the variance of $X$, then the “obvious” way to estimate these quantities would be to use the sample mean and sample variance of the $x_i$’s. i.e.

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

However, in more complicated situations the model parameters do not often have such convenient interpretations, so that it may be difficult or impossible to obtain such “obvious” estimates. In any case it is more satisfactory to have some general methods that can be used to estimate parameters for a wide range of models, rather than relying on working out an estimation method for each model that we encounter. General methods also have the advantage that once their properties have been derived in general they apply to each particular model to which they are applied: this saves work, and means that we only need to understand and remember a little bit of general theory, rather than a mass of details relating to a large number of different models.

The rest of this section examines 3 general methods of point estimation.

4.2 The Method of Moments (MOM)

Let $x_1, x_2, \ldots, x_n$ be modelled as observations of a random variable $X$ the distribution of which is known except for the values of some parameters.

**Definition.** The $j^{th}$ moment of random variable $X$ is $E(X^j)$.

**Definition.** The $j^{th}$ sample moment of the sample of data $x_1, x_2, \ldots, x_n$ is $(1/n) \sum_{i=1}^{n} x_i^j$.

The moments are a way of characterising $X$’s distribution, and the method of moments is a way of estimating the unknown parameters of $X$’s distribution using the following simple idea:

If the parameters of $X$’s distribution are estimated correctly then moments of $X$ according to the model distribution should be close to the sample moments of the $x_i$’s.

To use this idea in practice we proceed as follows:

1. If there are $p$ parameters to estimate then find expressions for the first $p$ moments of $X$ (i.e. $E(X)$, $E(X^2)$, \ldots, $E(X^p)$), in terms of these unknown parameters.
2. Calculate the first $p$ sample moments using the data $x_1, x_2, \ldots, x_n$.
3. Equate the expressions for the first $p$ moments of $X$ to the calculated sample moments, to obtain $p$ simultaneous equations.
4. Solve the simultaneous equations to obtain the “method of moments” estimates of the parameters.

In practice this is usually very simple, as the following examples show.

*Note that the following two statements are exactly equivalent:

$x_1, x_2, \ldots, x_n$ are independent observations of r.v. $X$ with p.d.f. $f(x)$.
and

$x_1, x_2, \ldots, x_n$ are observations of independent identically distributed (i.i.d.) r.v.s $X_1, X_2, \ldots, X_n$ with p.d.f. $f(x)$.

They will be used interchangeably.
Suppose that you have data $u_1, u_2, \ldots, u_n$ which you want to model as observations of a random variable $U \sim \text{exp}(\lambda)$, where the parameter $\lambda$ is unknown and has to be estimated.

1. There is one parameter to estimate ($\lambda$), so we need only find the first moment of $U$ in terms of $\lambda$. From section 3.4.3: $E(U) = 1/\lambda$.

2. The first sample moment is $\sum_{i=1}^{n} u_i/n$.

3. Equating the first sample moment and the first moment of $U$ gives the equation to solve in order to find the estimate $\hat{\lambda}$ of $\lambda$:

   $$ \frac{1}{\lambda} = \frac{1}{n} \sum_{i=1}^{n} u_i $$

4. Solving the equation gives:

   $$ \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} u_i} $$

The following figure plots a histogram of a sample of 100 $u_i$’s on the left. The MOM estimate of $\lambda$ for these data was 0.98. The second plot shows the p.d.f. that would result from a $\lambda$ of roughly twice that value, the third plot is the p.d.f. corresponding to the MOM estimated $\lambda$, while the final plot shows the p.d.f. for a $\lambda$ value that is approximately half the MOM estimate. Clearly the MOM estimated p.d.f. provides quite a good match to the shape of distribution suggested by the histogram.

Now suppose that you have data $x_1, x_2, \ldots, x_n$ that can be modelled as observations of random variable $X \sim N(\mu, \sigma^2)$ where $\mu$ and $\sigma^2$ are unknown parameters that need to be estimated from the data.

1. In this case there are two parameters to be estimated, so you need to find the first two moments of $X$ in terms of these parameters. The first moment is easy:

   $$ E(X) = \mu $$

To find the second moment $E(X^2)$, note that $\text{var}(X) = E(X^2) - [E(X)]^2 \Rightarrow E(X^2) = \text{var}(X) + [E(X)]^2$. We know that $\sigma^2 = \text{var}(X)$, for a normal r.v., so:

$$ E(X^2) = \sigma^2 + \mu^2 $$

2. The first and second sample moments are: $\sum_{i=1}^{n} x_i/n$ and $\sum_{i=1}^{n} x_i^2/n$.

3. So equating moments of $X$ to the sample moments of the $x_i$’s yields the two equations that must be solved to find the parameter estimates $\hat{\sigma}^2$ and $\hat{\mu}$:

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

and

$$ \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 $$
4. The first equation already gives $\hat{\mu}$, while substituting the first equation into the second gives:

$$
\hat{\sigma}^2 + \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n^2} \left( \sum_{i=1}^{n} x_i \right)^2
$$

- Now suppose that you have a observations: 1, 3, 0, 7 which can be modelled as observations of a random variable $V$ with distribution:

$$
f(v) = \begin{cases} 
1/\alpha & 0 < v < \alpha \\
0 & \text{otherwise}
\end{cases}
$$

where $\alpha$ is an unknown parameter. Use the method of moments to estimate $\alpha$ from the data.

1. There is one unknown parameter, so the first moment of $V$ must be evaluated in terms of that parameter:

$$
E(V) = \int_{-\infty}^{\infty} f(v) dv = \int_{0}^{\alpha} \frac{v}{\alpha} dv = 
$$

Hopefully, what these examples have illustrated is the way in which the method of moments provides a general and systematic approach to estimating model parameters in order to make the model match the data well. The method is often used for simple problems and in situations were other methods may be cumbersome or difficult to use. However it lacks the general theoretical results that lend to support to the other two methods discussed in the next two sections.

### 4.3 Least Squares

The method of least squares was introduced in the climate change case study. Like the method of moments it is based on the idea that the best estimates of model parameters will be those that make the model best match the data being modelled. The difference between least squares and the method of moments is in the way in which “best match” is judged. Consider a set of observations $x_1, x_2, \ldots, x_n$ which can be modelled as observations of r.v.s $X_1, X_2, \ldots, X_n$ the distributions of which have some unknown parameters. The method of least squares is based on the following idea:

**Model parameters should be chosen to minimise the sum of squared differences between each observation $x_i$ and its expected value according to the model: $E(X_i)$.**

i.e. the parameters are chosen to

$$
\text{minimise } S = \sum_{i=1}^{n} (x_i - E[X_i])^2
$$

w.r.t. the unknown model parameters. This is intuitively reasonable: a good model ought to be capable of producing $E(X_i)$’s that are close to the data, $x_i$’s, whereas a poor model will not be.

To see how the method works consider some examples:
1. First consider the case of estimating the parameters of a normal model, from independent observations \(x_1, x_2, \ldots, x_n\) of r.v. \(X \sim N(\mu, \sigma^2)\). \(E(X) = \mu\), so we seek to:

\[
\text{minimise } S = \sum_{i=1}^{n} (x_i - \mu)^2
\]

w.r.t. the unknown parameter \(\mu\). Note that since \(E(X)\) does not depend on \(\sigma^2\) it will not be possible to estimate \(\sigma^2\) using this method. \(S\) is a quadratic equation in \(\mu\). For example, if \(n = 5\) and the observations \(x_1, \ldots, x_5\) are 1.2, 3.4, 2.1, 2.3, 2.7 then a plot† of \(S\) against \(\mu\) looks like this:

![Plot of S against \(\mu\)](image)

--- from which it is possible to read off the least squares estimate of \(\mu\) for the 5 data given: \(\hat{\mu} \approx 2.3\).

Generally it is better to take a mathematical approach to minimising \(S\) w.r.t. \(\mu\). To find the turning point of \(S\) it must be differentiated‡ w.r.t. \(\mu\):

\[
\frac{dS}{d\mu} = \sum_{i=1}^{n} -2(x_i - \mu)
\]

and the result set to zero in order to obtain an equation to solve for \(\hat{\mu}\):

\[
\sum_{i=1}^{n} -2(x_i - \hat{\mu}) = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\hat{\mu} \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}
\]

This estimate is the same as the one that we obtained using the method of moments, and is also the “obvious” estimate. Applying the formula to the 5 measurements used to generate the plot \(S\) against \(\mu\) shows that in that case \(\hat{\mu} = 2.34\).

2. Now consider an example that would be quite difficult to approach using the method of moments. Suppose that we have data \(y_1, y_2, \ldots, y_n\) and some related data \(x_1, x_2, \ldots, x_n\) and that we have a model that says that the \(y_i\) are observations of random variables \(Y_i \sim N(\beta x_i, \sigma^2)\), where \(\beta\) (and \(\sigma^2\)) are unknown parameters. i.e. the model says that the expected value of \(Y_i\) is related to the value of the variable \(x_i\). As in the previous case, \(\sigma^2\) can not be estimated by least squares, but the parameter \(\beta\) can be.

According to the model \(E(Y_i) = \beta x_i\), so minimising:

\[
S = \sum_{i=1}^{n} (y_i - E(Y_i))^2
\]

†Produced using the following R code, incidentally:

```r
> x<-c(1.2, 3.4, 2.1, 2.3, 2.7)
> mu<-0:100/20;S<-0
> S<-0;for (i in 1:101) S[i]<-sum((x-mu[i])^2)
> plot(mu,S,xlab=expression(mu),type="l")
```

‡Note that

\[
\frac{\partial \sum_{i} x_i}{\partial y} = \sum_{i} \frac{\partial x_i}{\partial y} \quad \text{since:} \quad \frac{\partial \sum_{i} x_i}{\partial y} = \frac{\partial x_1}{\partial y} + \frac{\partial x_2}{\partial y} + \cdots + \frac{\partial x_n}{\partial y} = \sum_{i} \frac{\partial x_i}{\partial y}
\]
w.r.t. $\beta$ to find $\hat{\beta}$ is equivalent to minimising:

$$S = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

w.r.t. $\beta$. Again $S$ is quadratic in the unknown parameter. Given a particular set of $x_i, y_i$ values $S$ could be plotted against $\beta$ to illustrate this quadratic dependence, and $\hat{\beta}$ could be found graphically. Again, to find the exact least squares estimate of $\beta$ we seek the location of the minimum of $S$. $S$ is differentiated w.r.t. $\beta$:

$$\frac{\partial S}{\partial \beta} = \sum_{i=1}^{n} -2x_i(y_i - \hat{\beta}x_i)$$

and the result set to zero:

$$\sum_{i=1}^{n} -2x_i(y_i - \hat{\beta}x_i) = 0$$

to obtain the equation that must be solved to find $\hat{\beta}$:

$$\sum_{i=1}^{n} -2x_i(y_i - \hat{\beta}x_i) = 0 \Rightarrow \sum_{i=1}^{n} x_i y_i = \hat{\beta} \sum_{i=1}^{n} x_i^2 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Using least squares with this sort of model will be covered in much more detail later in the course.

3. Consider estimating $\lambda$ from independent observations $u_1, u_2, \ldots, u_n$ of r.v. $U \sim \exp(\lambda)$. In this case $E(U) = 1/\lambda$, so to estimate $\lambda$ from the data we seek to

minimise $S = \sum_{i=1}^{n} (u_i - 1/\lambda)^2$ \ w.r.t. $\lambda$

$S$ is not quadratic in $\lambda$ in this case, but this will not change the basic approach to finding $\hat{\lambda}$. First differentiate $S$ w.r.t. $\lambda$:

$$\text{then set } \frac{dS}{d\lambda} \text{ to zero and solve the resulting equation for } \hat{\lambda}:$$

Least squares will be covered in more detail later, as it is the best method for estimation in the context of linear regression modelling. Note that it can be used to estimate several parameters simultaneously: to do so involves finding the partial derivatives of the $S$ w.r.t. to each parameter and setting all of these to zero. Solving the resulting set of simultaneous equations gives the least square parameter estimates.
4.4 Maximum Likelihood estimation

Like the other two methods the maximum likelihood method is based on a very simple idea:

**Unknown model parameters should be chosen to maximise the probability of the observed data according to the model.**

The idea is that parameter values that make what actually happened appear probable are more “likely” than parameter values that make it look improbable. Specifically, if you have observations of random variables $X_1, X_2 \ldots X_n$ and can write down their joint p.d.f. or joint p.m.f., $f(x_1, x_2, \ldots, x_n)$, then you simply plug the observed values into $f(\cdot)$ and find the parameter values that maximise the resulting expression. Joint p.d.f.s or p.m.f.s with the data plugged in in this way are known as “likelihoods” of the parameters. Again, the easiest way to see how it works is to apply the idea to some examples.

Suppose that we have $n$ data $\tilde{x}_i$ which we model as observations of independent random variables $X_i$ with p.d.f.

$$f(x_i) = \begin{cases} 
(\beta + 1)x_i^\beta & 0 < x_i < 1 \\
0 & \text{otherwise}
\end{cases}$$

where $\beta$ is an unknown parameter to be estimated from the data, and the p.d.f. is shown here for the case $\beta = 2$:

![Graph showing the probability density function for $\beta = 2$]

Since the $X_i$’s are independent, the results of section (3.4.3) (see also section 3.2.1) can be used directly to obtain the joint p.d.f. of the $X_i$’s as the product\(^9\) of their marginal p.d.f.’s:

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} (\beta + 1)x_i^\beta \quad 0 < x_i < 1 \text{ for all } i$$

Plugging the observations, $\tilde{x}_i$, (of $X_i$) into this gives a result that indicates how probable observations are according to the model. Values of $\beta$ are not likely to be correct if they cause the model to suggest that what actually happened is improbable. Values of $\beta$ which cause the observed data to be probable under the model are much more likely to be right. So we can find the “most likely” value $\beta$: the one that maximises $f(\cdot)$ with the data plugged in.

The joint p.d.f., $f(\cdot)$, with the observed data plugged in and considered as a function of $\beta$ is called the **likelihood** of $\beta$:

$$L(\beta) = \prod_{i=1}^{n} (\beta + 1)\tilde{x}_i^\beta$$

\(\ldots L\) should be large for likely values of $\beta$ (values that make the observed data appear probable under the model), and small for unlikely values of $\beta$ (values that make the observed data improbable according to

\(^9\)Note the following definition of notation:

$$\prod_{i=1}^{n} v_i \equiv v_1 \times v_2 \times \ldots \times v_n$$
the model). The most likely value of $\beta$ is the one that maximises $L$. Notice that I have been quite careful to distinguish the dummy variables, $x_i$ that are arguments of the joint p.d.f. and the actual observed values $\tilde{x}_i$. Most textbooks are not so careful, and just write $x_i$ for both: once you are used to likelihood this doesn’t cause confusion.

To make this less abstract, suppose that we have 10 observations ($\tilde{x}_i$’s): 0.68, 0.75, 0.96, 0.90, 0.96, 0.62, 0.95, 0.98, 0.73, 0.91. Then we can plot $L$ as a function of $\beta$:

Use the graph to obtain an approximate maximum likelihood estimate for $\beta$:

Of course, we don’t have to resort to plotting $L$, but can find the $\beta$ that maximises $L$ mathematically. The process is made easier if we note that $\log(L)$ will be maximised by the same value of $\beta$ that maximises $L$ (whichever $\beta$ gives the biggest $L$ value, must automatically give the biggest $\log(L)$ value). The following plot of $\log(L)$ against $\beta$ illustrates this:

... the shape of the plot is very different, but the maximum is at the same value of $\beta$. Define $l \equiv \log(L)$. Repeated application of the rules $\log(AB) = \log(A) + \log(B)$, and $\log(A^b) = b \log(A)$ yields:

$$l(\beta) = \log(L) = \sum_{i=1}^{n} \log(\beta + 1) + \beta \log(x_i) = n \log(\beta + 1) + \beta \sum_{i=1}^{n} \log(x_i)$$

To find the maximum of $l$ w.r.t. $\beta$ we need to find the value of $\beta$ at which $dl/d\beta = 0$.

$$\frac{dl}{d\beta} = \frac{n}{\beta + 1} + \sum_{i=1}^{n} \log(x_i)$$

and setting this to 0 implies that:

$$\hat{\beta} = -\frac{n}{\sum \log(x_i)} - 1$$

---

*Which generalizes to:

$$\log \left( \prod_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} \log(x_i)$$
Use this expression to obtain the maximum likelihood estimate of $\beta$ from the 10 data given above:

$$f(x_i) = \begin{cases} 
\lambda e^{-\lambda x_i} & x_i \geq 0 \\
0 & x_i < 0 
\end{cases}$$

Suppose that you have observations of the time in seconds between 6 calls, 1.48, 0.13, 0.42, 0.39, 0.38, 0.07. Use these to obtain a maximum likelihood estimate of $\lambda$. (First form $L$ then $\log(L)$ and then find the maximum w.r.t. $\lambda$ — it’s usually best to only plug in the actual numbers right at the end)

Having worked through the mechanics of maximum likelihood estimation, it’s worth revisiting what it’s trying to do once more. The example that you have just done provides a useful way of illustrating this. The following 3 plots show the p.d.f. of the exponential for 3 different $\lambda$ values, with the 6 data that you have just used shown on the x axis.

On the left $\lambda$ is set very low, which means that all the data have rather low associated probability densities - the data don’t seem all that probable under the model relative to the next panel. In the middle panel the 5 fairly low valued data get quite high probability densities while the remaining point still gets assigned a reasonable probability density, so overall the data look quite probable under the model, suggesting that this $\lambda$ is quite likely. In the final panel $\lambda$ is too high: the lower 5 values still have quite high probability densities, but now the probability density of the final datum is close to zero, which makes the data set look rather improbable according to the model. The mathematics you have just been through is really about formalizing this approach.

Maximum likelihood estimation works in exactly the same way when we have several unknown parameters to estimate. The only difference being that we have to find the maximum of the likelihood w.r.t. several parameters, rather than just one. It also works just as well for discrete data as continuous data: all that changes is that p.m.f.s take the place of p.d.f.s.
4.5 Bias and variance in estimation

You have now met the three alternative general approaches to estimating model parameters from data. Which one should be used in any given situation? Sometimes the answer depends on considerations of practicality or convenience, but it is also possible to develop some theory for comparing estimates. To do so it is helpful to draw a distinction between an estimate and an estimator. An estimator is a function of the random variables that the data are observations of, and defines the recipe for generating estimates from data. Since the estimator is a function of random variables, it is itself a random variable, and we can discuss its mean, variance, distribution and other properties. An estimate is a particular observation of an estimator.

For example, suppose that we have \( n \) observations \( x_i \) of random variables \( X_i \sim N(\mu, \sigma^2) \) an estimator of \( \mu \) is:

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{estimator}
\]

You get a particular estimate of \( \mu \) by plugging the particular observations \( x_i \) into the estimator in place of the \( X_i \)'s:

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{estimate}
\]

e.g. if the \( n = 3 \), \( x_1 = 2.5 \), \( x_2 = 3.1 \) and \( x_3 = 3.4 \) then the estimate is 3.

Estimators allow us to study how we would expect estimates to behave if we calculated them repeatedly using different replicate sets of data.

4.5.1 Bias

Since an estimator is a random variable, with a distribution, we can evaluate its expected value. For example 1 we have that:

\[
E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n} n \mu = \mu.
\]

So the average value of the estimator in this case is the value of the thing being estimated. This is quite a nice property. It means that if we take repeated samples and apply the estimator to each, then the average of the resulting estimates will tend towards the population value of \( \mu \) as we increase the number of reps.

**Definition.** An estimator \( \hat{\beta} \) of a parameter \( \beta \) is said to be **unbiased** if:

\[
E(\hat{\beta}) = \beta
\]

otherwise the estimator is **biased**.

Intuitively, an unbiased estimator is one that “gets it right on average”. Clearly \( \hat{\mu} \), given above, is an unbiased estimator of \( \mu \) (actually the sample mean is an unbiased estimator of the population mean for any model, not just a normal distribution).

We note, however, that quite a few perfectly defensible estimators are biased. For example, the MOM estimate of \( \sigma^2 \) for a normal population was:

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

but it’s not too hard to show that this estimator is biased\(^\dagger\):

\[
E(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2
\]

\(^\dagger\)Like with the sample mean, this result also applies to estimating the variance of any r.v., not just a normal one.
i.e. on average the “obvious” estimator slightly underestimates $\sigma^2$. In this case it’s very easy to correct the biased estimator to get an unbiased one:

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

is easily seen to be unbiased, and is therefore generally the preferred estimator of the variance of a population.

### 4.5.2 Estimator variance

Unbiasedness only indicates that the expected value of an estimator is correct, it doesn’t indicate how close any particular estimate will be to the true value of the parameter of interest. In order to judge accuracy of an estimator we need to know how variable it is: we need to know its variance. Estimates that come from an unbiased estimator with small variance can be expected to be quite close to the truth, while estimates from an unbiased estimator with large variance are obviously less reliable. The figure below illustrates this graphically, using two estimators of the mean of a normal population for which the true mean is 2:

$\hat{\mu}_1$ is an unbiased estimator with high variance (it’s actually the average of the smallest and largest observation in the sample). $\hat{\mu}_2$ is an unbiased estimator with small variance (it’s the mean of all the data in the sample). Each point in both plots shows the estimate of the mean obtained by using the respective estimators on a sample from a normal population.

Returning again to the problem of estimating the mean, $\mu$, of a normal distribution from a sample of data we seek:

$$\text{var}(\hat{\mu}) = \text{var} \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

so the estimator becomes more accurate as the number of observations $n$ increases.

It is very important to remember that the variance relates to the estimator and not the parameter being estimated: the parameter being estimated is a fixed quantity, not a random variable.

So, in general we would like to have estimators that are unbiased and have low variance. But when searching for estimators, how low is it possible to get the variance? It turns out that there is a lower limit to the variance of unbiased estimators of any parameter: it is simply not possible to find an estimator with a smaller variance than this lower limit. The lower limit, known as the Cramer-Rao lower bound can be slightly complicated to work out in any given case, so we won’t give the formula here (see honours courses if you’re interested), but it does mean that there is a “gold standard” against which to judge estimators.

Do these considerations help in choosing between the three different estimation approaches covered in this section? To some extent yes. Maximum likelihood estimators turn out to be very appealing, because for almost any estimation problem they give estimators that tend rapidly towards un-biasedness and towards the Cramer- Rao lower bound as sample sizes increase. Maximum likelihood is also the method that works in the widest variety of circumstances. Least squares turns out to be unbiased and to produce...
the estimators with lowest possible variance for a large class of models called linear models, which we will return to later. The method of moments has less theoretical support, but is often the simplest approach to use.

4.6 Confidence Intervals

The problem with point estimators is that we can almost always be certain that they are wrong, in the sense that they don’t exactly equal the thing that they are trying to estimate. It would be nice to have estimators that we could be fairly confident are right, in some sense. The only way to achieve this is to move away from point estimates of the precise value of a parameter, to give instead a range of values which we can be quite confident will contain the true parameter. To make things more concrete, suppose that we have a population described by a model with parameter \( \beta \), which is to be estimated using a random sample from the population. Previously we saw how a sample of data can be used to provide a single estimate of the value of a parameter. Now we want to use the data to construct an interval which is very likely to include the true value of the parameter. Such intervals are termed confidence intervals.

**Definition.** A 100\( \alpha \)% confidence interval for a parameter \( \beta \) is a random interval that has a probability \( \alpha \) of including \( \beta \).

Note the crucial point that the interval will be calculated from a sample of observations of random variables and is therefore a random interval, in that it will differ between replicate samples. The parameter \( \beta \) on the other hand is a fixed property of the population (strictly of the model of the population). Hence the probabilities are associated with the interval and not the parameter.

The following figure shows 20 confidence intervals (vertical lines) for the mean of a normal population, calculated from 20 replicate samples (each of size 20) from the same population.

![Confidence Intervals](image)

The figure shows clearly that while the parameter \( \mu \) is fixed, the confidence intervals for \( \mu \) vary from sample to sample. However, since these are 95% CI’s we expect (on average) that 19 out of 20 of them will include the true value of \( \mu \) while 1 out of 20 will not - this is exactly what has happened in this example. Of course in any real situation we will have just one interval, but it is important to know what is meant by stating a specified degree of confidence in it.

Having defined the properties that a confidence interval ought to have, how can confidence intervals actually be calculated? Unfortunately, there is not a simple universally applicable answer, but we can get quite a long way towards a useful theory for interval estimation. First consider the situation in which you have a model with an unknown parameter \( \beta \) and an estimator of it, \( \hat{\beta} \). Suppose for the moment that \( \beta \) is known: then in principle the distribution of \( \hat{\beta} \) could be worked out (never mind the details of how, for the moment). Given this distribution it is possible to obtain an interval within which \( \beta \) will lie with specified probability, \( \alpha \), say: this interval can be used to construct another interval around \( \hat{\beta} \) which is a random interval with probability \( \alpha \) of including \( \beta \). The following figure illustrates exactly how this works for a 95% CI:
Work out distribution of $\hat{\beta}$
given model parameter $\beta$

Find interval containing
95% of $\hat{\beta}$'s.

Define 2 sub-intervals
and

Create an interval from below $\hat{\beta}$ to above $\hat{\beta}$:

Since $\hat{\beta}$ is a random variable, $\beta$ is a random interval.

Now, this approach would be fine, but for the obvious problem that we do not know $\beta$ and hence we can not work out the distribution of $\hat{\beta}$, exactly. In some circumstance we may be able to show that the length of the interval does not depend on $\beta$ and an interval can be calculated, but generally the length of the interval will depend on $\beta$. There are two possibilities for proceeding. The first is that we replace $\beta$ by $\hat{\beta}$ when working out how wide the interval should be: this can be a quite reasonable approximation if the sample size used for estimation is large, and such intervals are quite often used in modern computer intensive statistics. Unfortunately, small samples (less than 30, say, for a one parameter problem) are much more problematic, in that the estimator is itself likely to be so variable, that there is a serious danger of making the confidence interval much too short by chance, and thereby substantially reducing the probability of the interval including the truth.

4.6.1 How to obtain CI's in practice: an example

Consider the case in which you have an unbiased estimator $\hat{\beta}$ of a parameter $\beta$, and the variance of $\hat{\beta}$ is known to be $\sigma^2_{\hat{\beta}}$. Further suppose that the estimator has a normal distribution, so that:

$$\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}}).$$

Now consider the transformed r.v.

$$Z = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}}$$

In section 3.5 of the notes you showed that $Z \sim N(0, 1)$. It is straightforward to find an interval within which $Z$ lies with any specified probability, and then to derive from this an interval that will contain $\beta$ with the specified probability. Consider the case of a 95% CI:

$$\Pr[-1.96 \leq Z \leq 1.96] = 0.95$$

$$\Rightarrow \Pr[-1.96 \leq (\hat{\beta} - \beta)/\sigma_{\hat{\beta}} \leq 1.96] = 0.95$$

$$\Rightarrow \Pr[-1.96\sigma_{\hat{\beta}} \leq \hat{\beta} - \beta \leq 1.96\sigma_{\hat{\beta}}] = 0.95$$

$$\Rightarrow \Pr[-\hat{\beta} - 1.96\sigma_{\hat{\beta}} \leq -\beta \leq -\hat{\beta} + 1.96\sigma_{\hat{\beta}}] = 0.95$$

$$\Rightarrow \Pr[\beta - 1.96\sigma_{\hat{\beta}} \leq \beta \leq \beta + 1.96\sigma_{\hat{\beta}}] = 0.95$$

42
so, there is a 95% chance of the random interval: \((\hat{\beta} - 1.96\sigma_{\hat{\beta}}, \hat{\beta} + 1.96\sigma_{\hat{\beta}})\) containing \(\beta\): we have derived a 95% confidence interval for \(\beta\).

Note that the number 1.96 is the number which \(Z \sim N(0, 1)\) exceeds with probability 0.025, i.e. \(\Pr[Z > 1.96] = 0.025\). By the symmetry of the normal distribution it is also the case that \(\Pr[Z < -1.96] = 0.025\). As shown in section 3.6, such numbers can be found in \(R\) using the \texttt{qnorm()} function. For example:

\[
> \text{qnorm}(0.025, \text{mean}=0, \text{sd}=1)
\]

\[
[1] \quad 1.959964
\]

In general, let \(z_{\alpha/2}\) be the number such that \(\Pr[Z \leq -z_{\alpha/2}] = \alpha/2\) where \(Z \sim N(0, 1)\) (\(z_{\alpha/2}\) is known as the \(\alpha/2\) critical point of the \(N(0, 1)\) distribution). If \(\beta\) is a parameter with estimator \(\hat{\beta} \sim N(\beta, \sigma_{\hat{\beta}}^2)\) and \(\sigma_{\hat{\beta}}^2\) is known, then a 100\((1 - \alpha)\)% confidence interval for \(\beta\) is:

\[
(\hat{\beta} - z_{\alpha/2}\sigma_{\hat{\beta}}, \hat{\beta} + z_{\alpha/2}\sigma_{\hat{\beta}})
\]

This result is one that you should learn and know how to use.

**Example:** Suppose that we have data \(x_1, x_2, \ldots, x_n\) that can be modelled as observations of independent r.v.s \(X_i \sim N(\mu, \sigma^2)\), where \(\sigma^2\) is known, but \(\mu\) is not. An unbiased estimator of \(\mu\) is \(\hat{\mu} = \sum X_i/n \equiv \bar{x}\), and this estimator has variance \(\sigma_{\hat{\mu}}^2 = \sigma^2/n\) (see section 4.5.2). Furthermore since \(\hat{\mu}\) is a weighted sum of normal r.v.s it must follow a normal distribution (see section 3.5). i.e. \(\hat{\mu} \sim N(\mu, \sigma^2/n)\). Find a 99% CI for \(\mu\). Using the general result given above this will be:

\[
(\bar{x} - z_{0.005}\sigma/\sqrt{n}, \bar{x} + z_{0.005}\sigma/\sqrt{n})
\]

Using:

\[
> \text{qnorm}(0.005, \text{mean}=0, \text{sd}=1)
\]

\[
[1] \quad 2.575829
\]

we have that the 99% CI is:

\[
(\bar{x} - 2.575\sigma/\sqrt{n}, \bar{x} + 2.575\sigma/\sqrt{n})
\]

**Numerical example:** Suppose that in the above example \(n = 4\), \(\sigma = 1\) and the \(x_i\)'s are: 0.3, 0.5, 1.4 and 0.8. Find a 90% CI for \(\mu\) (your answer should be two actual numbers). One of the results in the following \(R\) snippet is useful:

\[
> \text{qnorm}(0.05, \text{mean}=0, \text{sd}=1)
\]

\[
[1] \quad -1.644854
\]

\[
> \text{qnorm}(0.025, \text{mean}=0, \text{sd}=1)
\]

\[
[1] \quad -1.959964
\]

### 4.6.2 Calculating CI’s in general: the ideal recipe

The example in the previous section is a very simple example of the statisticians “ideal recipe” for generating confidence intervals. Consider finding a 100\(\alpha\)% confidence interval for a parameter \(\beta\) of a population, from which you have a random sample of data.

1. Find a so called pivotal function of \(\beta\) and the data, with the property that the pivotal has a completely known distribution (i.e. a distribution with no unknown parameters). In the example in the previous section \(Z\) was such a pivotal quantity.

2. Find an interval within which the pivotal lies with probability \(\alpha\).

3. Manipulate the expression for the interval in 2 to get an interval for \(\beta\).

The key point is that if you can find a pivotal quantity, then you avoid the problems associated with trying to find confidence intervals directly from distributions which have unknown parameters. Having made this general recipe explicit the next section considers an example where it is immediately useful.
4.6.3 Confidence intervals when the variance is unknown

Now consider finding a confidence interval for a parameter $\beta$, given an estimator $\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}})$ where $\sigma^2_{\hat{\beta}}$ is not known. Suppose in this case that $\beta$ is one of $p$ unknown parameters (excluding $\sigma^2_{\hat{\beta}}$), and that the parameters have been estimated using $n$ data. If $\hat{\sigma}^2_{\hat{\beta}}$ is the unbiased estimator for $\sigma^2_{\hat{\beta}}$, then it turns out that:

$$ T = \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} $$

is a pivotal quantity following a so called $t$ - distribution**. Specifically:

$$ T = \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t_{n-p} $$

$t_{n-p}$ is “a $t$ distribution with $n - p$ degrees of freedom”. $t$ distributions are similar in shape to the $N(0,1)$ distribution, but with more probability in the tails of the distribution, and less in the middle. The following shows some $t$ distribution with different degrees of freedom and the $N(0,1)$ distribution. A $t_{\infty}$ distribution is an $N(0,1)$ distribution.

![Image of t and normal distributions](image)

To fix ideas, suppose that a 95% CI is required, $n = 20$ and we are only estimating $\beta$ and the variance from the data, so that $p = 1$. From the c.d.f. of the $t_{19}$ distribution it is easy to obtain an interval containing the pivotal quantity with probability 0.95, and this leads to the CI:

$$ \Pr[-2.09 \leq T \leq 2.09] = 0.95 $$

$$ \Rightarrow \Pr[-2.09 \leq (\hat{\beta} - \beta)/\hat{\sigma}_{\hat{\beta}} \leq 2.09] = 0.95 $$

$$ \Rightarrow \Pr[-2.09\hat{\sigma}_{\hat{\beta}} \leq \hat{\beta} - \beta \leq 2.09\hat{\sigma}_{\hat{\beta}}] = 0.95 $$

$$ \Rightarrow \Pr[-\hat{\beta} - 2.09\hat{\sigma}_{\hat{\beta}} \leq -\beta \leq -\hat{\beta} + 2.09\hat{\sigma}_{\hat{\beta}}] = 0.95 $$

$$ \Rightarrow \Pr[\hat{\beta} - 2.09\hat{\sigma}_{\hat{\beta}} \leq \beta \leq \hat{\beta} + 2.09\hat{\sigma}_{\hat{\beta}}] = 0.95 $$

i.e. a 95% CI in this case is:

$$ (\hat{\beta} - 2.09\hat{\sigma}_{\hat{\beta}}, \hat{\beta} + 2.09\hat{\sigma}_{\hat{\beta}}) $$

The following figure shows 20 of these intervals calculated from 20 replicate samples.

**There are some technical restrictions required for this result to be exact: the parameter estimates must be linear in the data, and $\sigma^2_{\hat{\beta}}$ must be estimated using an unbiased estimator of the data variance, from the model residual (see section 5). In practice, these restrictions are always met in the cases considered in this course.
95% confidence intervals for \( \mu \) from 20 samples of size 20 from \( N(1,1) \)

\( \sigma^2 \) assumed unknown for interval calculation

Notice how ingenious these intervals are: they are all different lengths and located in different places, but still have a probability of exactly 0.95 of including the truth.

**In general:** Let \( t_{n-p,\alpha/2} \) be the \( \alpha/2 \) critical point of \( t_{n-p} \) (i.e. \( \Pr[T \leq -t_{n-p,\alpha/2}] = \alpha/2 \) where \( T \sim t_{n-p} \)). Let \( \beta \) be a parameter with estimator \( \hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}}) \) where \( \sigma^2_{\hat{\beta}} \) is **unknown** and assume that \( \beta \) is one of \( p \) parameters being estimated using \( n \) data. A 100(1 - \( \alpha \))% confidence interval for \( \beta \) is:

\[
(\hat{\beta} - t_{n-p,\alpha/2} \hat{\sigma}_{\beta}, \hat{\beta} + t_{n-p,\alpha/2} \hat{\sigma}_{\beta})
\]

This result is one that you should learn and know how to use!

**Example:** Consider data \( x_1, x_2, \ldots, x_9 \) which can be modelled as observations of independent r.v.s \( X_i \sim N(\mu, \sigma^2) \), where \( \mu \) and \( \sigma \) are unknown. Suppose that the sample mean of these data is 0.7, while the sample standard deviation is 0.6, and a 98% CI for \( \mu \) is required. \( \mu \) can be estimated using the sample mean (i.e. \( \hat{\mu} = \sum x_i/9 \equiv \bar{x} \)), while \( \sigma \) can be estimated using the sample standard deviation (equivalently \( \hat{\sigma}^2 = \sum(x_i - \bar{x})^2/8 \)). Hence \( \hat{\sigma}_{\mu}^2 = \hat{\sigma}^2/9 \), and \( \hat{\sigma}_{\mu} = 0.6/3 = 0.2 \), while \( \hat{\mu} = 0.7 \). We also require the \( t_{8,0.01} \) critical points:

\[
> -qt(0.01,8) \\
\text{[1]} 2.896459
\]

So a 98% CI for \( \mu \) is:

\[
(\hat{\mu} - t_{8,0.01} \hat{\sigma}_{\mu}, \hat{\mu} + t_{8,0.01} \hat{\sigma}_{\mu}) = (0.7 - 2.896 \times 0.2, 0.7 + 2.896 \times 0.2) = (0.12, 1.28)
\]

**Exercise:** Fifteen tomato plants have their heights measured after a period of growth in a greenhouse. The mean height in the sample is 83 cm and the standard deviation of height in the sample is 5.8 cm. Find 95% confidence limits for the mean plant height in the population from which the random sample was drawn, assuming height to be normally distributed in the population. One part of the following R code is useful!

\[
> -qt(0.05,14) \\
\text{[1]} 1.76131
\]

\[
> -qt(0.05,13) \\
\text{[1]} 1.770933
\]

\[
> -qt(0.025,14) \\
\text{[1]} 2.144787
\]

\[
> -qt(0.025,13) \\
\text{[1]} 2.160369
\]
4.6.4 Discussion and approximations

The pivotal method produces excellent confidence intervals, but to use it we have to find a pivotal function, and in very many cases this is either very difficult or impossible. It might seem then, that our search for a way of finding confidence intervals in general has not been very successful. In fact this is over-pessimistic: the two cases that we have managed to do properly can be used to find approximate confidence intervals in all sorts of situations as a direct result of the central limit theorem. In particular, if you can treat a parameter estimator as being *approximately* normally distributed, then you can directly use one of the two results in this section to obtain an *approximate* confidence interval for the parameter. All you need in order to do this is the estimate of the parameter, the associated (estimated) standard deviation and the relevant critical points from the appropriate $t$ or $N(0,1)$ distribution.

4.7 Hypothesis testing

As you saw in the climate change case study and as we will see again in section 5, a common problem when constructing statistical models of data is that we have to decide which terms to include in a model, and which terms to leave out. It is good to keep models simple, because models with few parameters generally lead to more precise answers than models with many parameters, and because simple models are often easier to interpret than more complicated models. However, if we try to use models with fewer parameters than are necessary to describe the process that really generated the data then we can not expect to get meaningful results from our efforts.

Hypothesis testing is a mathematical approach for helping to decide which terms are necessary in a model, and which can be omitted. In these notes we will assume that omitting a term from a model is always equivalent to setting one (or more) model parameter(s) to zero. We will therefore only consider testing hypotheses about whether particular model parameters are zero or not. Our main concern will be to develop a meaningful measure of the evidence that a parameter is not zero, that will provide a standard scale of measurement for judging when to include and when to exclude terms from models.

4.7.1 p-values: measuring the evidence against a hypothesis

To fix ideas, consider the case of an unknown model parameter $\beta$, which can be estimated from data using an unbiased parameter estimator $\hat{\beta}$. To simplify matters, assume that the variance of $\hat{\beta}$ is known, and that $\hat{\beta} \sim N(\beta, \sigma_\beta^2)$. We would like to test whether the parameter $\beta$ is really necessary in this model or not. Formally we want to test the null hypothesis:

$$H_0 : \beta = 0$$

against the alternative (and more general) hypothesis:

$$H_1 : \beta \neq 0 \quad \text{(but unknown)}$$

How can we assess whether $H_0$ or $H_1$ is more appropriate? The obvious approach is to see whether $\hat{\beta}$ is closer to the $\beta$ value specified by $H_0$, or to that specified by $H_1$, but this is not satisfactory. Because $H_1$ is phrased more generally than $H_0$, $\hat{\beta}$ is always closer to $\beta$ consistent with $H_1$ than to the single value of $\beta$ consistent with $H_0$. The only exception to this is if $\hat{\beta} = 0$: but, even if $H_0$ is true, the probability that $\hat{\beta} = 0$ is vanishingly small or zero. Hence, if we want to avoid always selecting the more complicated model specified under $H_1$, we will have to devise a procedure that accepts $H_0$ by default and only rejects it in those cases in which the evidence points strongly towards $H_1$ being more appropriate.

A sensible procedure would be one in which $H_0$ is accepted when $\hat{\beta}$ is close to zero, but rejected in favour of $H_1$ in those cases in which $\hat{\beta}$ is too far from zero. But how can we judge whether $\hat{\beta}$ is close enough to zero or too far from zero? A usefully general method is based on asking the following question:

If $H_0$ were true, with what probability would the estimator $\hat{\beta}$ be at least as favourable to $H_1$ as the observed estimate $\beta$?
The answer to this question is known as the **p-value** of the test of $H_0$ against $H_1$. For example, the black filled in area on the following plot would be the p-value for the example used to motivate this section, if the observed $\hat{\beta}$ was the one shown:

![Plot showing p-value](image)

Try to answer these two questions about the p-value:

1. What would a **small** p-value lead you to conclude about which hypothesis to accept?
2. What would a **large** p-value lead you to conclude about which hypothesis to accept?

### 4.7.2 Specific example: testing the mean of a normal population

Suppose that we have data $x_1, x_2, \ldots, x_n$ which can be modelled as observations of independent r.v.s $X_i \sim N(\beta, \sigma^2)$, where $\sigma^2$ is known. Suppose that we would like to test:

$$H_0 : \beta = 0 \text{ against } H_1 : \beta \neq 0 \text{ but unknown.}$$

In this case $\hat{\beta} = \frac{1}{n} \sum X_i$ is an unbiased estimator. Since $\hat{\beta}$ is a sum of normal r.v.s it will also be normally distributed (see section 3.5), and we have already seen that $\text{var}(\hat{\beta}) = \sigma^2/n$. In summary $\hat{\beta} \sim N(\beta, \sigma^2/n)$.

We now need to calculate the p-value for this test. Assuming $H_0$ is true, the probability that the estimator $\hat{\beta}$ is at least as favourable to $H_1$ as the observed estimate $\hat{\beta}$ is:

$$p = \Pr(|B| \geq |\hat{\beta}|) \text{ where } B \sim N(0, \sigma^2/n)$$

By the symmetry of the normal distribution, this is readily seen to be:

$$p = 2 \times \Pr(B \leq -|\hat{\beta}|) \text{ where } B \sim N(0, \sigma^2/n)$$

and the required probability can readily be evaluated from the c.d.f. of $N(0, \sigma^2/n)$, using R, for example.

**Numerical Example:** Assuming that $0.63, 0.86, 3.12, 0.46, 2.05, 0.16, 2.04, 1.64, 0.09, 0.82$ are independent observations of $N(\beta, 1)$ r.v.s, let us test:

$$H_0 : \beta = 0 \text{ against } H_1 : \beta \neq 0 \text{ but unknown.}$$

First evaluate $\hat{\beta}$:

```r
> x <- c(0.63, 0.86, 3.12, 0.46, 2.05, 0.16, 2.04, 1.64, 0.09, 0.82)
> b.est <- mean(x); b.est # estimate of beta
[1] 1.187
```
Now we know that under the null hypothesis $\hat{\beta} \sim N(0, \sigma^2 = 0.1)$, so the p-value is readily evaluated:

```r
> p.value<-2*pnorm(-abs(b.est),mean=0,sd=0.1^0.5)
> p.value
[1] 0.0001742965
```

There appears to be quite strong evidence against the null hypothesis and for the alternative.

**Another example:** A single parameter model is fitted to 10 data with known variance. The single parameter is $\theta$ and its estimate $\hat{\theta}=-0.12$. The variance of the estimator is 0.01, and it is normally distributed. Calculate the p-value of the test $H_0: \theta = 0$ against $H_1: \theta \neq 0$ but unknown. using the following R output, only one part of which is relevant!

```r
> pnorm(0.12,mean=0,sd=0.1) > pnorm(-0.12,mean=0,sd=0.1/10^0.5)
[1] 0.8849303 [1] 7.390116e-05
> pnorm(0,mean=0.12,sd=0.1)
[1] 0.1150697
```

Recall that `pnorm(a,mean=b,sd=c)` return $\Pr[X \leq a]$ where $X \sim N(b, c^2)$.

### 4.7.3 More generality: what to do when variance is unknown

In the previous examples the distribution of $\hat{\beta}$ was completely known if $H_0$ was true, but this will not always be the case. For example, consider the case in which parameter $\beta$ is estimated from data using an unbiased estimator $\hat{\beta} \sim N(\beta, \sigma^2_\beta)$ where $\sigma^2_\beta$ is now unknown and must be estimated. This presents a problem: if we try to calculate a p-value using the estimated variance of $\hat{\beta}$ as if it were the true value we will obtain an underestimate as a result of neglecting the uncertainty in $\hat{\sigma}^2_\beta$. To address this problem we can try to find a test statistic, which can be used to distinguish between the hypotheses in the same way as $\hat{\beta}$ does, but which has a completely known distribution under $H_0$. For example, we could test

$$H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta \neq 0 \quad \text{but unknown.}$$

using the test statistic:

$$T = \frac{\hat{\beta}}{\hat{\sigma}_\beta}.$$ 

$T$ will clearly take values close to zero if $H_0$ is true, but will tend to have a larger magnitude if $H_1$ is true. It also turns out that under $H_0$ the distribution of $T$ is completely known. If $\beta$ is part of a model in which we must estimate $p$ parameters (not counting $\sigma^2_\beta$), from $n$ data then:

$$\hat{\beta}/\hat{\sigma}_\beta \sim t_{n-p} \quad \text{if} \quad H_0 \quad \text{is true}$$

where $t_{n-p}$ is a $t$ distribution with $n-p$ degrees of freedom\(^\dagger\). To find a p-value in this case we must evaluate:

The probability of the test statistic, $T$, being at least as favourable to $H_1$ as the observed test statistic, if $H_0$ is true.

\(^\dagger\)There are some technical restrictions on the range of models for which this result is exact: basically we require that the parameters estimators be linear in the data and that the variance of the data has been estimated using the sample mean of the model residuals (see section 5, for example): in this course the restrictions are always met, so you do not need to worry about them.
i.e. we must evaluate:

\[ p = \Pr[|T| \geq |\hat{\beta}/\hat{\sigma}_\beta|] \text{ where } T \sim t_{n-p} \]

By the symmetry of the t distribution, this is readily seen to be:

\[ p = 2 \times \Pr[T \leq -|\hat{\beta}/\hat{\sigma}_\beta|] \text{ where } T \sim t_{n-p} \]

and this quantity can be evaluated using the c.d.f. of \( t_{n-p} \).

**Numerical example:** Assume that 0.63, 0.86, 3.12, 0.46, 2.05, 0.16, 2.04, 1.64, 0.09, 0.82 are independent observations of \( N(\beta, \sigma^2) \) r.v.s where \( \sigma^2 \) is unknown. Let us test:

\[ H_0 : \beta = 0 \text{ against } H_1 : \beta \neq 0 \text{ but unknown.} \]

First evaluate the test statistic \( \hat{\beta}/\hat{\sigma}_\beta \):

```r
> x<-c(0.63,0.86,3.12,0.46,2.05,0.16,2.04,1.64,0.09,0.82)
> b.est<-mean(x); b.est  # estimate of beta
[1] 1.187
> s.est<-(var(x)/10)^0.5; s.est  # estimate of s.d. of beta estimator
[1] 0.3118513
> t.est<-b.est/s.est; t.est  # the estimate of the test statistic
[1] 3.806301
```

There are 10 data and 1 parameter, so we know that under the null hypothesis \( \hat{\beta}/\hat{\sigma}_\beta \sim t_9 \). Hence the p-value is readily evaluated:

```r
> p.value<-2*pt(-abs(t.est),9)
> p.value
[1] 0.004176763
```

There appears to be quite strong evidence against the null hypothesis and for the alternative. Notice however that the evidence is less strong than in the example in which we treated these same data as coming from a population with known variance.

**Another example:** consider data \((x_i, y_i), i = 1 \ldots 10\), where the \( y_i \) can be modelled as observations of independent r.v.s \( Y_i \sim N(\beta x_i, \sigma^2) \) where \( \sigma^2 \) and \( \beta \) are unknown. The parameters of the model are estimated from the data and this yields; \( \hat{\beta} = 0.1 \) and \( \hat{\sigma}_\beta = 0.05 \). Using the following R code find the p-value for the test:

\[ H_0 : \beta = 0 \text{ against } H_1 : \beta \neq 0 \text{ but unknown.} \]

```r
> 2*pt(-0.1,9) > 2*pt(-0.1,8)
[1] 0.9225364 [1] 0.922805
> 2*pt(-0.1/0.05,9) > 2*pnorm(-0.1,mean=0,sd=0.05)
[1] 0.07655282 [1] 0.04550026
> 2*pnorm(-0.1,mean=0,sd=0.05/10^0.5) > 2*pt(-0.1/0.05,8)
[1] 2.539629e-10 [1] 0.08051624
```

4.7.4 **Another example: a one-sided test**

Suppose that you are conducting a clinical trial to test out a new drug. In this case you would typically want to test the null hypothesis that the drug is useless against the alternative that it is useful. This is an example where it is very sensible to follow the hypothesis testing approach of favouring the null hypothesis until there is strong evidence against it. Given the expense of fully safety testing, manufacturing and marketing a new drug, it is important to be very sure that it really works before doing so! However, this
is also an example where it makes sense to consider a slight modification of the alternative hypotheses considered so far. Suppose for example that you start with a small 30 patient study, and record a measure \( x_i \) of drug efficacy for each patient, where the \( x_i \) can be treated as observations of independent r.v.s \( X_i \sim N(\beta, \sigma^2) \), \( \beta \) having the interpretation of being average drug efficacy in the population from which your patients have come, and \( \sigma^2 \) being an unknown parameter describing the variability in efficacy in the population. In this case the null hypothesis (the drug is useless) corresponds to the usual:

\[
H_0 : \beta = 0 \quad \text{‡‡}
\]

but the alternative, corresponding to the drug being useful, is now:

\[
H_1 : \beta > 0
\]

This will change the range of test statistic values favouring \( H_1 \). Specifically if the test statistic is:

\[
T = \frac{\hat{\beta}}{\hat{\sigma}_\beta}
\]

then only large positive values of \( T \) now favour \( H_1 \). Hence when we evaluate:

\[
\text{the probability of the test statistic, } T, \text{ being at least as favourable to } H_1 \text{ as the observed test statistic, if } H_0 \text{ is true},
\]

we will only be interested in the probability of obtaining \( T \) values greater than the one observed, rather than the probability of \(|T|\) being greater than the magnitude of the observed test statistic. So in this case the p-value would be:

\[
\Pr[T \geq \frac{\hat{\beta}}{\hat{\sigma}_\beta}] \text{ where } T \sim t_{29}
\]

If \( \hat{\sigma}_\beta = 0.7 \), evaluate the p-values for this case (i) if the study yielded \( \hat{\beta} = 1.4 \) and (ii) if it yielded \( \hat{\beta} = -1.4 \). Some of the following R session is helpful:

```r
> pt(1.4/0.7,29)
[1] 0.9725282
> pt(-1.4/0.7,29)
[1] 0.02747182
> pt(1.4/(0.7/sqrt(29)),28)
[1] 1
> pt(-1.4/(0.7/sqrt(29)),28)
[1] 9.122473e-12
> pt(-1.4,29)
[1] 0.08606059
> pt(1.4,29)
[1] 0.9139394
```

### 4.7.5 Significance levels

If we decide in advance that we will reject the null hypothesis if the p-value is less than some constant \( \alpha \), then \( \alpha \) is known as the **significance level** of the test. Popular choices for \( \alpha \) are 0.05, 0.01, 0.005 and 0.001, with the first being the most common.

‡‡We could also have written the null as \( H_0 : \beta \leq 0 \), but then we’d have ended up using a rather involved argument which shows that to test that null against our \( H_1 \), we actually need only consider the \( H_0 : \beta = 0 \)

50
4.7.6 Hypothesis testing in general

The general recipe for all hypothesis testing is:

1. Choose a test statistic, which can be calculated from the available data and will discriminate between \( H_0 \) and \( H_1 \), by taking one range of values if the null is true and a different range if the alternative is true.

2. Work out the p-value for the test. This is the probability of obtaining a test statistic at least as favourable to \( H_1 \) as the one actually observed, \textit{if \( H_0 \) is true}.

3. Small p-values (e.g. less than 0.05) suggest that since the observed test statistic would be quite improbable under \( H_0 \), it should be rejected in favour of \( H_1 \).

4. Large p-values suggest that the observed test statistic could easily have been generated under \( H_0 \), so there is no reason to reject it.

In the simplest examples we have looked at the test statistic was just the estimator of the parameter that the hypothesis test related to, but we also considered one more complicated case. However the general recipe given above can be used in cases in which the hypotheses involve several parameters simultaneously, and even in cases in which the hypotheses are not phrased in terms of parameters at all — the key is always to find a suitable test statistic, for which it is possible to work out the distribution under the null.

Finally note that we can also use the hypothesis testing approach with null hypotheses specifying values other than zero for the parameter(s), but we will not pursue this further in this course.

5 Linear modelling

So far we have only examined quite simple models in theoretical detail. For example, a model of observations of a single variable with a Normal distribution. However the \textit{methods} that we developed for those very simple models apply just as well to more ambitious models that describe the relationships between different variables. In this section we will look at a very useful class of models of this sort: \textbf{linear models} (sometimes known as “linear regression models”). This section introduces linear regression models via a simple example that allows all the key results to be derived and understood quite easily, and then introduces more general linear models.

5.1 A simple linear model

Let us start with an example that we’ll be using throughout the section to illustrate many of the relevant concepts in linear modelling. The dataset \texttt{hprice} (in the \texttt{mt1007} R package) contains prices (in Canadian dollars) of 100 houses sold in Windsor, Canada, during the summer of 1987. For each house sold we have price and several other variables related to the characteristics of the house. It would be good to be able to predict the market value of a house from its characteristics: with a suitable statistical model this may be possible.

To start with consider just one characteristic: the area of the plot of land that the house is built on (array \texttt{area}, measurements in square feet). It seems quite likely that the price of a house is related to this area, and if the data are plotted against each other this seems to be so:
Looking at this plot it doesn’t seem too unreasonable to suppose that the relationship between area and house price could be described by a straight line plus some random variation (or “random error”). Since we wouldn’t expect to pay anything for a house and plot of land of zero area, it’s reasonable to suppose that this straight line should “go through the origin”. So our first model might be:

\[
\text{price} = \beta \times \text{area} + \text{“random variation”}
\]

Where \( \beta \) is a parameter of the model, which will have to be estimated. The “random variation” term is used to absorb all those things that might affect house price in addition to plot area, which in this simple model we’re treating as being essentially unpredictable from one house to the next.

More formally, let \( y_i \) be the price of the \( i^{th} \) house in the sample and let’s model this as being an observation of a random variable \( Y_i \). Also let \( x_i \) be the corresponding plot area. The model becomes:

\[
Y_i = \beta x_i + \epsilon_i
\]

where \( \epsilon_i \) is the \( i^{th} \) random variation term - the part of \( y_i \) that can’t be explained by \( x_i \). The \( \epsilon_i \)'s are modelled as mutually independent* random variables, all with mean zero and the same variance \( \sigma^2 \). i.e. \( E[\epsilon_i] = 0, \text{var}[\epsilon_i] = \sigma^2 \). The model is an example of a linear model (or linear regression model) because the r.v.’s \( \epsilon_i \) and the parameter \( \beta \) enter the model in a linear way. \( y \) is known as the response variable and \( x \) the predictor variable or explanatory variable, the idea being that we are predicting (or explaining) \( y \) using \( x \), or equivalently that \( y \) is responding to \( x \). Parameters like \( \beta \) (but not \( \sigma \)) are known as model coefficients or regression coefficients.

As with all models, this one isn’t supposed to exactly represent the real data: it’s a mathematical caricature that we hope will do “well enough” at describing reality. The following diagram shows the model schematically:

\[
Y
\]

So the model is saying that the house prices \( y_i \) are given by a simple linear transformation of plot area \( x_i \) plus a random “error” term, but at present we don’t know what values the model parameters \( \beta \) and \( \sigma \) might take. They need to be estimated.

*Don’t forget that mutual independence of the \( \epsilon_i \)'s means that, for \( i \neq j \), knowing \( \epsilon_i \) doesn’t give you any information about the value that \( \epsilon_j \) might have
5.1.1 Parameter estimation

Least squares estimation is the approach to use for the parameter $\beta$. Recall from section 4.3 that this means finding the parameters that minimise:

$$S = \sum_{i=1}^{n} (y_i - E[Y_i])^2$$

(To avoid clutter, from now on the range of summation will be dropped from summation signs; whenever a $\sum$ appears in this section, the range of summation will be $i = 1 \ldots n$.) This means that we need to know $E(Y_i)$:

$$Y_i = \beta x_i + \epsilon_i \implies E(Y_i) = \beta x_i + E(\epsilon_i) = \beta x_i$$

This uses the fact that $\beta x_i$ is not a random variable, so its expected value is just $\beta x_i$, while $E(\epsilon_i) = 0$. Hence:

$$S = \sum (y_i - \beta x_i)^2$$

$S$ varies with $\beta$ as is clearly shown by this plot$^\dagger$:

The least squares estimate of $\beta$ will be the value at which $S$ is smallest. From the plot, this seems to be around 12, but a precise estimate can be obtained by mathematically finding the minimum of $S$ (which is the point at which $\partial S/\partial \beta = 0$):

$$\frac{\partial S}{\partial \beta} = \sum -2x_i(y_i - \beta x_i)$$

Setting this to zero gives the equation that must be solved to find the estimate of $\beta$:

$$\sum -2x_i(y_i - \hat{\beta} x_i) = 0 \Rightarrow \sum y_i x_i - \sum \hat{\beta} x_i^2 = 0$$

Which with trivial re-arrangement, leads swiftly to the estimate:

$$\hat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2}$$

Although R has excellent facilities for calculating estimates like $\hat{\beta}$ automatically, for now let’s calculate it the long way and plot the result (assuming hprice has been loaded and attached):

```r
> plot(area,price)
> beta<-sum(price*area)/sum(area^2);beta
[1] 11.72646
> abline(0,beta) # plots a straight line with intercept 0 and slope beta
```

$^\dagger$Produced using the following R code:

```r
> beta<-0:250/10;S<-0;
> for (i in 1:251) S[i]<-sum((price-beta[i]*area)^ 2);
> plot(beta,S,type="l")
```
... quite reasonable, though far from perfect (see later).

Now consider estimating $\sigma^2$. This is the variance of the $\epsilon_i$ random variables. As we have seen in section 4.3, $\sigma^2$ can’t be estimated by least squares. Hence we might as well estimate $\sigma^2$ using the variance of the estimated values of the $\epsilon_i$’s:

$$\hat{\epsilon}_i = y_i - \hat{\beta}x_i,$$

which leads to the following estimate of $\sigma^2$:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum \hat{\epsilon}_i^2 = \frac{1}{n-1} \sum (y_i - \hat{\beta}x_i)^2$$

(Note that $E(\epsilon_i) = 0$ and $\sum \hat{\epsilon} = 0$, leading to this simple form for the variance of the $\hat{\epsilon}_i$’s.) The sum of squared terms is divided by $n-1$, to get an unbiased estimator of $\sigma^2$ — the 1 in the $n-1$ is because we have estimated one parameter ($\beta$) in addition to $\sigma^2$. If we had estimated two parameters (e.g. slope and intercept), we would need a divisor of $n-2$. Again, R has more efficient ways of doing this automatically, but let’s calculate $\hat{\sigma}^2$ the long way:

```r
> sig2<-sum((price-beta*area)^2)/99; sig2
[1] 331095738
```

This corresponds to a $\hat{\sigma}$ of about 18,200 Canadian dollars.

### 5.1.2 How good is $\hat{\beta}$?

Now let’s briefly consider how $\hat{\beta}$ does as an estimator. (For example, what properties would we expect it to have if used on a large number of independent replicate data sets on house prices and plot area?) Firstly is it unbiased? To find out, we just need to find the expected value of $\hat{\beta}$ (written now in estimator form):

$$E(\hat{\beta}) = E \left( \frac{\sum Y_i x_i}{\sum x_i^2} \right) = \frac{\sum E(Y_i) x_i}{\sum x_i^2}$$

According to the model $E(Y_i) = \beta x_i$, so:

$$\ldots$$

The other thing it’s useful to know is the variance of $\hat{\beta}$. This is quite easy to work out using results 6 and 7 from section 3.4.4: if $Y_1, Y_2, \ldots, Y_n$ are independent r.v.s and $a_1, a_2, \ldots a_n$ are constants, then:

$$\text{var}(a_1 Y_1 + a_2 Y_2 + \ldots + a_n Y_n) = a_1^2 \text{var}(Y_1) + a_2^2 \text{var}(Y_2) + \ldots + a_n^2 \text{var}(Y_n)$$
or more compactly:

\[
\text{var} \left( \sum a_i Y_i \right) = \sum a_i^2 \text{var}(Y_i)
\]

In the case of \( \hat{\beta} \):

\[
a_i = \frac{x_i}{\sum x_i^2} \quad \text{and} \quad \text{var}(Y_i) = \sigma^2 \text{ for all } i
\]

(var(\(Y_i\)) = \text{var}(\epsilon_i), \text{since } Y_i = \beta x_i + \epsilon_i \text{ and } \beta x_i \text{ is fixed}) so . . .

\[
\text{var}(\hat{\beta}) = \sum \left( \frac{x_i}{\sum x_i^2} \right)^2 \sigma^2 = \frac{\sum x_i^2}{\left( \sum x_i^2 \right)^2} \sigma^2 = \frac{\sigma^2}{\sum x_i^2}
\]

Notice that this only works because the \( Y_i \)'s are assumed to be independent, each with the same variance, \( \sigma^2 \). For later convenience, let’s denote var(\(\hat{\beta}\)) by \( \sigma^2_{\hat{\beta}} \). Of course \( \sigma^2_{\hat{\beta}} \) involves the unknown \( \sigma^2 \), but the obvious estimator to use is:

\[
\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum x_i^2}
\]

Again, let’s calculate this the slow way:

\[
> \text{sb2<-sig2/sum(area^2);sb2}
> [1] 0.1491593
\]

corresponding to \( \hat{\sigma}_{\hat{\beta}} = 0.3862 \). So it seems that despite the scatter in the data, \( \beta \) is quite precisely estimated.

### 5.1.3 Adding a distributional assumption to the model

To be able to obtain confidence intervals for, and test hypotheses about, the parameters of a linear model, another element needs to be added to the model: a probability distribution for the \( \epsilon_i \)'s. The assumption is:

\[\epsilon_i \sim N(0, \sigma^2) \text{ for all } i\]

which translates (see section 3.5) to:

\[Y_i \sim N(\beta x_i, \sigma^2)\]

This means that the distribution of \( \hat{\beta} \) follows directly from the results of section 3.5. As we have seen, \( \hat{\beta} \) is a weighted sum of the \( Y_i \)'s. If these are independent and normally distributed, then \( \hat{\beta} \) must be normally distributed as well. i.e.

\[\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}})\]

Since \( \sigma^2_{\hat{\beta}} \) is unknown and has to be estimated, this result isn’t directly useful, but just as in section 4.6, it is possible to construct a pivotal quantity with known distribution:

\[\frac{\hat{\beta} - \beta}{\hat{\sigma}^2_{\hat{\beta}}} \sim t_{n-1}\]

The \( t \) distribution has \( n - 1 \) degrees of freedom because we have \( n \) data but have estimated one regression coefficient, \( \beta \). Given this distributional result, the results of section 4.6 can be used directly to find a confidence interval for \( \beta \). Suppose that we want a 95% CI for \( \beta \). First find the the points between which 95% of \( t_{99} \) r.v.’s will lie: \( t_{99}(0.025) \) and \( t_{99}(0.975) \), say:

\[
> \text{qt(0.025,99); qt(0.975,99)}
> [1] -1.984217
> [1] 1.984217
\]
Since the $t$ distribution is symmetric, $t_{99}(0.025) = -t_{99}(0.975)$. Using the results from section 4.6, the 95% is:
\[
\hat{\beta} \pm t_{99}(0.025)\hat{\sigma}_\beta = 11.7265 \pm 1.9842 \times 0.3862
\]
i.e. the 95% CI for $\beta$ is (10.96,12.49).

You can also use the distributional result to test hypotheses about $\beta$. As we will see, the most useful one to test is $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. If $H_0$ is true, then:
\[
\frac{\hat{\beta}}{\hat{\sigma}_\beta} \sim t_{n-1}
\]
and in the usual way we then see how likely the observed value of $\hat{\beta}/\hat{\sigma}_\beta$ is under $H_0$, by obtaining a $p$-value:
\[
> \text{sb}<-\text{sb}^20.5
> \text{pt(bbeta/sb,99)}
\]
[1] 1

...so to the limits of R’s calculation capabilities, $\hat{\beta}/\hat{\sigma}_\beta$ would certainly be less than the value we actually observed, if $H_0$ was true. So $H_0$ cannot be true‡.

### 5.1.4 Letting R do all the calculations

Most of the calculations covered so far are built into the R function `lm()`, which is straightforward to use. If the house prices are in array `price` and the plot areas are in array `area`, then:

\[
> \text{mod1<-lm(price\~area-1)}
\]
estimates the model parameters (also known as “fitting” the model), and stores these estimates, together with all the information needed for further analyses, in an object which I’ve called `mod1`. The only thing that requires explanation is the model formula `price\~area-1`, which tells `lm()` to fit a linear model to the data `price` and `area`, treating `price` as the response and `area` as the explanatory variable. The “-1” term tells `lm()` that the model should not have an extra “intercept” parameter, which would otherwise be included by default.

There are a number of ways of extracting information from a fitted model object like `mod1`. Much of what we need can be displayed using the `summary()` function:

\[
> \text{summary(mod1)}
\]

Call:
`lm(formula = price ~ area - 1)`

Residuals:
Min 1Q Median 3Q Max
-42843 -7935 1807 13164 75990

Coefficients:
| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| area 11.7265 | 0.3862 | 30.36 | <2e-16 *** |

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18200 on 99 degrees of freedom

‡Notice that in this case it was not necessary to complete the calculation of the $p$-value, since the result was so clear (and the actual $p$-value, $2 \times (1 - 1)$, is obviously 0).
Multiple R-Squared: 0.903,  Adjusted R-squared: 0.902
F-statistic: 921.9 on 1 and 99 degrees of freedom,  p-value: 0

For the moment, concentrate on the output following **Coefficients**: Since \( \beta \) is the parameter multiplying area, it is area that is used to identify it in the output. What follows on the same line are: \( \hat{\beta} \), \( \hat{\sigma} \), \( \hat{\beta}/\hat{\sigma} \), and the \( p \)-value associated with the test of \( H_0 : \beta = 0 \). Finally, there is a code indicating the smallest of several standard levels that \( H_0 : \beta = 0 \) could be rejected at. All the values match the ones that we calculated more laboriously above.

The estimate of \( \sigma \) is provided by the **Residual standard error**. Again, it matches the value obtained above. **99 degrees of freedom** refers to the fact that we fitted the model to 100 data points, but estimated one regression coefficient (\( \beta \)) in the process. For the moment, ignore the other output; most of it will be covered later.

### 5.2 Generalising a bit

The previous section introduced all the main ideas of linear modelling using a very simple linear model as an example. As you will begin to see in this section, all these results generalise very easily to linear models with more parameters. Considering the house price example again: the argument that the linear model relating price and area should pass through the origin was perhaps a bit bogus. We don’t have any data for properties much below 2000 square feet, and constraining the model to have a particular behaviour for values of area that are of purely theoretical interest is a bit dubious. We might therefore seek to fit a slightly more flexible two-parameter linear model:

\[
Y_i = \alpha + \beta x_i + \epsilon_i
\]

where again it is assumed that the \( \epsilon_i \)'s are independent r.v.'s with \( E(\epsilon_i) = 0 \) and \( \text{var}(\epsilon_i) = \sigma^2 \) for all \( i \).

Again the model coefficients will be estimated by least squares, that is by finding the values of \( \alpha \) and \( \beta \) that minimise:

\[
S = \sum (y_i - E(Y_i))^2 = \sum (y_i - \alpha - \beta x_i)^2
\]

Here is a plot\(^8\) of \( S \) against a range of \( \alpha \) and \( \beta \) values (labels on the contours should be multiplied by \( 10^{10} \)):
Find approximate least squares estimates of $\alpha$ and $\beta$ from the above plot:

More precise estimates can be found mathematically, by finding values of $\alpha$ and $\beta$ such that:

$$\frac{\partial S}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial S}{\partial \beta} = 0$$

Solution of the resulting pair of simultaneous equations is routine but tedious. While it provides good algebra practice, it doesn’t offer any statistical insight, so it will be omitted here. We obtain:

$$\hat{\beta} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

(Both are unbiased.) The estimate for $\sigma^2$ is:

$$\hat{\sigma}^2 = \frac{1}{n - 2} \sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Note that we now have divisor $n - 2$, because there are now two estimated parameters.

Because both estimates can again be written as weighted sums of the response data, we can derive variance estimates $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\beta^2$ for $\hat{\alpha}$ and $\hat{\beta}$. As before, the additional assumption of normality of the $\epsilon_i$’s means that the parameter estimators are normally distributed, and this leads to the useful results:

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_\alpha} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\beta} - \beta}{\hat{\sigma}_\beta} \sim t_{n-2}$$

(the $n - 2$ again arising because we have $n$ observations and have estimated two parameters). Exactly as for the previous model, these results allow confidence intervals to be calculated and hypothesis tests to be conducted.

The model fitting and associated calculations are easily performed in R:

```r
> mod2<-lm(price~area) # fit model: price = a + b area
> coef(mod2) # print regression coefficients

(Intercept)  area
19305.673696  7.881507
```

... so $\hat{\alpha} = 19306$ and $\hat{\beta} = 7.882$. Let’s see how the fitted model and data compare this time:

```r
> plot(area,price);abline(coef(mod2))
```

As soon as you have more than one parameter in a linear model, it’s actually rather silly to use the approach taken here, and much better to re-write both $S$ and the model using matrix algebra - it’s then very straightforward to find an expression for the least squares estimators that applies however many coefficients the model has, and similarly to find expressions for the variances of those estimators. Such results look very similar to the results that we obtained for the simple model $Y_i - \beta x_i + \epsilon_i$, just written in matrix form. It is these general results that are used by `lm` in R. Derivation of these general results requires second-year linear algebra.
Here is some of the information from the fitted model summary.

```r
> summary(mod2)
```

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 19305.674| 4917.325   | 3.926   | 0.000161 *** |
| area           | 7.882    | 1.044      | 7.551   | 2.27e-11 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17000 on 98 degrees of freedom
Multiple R-Squared: 0.3678, Adjusted R-squared: 0.3614

Exercise: Using

```r
> qt(0.995,98)
[1] 2.626931
```

find a 99% confidence interval for \( \beta \ldots \)

5.3 How well does the model fit?

There are two aspects to the question of how well the model fits. One is how well does it do at matching the data, and the other is how well do the assumptions of the model appear to be met? The latter question is very important — if the data do not appear to match the modelling assumptions reasonably closely, then our ability to use the model to draw conclusions (about the population that the data came from) will be severely curtailed.

5.3.1 Residual checking

The key model assumptions, and why they matter, are:

- The response data are observations of independent r.v.s all with the same variance. If these assumptions are not met, then the estimators of the variances of the data and of the parameter estimators will be wrong. Without good variance estimates, it is not possible to calculate reliable confidence intervals or test hypotheses about the models.

- The response data are observations of normally distributed r.v.s. If this assumption is violated, then confidence intervals will be suspect, as will the \( p \)-values associated with hypothesis tests. However, this assumption is usually not as critical as the independence and constant variance assumptions: this is because the estimators are weighted sums of the response variables, and as a result of the central limit theorem, may tend towards normality whatever the distribution of the response data, provided sample sizes are large relative to the number of model coefficients.

The only way to check these assumptions is to check the model residuals:

\[
\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta}x_i
\]

These should look like observations of independent r.v.s all with the same variance, and approximately normally distributed. The constant variance assumption is often violated because there is a relationship
between $E(Y_i)$ and $\text{var}(\epsilon_i)$. To diagnose this, you should always plot the residuals against the model estimates of $E(Y_i)$, namely the fitted values:

$$\hat{\alpha} + \hat{\beta}x_i$$

R allows you to extract both these quantities easily from the fitted model object using `resid()` and `fitted()`. Thus for the house price data we can:

```r
plot(fitted(mod2), resid(mod2))
```

This plot shows a clear problem: the variance of the residuals increases markedly as the fitted values increase.

A common way to violate the independence assumption is to get the model structure wrong, so that the relationship between $y$ and $x$ is modelled incorrectly, or perhaps a useful explanatory variable has been left out of the model. This sometimes appears as a systematic pattern in the residuals when plotted against fitted values, but it is usually clearer if the residuals are plotted against the explanatory variables, and potential explanatory variables. For the house price model, the command

```r
> plot(area, resid(mod2))
```

produces the plot on the left below, which shows no systematic pattern, but rather a random scatter about zero. Hence the independence assumption seems OK (although the non-constant variance problem is clearly visible in this plot too). For comparison, the plot on the right shows an example from another model fit to another set of data where there is a clear systematic pattern in the residuals: a violation of the assumption of independence.

The assumption of normality in the $\epsilon_i$'s can be checked with a simple plot as well. The plot, known as a “normal Q-Q plot” (where “Q-Q” stands for “Quantile-Quantile”), is based on the fact that if you take two samples of the same size from any two normal distributions, sort both into order and then plot one

[The problem in this plot could be tackled by adding an $x^2$ term to the model, something that can be done using the more general linear model covered later]
against the other, you’ll get a straight line with a little random scatter around it. If you do this with two samples, one of which is not normal, you’ll get a curve. A normal Q-Q plot for a set of data is produced by generating an idealised ordered sample from a N(0,1) distribution and plotting the original ordered data against this. If the data are normally distributed, we get a straight line (plus scatter); otherwise, we get a curve (plus scatter).

In the house price case, there’s not a great deal of point producing a Q-Q plot when we already know that there is a more serious problem (the non-constant variance), but here’s how it’s done anyway:

```r
> qqnorm(resid(mod2))
```

![Normal Q-Q Plot](image)

This produces a curve, indicating problems. At this stage, it might be worth looking at a histogram of the residuals as well, in order to help diagnose what’s going wrong.

Thus for the house price data, the independence assumption looks sound, but the constant variance assumption is not, and we will need to fix it.

### 5.3.2 How close is the model fit? $R^2$

One visual way to check how well the model does at predicting the response from the predictor variables is to plot the response data against the fitted values:

```r
> plot(fitted(mod2),price)
```

![Plot of fitted vs. price](image)

In this case, the model clearly gets the broad pattern right, but it can be a long way off in particular cases.

It’s useful to have a quantitative measure of how well the model is doing at explaining the response variable, so that the fit of different models can be compared. A common measure of how closely a model fits is $R^2$. The variance of the response variable can be estimated by:

$$
\hat{\sigma}_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2
$$
The variance left “unexplained” by the model is the variance of the $\epsilon_i$’s, which we have seen can be estimated by:

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum \epsilon_i^2$$

where $p$ is the number of regression coefficients in the model. Hence the proportion of the variance left unexplained by the model is $\hat{\sigma}^2/\hat{\sigma}_y^2$. This leads to the result that $R^2$, defined as the proportion of variance explained by the model, is given by:

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}$$

In fact, this gives the “adjusted $R^2$”, since we have used an unbiased estimator of $\hat{\sigma}^2$. The unadjusted $R^2$ uses the estimator

$$\hat{\sigma}^2 = \frac{1}{n - 1} \sum \epsilon_i^2$$

instead: the unadjusted estimator is the one referred to in the output from `lm()` as Multiple R-squared**.

Our second linear model for the house price data managed to explain 0.36 (36%) of the variance in house price — not that great, but then it is a very simple model — we really wouldn’t expect to be able to predict house price very accurately based only on plot area.

### 5.4 Transformation

When we checked the house-price model, there was a problem that the residual variance clearly depended on fitted values. Sometimes problems like this can be lessened substantially by transforming the response data, so that we effectively work on a different measurement scale. A useful type of transformation is obtained by raising the response data to some power.

Use $y^b$ as the response variable in place of $y$, where $b$ is some constant.

The problem with doing this is that, while it may fix problems with the assumptions about the $\epsilon_i$’s, it carries the side effect that the rest of the model structure changes too:

$$y^b = \alpha + \beta x$$

is not the same as

$$y = \alpha + \beta x$$

The problem is that we can’t raise both sides of $y = \alpha + \beta x$ to the power $b$ and retain a linear model. Further, if we fit this model, we’ll obtain unbiased estimates of the $E(Y_i^b)$’s, but not of the $E(Y_i)$’s.

Despite the problems, it’s possible to do something sensible for the house-price data. We saw that the simple model $Y_i = \beta x_i + \epsilon$ was not too bad for these data, and if we raise both sides $y = \beta x$ to the power $b$, and define a new parameter $\gamma = \beta^b$ we get:

$$y^b = \beta^b x^b \Rightarrow y^b = \gamma x^b$$

That is, we have a model that is linear in $\gamma$, and corresponds to a straight line on the original measurement scale. This suggests trying the linear model:

$$Y_i^b = \alpha + \delta x_i^b + \epsilon_i$$

where the usual assumptions apply to the $\epsilon_i$’s. A sensible value for $b$ might be obtained by trial and error* — try a range of values until the residual plots look OK.

For the house-price data, it turns out that $b = 0.2$ does quite a good job . . .

---

**The default $R^2$ returned by R’s `lm()` when the model contains no intercept is different again, and can be quite misleading — $\hat{\sigma}_y^2$ is estimated as $\sum y_i^2/(n - 1)$. I suggest ignoring the calculated $R^2$ for models when your model formula included “-1”.

*It is possible to develop maximum likelihood methods for estimating $b$, but we do not pursue this here.
> priceb<-price^0.2; areab<-area^0.2
> mod3<-lm(priceb~areab)
> par(mfrow=c(1,2))
> plot(fitted(mod3), resid(mod3))
> qqnorm(resid(mod3))

Still not perfect, but much better. Here is the (edited) fitted model object summary:

> summary(mod3)

Residuals:

 Min  1Q median  3Q Max
-1.03067 -0.31398 -0.06517 0.27107 1.64107

Coefficients:

 Estimate Std. Error t value Pr(>|t|)
 (Intercept) 3.5807 0.6466 5.538 2.56e-07 ***
 areab 0.9766 0.1216 8.032 2.17e-12 ***

---

Residual standard error: 0.4665 on 98 degrees of freedom
Multiple R-Squared: 0.397, Adjusted R-squared: 0.3908

This output includes a little information about the distribution of the residuals: 25% of residuals are smaller than 1Q, 50% are smaller than the Median and 75% are smaller than 3Q (all % approx.). This information should be enough to enable you to spot gross violations of the normality assumption — like severe lack of symmetry.

A more extreme option than the above, and one that works quite well here, is to take a log transform of both y and x. In fact, if we were to try the above with smaller and smaller values of b (but with b > 0), our results would tend towards those obtained here, using log transforms:

> logprice<-log(price); logarea<-log(area)
> mod4<-lm(logprice~logarea)
> par(mfrow=c(1,2))
> plot(fitted(mod4), resid(mod4))
> qqnorm(resid(mod4))
The analyses are very similar to those with $b = 0.2$, as expected, but both the constant variance and the normality assumption seem slightly improved.

5.5 Linear Models in General

Often there are several possible explanatory variables that we could use to model a response variable, and several of them together may be required to model the response adequately. In this case, linear multiple regression models (also known as general linear models\(^\dagger\)) are useful. For example, suppose that we have observations of a response variable $y$ and 4 predictor variables $x_1, x_2, x_3$ and $x_4$. A general (not generalized\(^\dagger\)) linear model for these might be:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$$

where $\beta_0, \beta_1, \ldots, \beta_4$ are the model coefficients, $Y_i$ is the r.v. of which $y_i$ is an observation, and the $\epsilon_i$'s are independent r.v.s such that $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ for all $i$. As before, if we want to find confidence intervals or test hypotheses about the population being modelled, we also need to assume that $\epsilon_i \sim N(0, \sigma^2)$.

To make this more concrete, let’s apply it to the house-price data, according to the following table:

\(^\dagger\)Do not confuse these with generalized linear models which are a much larger class of models.
Model Data Frame Description

<table>
<thead>
<tr>
<th>Model term</th>
<th>Data Frame array</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>(pb&lt;-price^0.2)</td>
<td>(House price)(^{0.2}) in (Canadian Dollars)(^{0.2}).</td>
</tr>
<tr>
<td>(x_1)</td>
<td>(ab&lt;-area^0.2)</td>
<td>(Plot area)(^{0.2}) in (Feet)(^{0.4}).</td>
</tr>
<tr>
<td>(x_2)</td>
<td>(beds)</td>
<td>Number of bedrooms.</td>
</tr>
<tr>
<td>(x_3)</td>
<td>(floors)</td>
<td>Number of floors (storeys).</td>
</tr>
<tr>
<td>(x_4)</td>
<td>(airco)</td>
<td>1 if house has air conditioning, 0 if it doesn’t.</td>
</tr>
</tbody>
</table>

As for the simpler models, the regression coefficients are estimated by least squares. That is, by finding the values of \(\beta_0, \beta_1, \ldots, \beta_4\) which minimise:

\[
S = \sum (y_i - E(Y_i))^2 = \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i})^2
\]

The mathematical principles by which this function is minimised are the same as the ones that we have seen in the simpler examples. Residual variance (\(\sigma^2\)) and coefficient estimator variances (\(\sigma^2\hat{\beta}\)) can likewise be obtained by straightforward generalizations of the results for the simple one-coefficient case (although you need a little second-year matrix algebra to do the generalization).

The calculations required to perform the minimisation and estimate variances are built into the R function \texttt{lm}(). Here is the R code required to fit the model and get a summary of the results (remembering that \texttt{areab} and \texttt{priceb} already contain suitably transformed area and price data):

\[
\texttt{> full.mod<-lm(priceb~areab+beds+floors+airco)} \\
\texttt{> summary(full.mod) # some bits edited out below}
\]

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.824505</td>
<td>-0.275123</td>
<td>-0.008602</td>
<td>0.239163</td>
<td>1.346944</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 3.94531 | 0.61284 | 6.438 | 4.92e-09 *** |
| areab | 0.80288 | 0.11891 | 6.752 | 1.15e-09 *** |
| beds | 0.11178 | 0.06838 | 1.635 | 0.10545 |
| floors | 0.10004 | 0.07229 | 1.384 | 0.16963 |
| airco | 0.31474 | 0.09915 | 3.174 | 0.00202 ** |

---

Residual standard error: 0.4281 on 95 degrees of freedom
Multiple R-Squared: 0.5076, Adjusted R-squared: 0.4869

Note that the adjusted \(R^2\) has now gone up to 0.49 — so this model is doing a bit better at predicting the response variable from the predictor variables. Notice also that \(\hat{\beta}_1\), the estimated coefficient for \texttt{areab}, differs from the estimate \(\hat{\beta}\) that we had previously when \texttt{areab} was the only predictor variable. This is hardly surprising, since we have changed the model since then, but it’s worth noting that some linear models for data from carefully designed experiments have the interesting property that the estimates of the parameters associated with a predictor variable do not depend on which other predictor variables are in the model.

**Exercise:** According to the estimated model, by how much does air-conditioning increase a house’s price by? Refer to the description of variable \texttt{airco} in the table, before answering this, and then think about the difference in predicted price between when it is 0 and when it is 1. Also, you will need to
back-transform your estimate from \((\text{Canadian Dollars})^{0.4}\) to Canadian dollars.

**Before** proceeding to test hypotheses or find confidence intervals, we need to check the assumptions on which these calculations will be based, by looking at residual plots:

```r
> par(mfcol=c(2,3))
> plot(fitted(full.mod), resid(full.mod)); qqnorm(resid(full.mod))
> plot(area, resid(full.mod)); plot(beds, resid(full.mod))
> plot(floors, resid(full.mod)); plot(airco, resid(full.mod))
```

Working from left to right: the constant variance and independence assumptions look OK (although there may be a problem for fitted values above 9). The residuals are clearly not exactly normal, but in practice this is our least important assumption, so the degree of departure is probably tolerable. The remaining plots show the residuals plotted against predictor variables: there doesn’t seem to be any systematic pattern in the results which might lead to violation of the independence assumption. Furthermore, the variances look reasonably constant across the values of the predictor variables. (You shouldn’t read much into the variability of 1, 2 or 3 points where that is all you have to go on!) So, the only potential cause for concern is the possible pattern at high fitted values, but we’ll ignore this issue here!

It’s now reasonable to think about hypothesis testing and confidence interval estimation. As in the simpler cases, we have

\[
\frac{\hat{\beta}_i - \beta}{\hat{\sigma}_\beta} \sim t_{n-p}
\]

where \(p\) is the number of parameters in the regression model: five in this case. We can now test \(H_0 : \beta_i = 0\) against \(H_1 : \beta_i \neq 0\) by noting that under \(H_0:\)

\[
\frac{\hat{\beta}_i}{\hat{\sigma}_\beta} \sim t_{n-p}
\]

As we have seen before, the resulting \(p\)-values are what R puts in the column `Pr(>|t|)` of a model object summary. High \(p\)-values indicate that the data are consistent with the null hypothesis that the
corresponding parameter† is zero. If we can’t reject the hypothesis that a model parameter is zero, then we might as well set it to zero in the model, i.e. remove the corresponding predictor variable.

**NB** It’s important to remove variables one at a time. This is because the estimated values of other parameters will change if we change the model, and so will their associated $p$-values. Hence the sensible strategy is to remove the term with the highest $p$-value (assuming it’s above some threshold, like the traditional 0.05) and then re-fit the model before deciding whether or not to remove another variable.

In the current case, *floors* has the highest $p$-value (0.17, which exceeds 0.05), so let’s remove it and examine the resulting model object:

```r
> small.mod<-lm(priceb~areab+beds+airco)
> summary(small.mod)
```

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.8445055</td>
<td>-0.3143635</td>
<td>0.0009016</td>
<td>0.2248981</td>
<td>1.3202308</td>
</tr>
</tbody>
</table>

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|----------|
| (Intercept)      | 4.00365  | 0.61429    | 6.517   | 3.31e-09 *** |
| areab            | 0.79646  | 0.11939    | 6.671   | 1.62e-09 *** |
| beds             | 0.15913  | 0.05949    | 2.675   | 0.00879 **  |
| airco            | 0.33188  | 0.09884    | 3.358   | 0.00113 **  |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4301 on 96 degrees of freedom
Multiple R-Squared: 0.4977, Adjusted R-squared: 0.482

Now all the terms are significant at least the 1% level (meaning that for each parameter, we can reject the null hypothesis that its value is zero). The change in the $p$-value of the *beds* term has arisen because the estimate of this coefficient has increased while its standard deviation has decreased. Why has this happened? The answer is that houses with more floors tend to have more bedrooms — both these variables carry rather similar information about the house. The problem is that, if two variables carry much the same information, they’ll both have much the same ability to “explain” the response variable. Hence we could obtain very similar fitted values to those obtained with the least squares estimates by, say, increasing the coefficient for one and reducing the coefficient for the other one. Our uncertainty about both coefficients is high if both (correlated) terms are included, but is much reduced if only one of them is included.

Given the above considerations, it is often useful to have a quantitative measure of the relatedness of predictor variables. A suitable measure was covered in some detail in section 4.2.2 of the Climate Change case study — the sample correlation coefficient $r$ between variables $w_i$ and $v_i$ measures how related they are. It is the sample covariance of the variables divided by the product of their sample standard deviations, leading to:

$$r = \frac{\sum (w_i - \bar{w})(v_i - \bar{v})}{\sqrt{\sum (v_i - \bar{v})^2 \sum (w_i - \bar{w})}}$$

Recall that $-1 \leq r \leq +1$. **R** function `cor()` finds $r$ for any two r.v.s. For example:

```r
> cor(beds,floors)
[1] 0.5105173
```

†Don’t forget: the null hypothesis specifies the value of the parameter for the population that we are modelling (all possible house prices in Windsor); it says nothing about the estimate of the parameter — we know that the estimate is not zero.
...confirming that number of bedrooms is moderately correlated with the number of floors. This is the highest correlation amongst the predictor variables in this data set, so it’s perhaps not surprising that the model only requires one of floors or beds.

While we’re on the subject of correlation, note the following:

\[ \text{cor(fitted(small.mod), priceb)}^2 \]
\[ 0.4977229 \]

This is the unadjusted $R^2$ for the model, i.e. the unadjusted $R^2$ is the square of the correlation between the model fitted values and the response data.

Our chosen model for the house-price data is now:

\[ Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_4 x_{4i} + \epsilon_i \]

Assuming $\epsilon_i \sim N(0, \sigma^2)$, complete this distributional statement:

\[ \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim \]

Hence find a 95% confidence interval for $\beta_1$: