School of Science and Technology
R.D. Murison

STAT260
Probability and Simulation Notes

Printed at the University of New England, December 8, 2007
STAT260 Semester I, 2009

Proposed Lecture Schedule

<table>
<thead>
<tr>
<th>WEEK</th>
<th>Module</th>
<th>Week Starting</th>
<th>Assignment Due</th>
<th>Day of Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Set up R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>23 Feb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.2–2.3</td>
<td>2 March</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.1–3.6</td>
<td>9 March</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.1–4.4</td>
<td>16 March</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4.5–4.6</td>
<td>23 March</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5.1–5.3</td>
<td>30 March</td>
<td>1</td>
<td>Monday March 30</td>
</tr>
<tr>
<td>7</td>
<td>5.4–5.5</td>
<td>6 April</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.1–6.4</td>
<td>4 May</td>
<td>2</td>
<td>Monday May 4</td>
</tr>
<tr>
<td>9</td>
<td>6.5–6.8</td>
<td>11 May</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7.1–7.5</td>
<td>18 May</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>8.1–8.3</td>
<td>25 May</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>8.4–8.5</td>
<td>1 June</td>
<td>3</td>
<td>Monday June 1</td>
</tr>
<tr>
<td>13</td>
<td>– –</td>
<td>8 June</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table of Contents

1 Probability basics 1
2 Distribution of random variables 57
3 Mathematical expectation 87
4 Standard discrete distributions 117
5 Standard continuous distributions 153
6 Bivariate distributions 207
7 Transformations of random variables 257
8 Sampling distributions 281
# Module 1

## Probability basics

### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Sample space, sample points and events</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Some definitions</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Assigning probabilities</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Empirical (relative frequency) approach</td>
<td>7</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Classical approach</td>
<td>9</td>
</tr>
<tr>
<td>1.3.3</td>
<td>Subjective approach</td>
<td>10</td>
</tr>
<tr>
<td>1.3.4</td>
<td>Axiomatic approach</td>
<td>10</td>
</tr>
<tr>
<td>1.4</td>
<td>Rules of probability</td>
<td>12</td>
</tr>
<tr>
<td>1.4.1</td>
<td>Equally-likely outcomes</td>
<td>13</td>
</tr>
<tr>
<td>1.5</td>
<td>Conditional probability and independence</td>
<td>13</td>
</tr>
<tr>
<td>1.5.1</td>
<td>Conditional probability</td>
<td>13</td>
</tr>
<tr>
<td>1.5.2</td>
<td>Multiplication rule</td>
<td>15</td>
</tr>
<tr>
<td>1.5.3</td>
<td>Independent events</td>
<td>16</td>
</tr>
<tr>
<td>1.5.4</td>
<td>Independence and mutually exclusive events</td>
<td>17</td>
</tr>
<tr>
<td>1.5.5</td>
<td>Partitioning the sample space</td>
<td>19</td>
</tr>
<tr>
<td>1.5.6</td>
<td>Bayes’ theorem</td>
<td>20</td>
</tr>
<tr>
<td>1.6</td>
<td>Computing probabilities</td>
<td>22</td>
</tr>
<tr>
<td>1.6.1</td>
<td>Tree diagrams</td>
<td>22</td>
</tr>
<tr>
<td>1.6.2</td>
<td>Tables</td>
<td>24</td>
</tr>
</tbody>
</table>
Module objectives

Upon completion of this module students should be able to:

- understand the concepts of probability, including sample space and sample points of an experiment
- define probability from different perspectives and apply them to compute probabilities in various situations
- explain the concept of conditional probability and independence of two events
- differentiate between mutually exclusive events and independent events
- apply Bayes’ Theorem to find the ‘reverse’ probability by using the formula and tree diagrams
- use the rules of combinations and permutations to compute the probabilities of various events involving counting problems
- recall and make use of the properties of arithmetic and geometric series
- recall and recognise the exponential, logarithmic and binomial power series
1.1 Introduction

What is the chance of snow falling in the next 12 months where you live? In some places, the answer is easy. In Townsville, Queensland, Australia, the answer is almost no chance. In London, the answer is that snow is almost certain to fall. In other places, however, the answer may be not so obvious.

If you live in Armidale, NSW, your answer may be that the chance is very slim. In Hobart, Tasmania, it may be unlikely. In Canberra, ACT, it may be reasonably likely. On Mt Kosciusko, it is almost certain. The language used is rather vague: ‘unlikely’, ‘very slim’, ‘reasonably likely’, ‘almost certain’. Probability simply tries to make these statements more precise by assigning numerical quantities to the events.

In most scientific, engineering and business applications, there are situations where it is almost impossible to be certain, and so probability is necessary.

- What is the chance that an investment will make money in the next 12 months?
- What is the chance that a medical procedure will be successful on a particular patient?
- How likely is it that the water level at a certain river crossing will rise above 2 meters in the next 20 years? And if it does, what is the chance that the bridge will not be washed away?
- What is the chance that two DNA sequences may come from identical parents if their genome sequences is almost identical?
- What is the chance of rain on the weekend?
- Given the evidence, what is the chance that the defendant is guilty?

To set about answering questions such as these we need a framework to work in. We need to define the concept of probability and provide a notation and theory consistent with our common experience in dealing with uncertainty.

As well as being important in modelling real-world phenomena, the tools developed in this course provide a grounding in the theory of statistics. An understanding of the analysis of data cannot be separated from an understanding of probability theory. A probability course such as this is a foundation course in any statistics major.

In this module we discuss the concept of probability, learn some notation and definitions, and develop some theory useful in manipulating probabilities.
A basic grounding in *Set Theory* is necessary for the understanding of probability theory. Section 2.3 of WMS and section 1.4 of DGS give a brief review of set notation and some basic definitions and ideas.

**Reading 1.1** Read DGS, Section 1.4 or WMS, Sections 2.1, 2.2 and 2.3.

(Throughout this Study Book, WMS denotes Wackerly, Mendenhall & Schaeffer (5th or 6th edition) and DGS denotes DeGroot & Schervish (3rd edition).)

Be aware of variations in notation. For example, the complement of an event $A$ (everything that is *not* in event $A$) is denoted $\overline{A}$ in WMS and $A^c$ in DGS. A further common notation is $A'$. Also the probability of an event $E$ occurring is either written as $P(E)$ or $\Pr(E)$.

### 1.2 Sample space, sample points and events

**Reading 1.2** DGS, Sections 1.3 and 1.6; WMS, Section 2.4 up to the paragraph under Definition 2.5.

The concepts of *sample space* and *sample point* are very important to understand. Without an understanding of these concepts, the notion of probability becomes vague. Note, however, that there are times when it is difficult to identify the sample space precisely.

Also note that the definition for *experiment* or *random experiment* is fairly vague. The intention is to think of an experiment as a process by which data is gathered. Definition 1.1 in the next section captures the essentials.

One important, special event is the null or empty event, $\emptyset$. This set is empty in that it contains no sample points at all. For example, the set of all years that snow has fallen in Singapore is the null or empty set.

Definition 2.4 in WMS needs some explaining. One part of the definition states that a discrete sample space is one that contains a finite set of points. For example, after rolling a standard die, there are exactly six possible outcomes. The other part indicates that a discrete sample space can contain a countable number of distinct points. Some books even use the phrase ‘countably infinite’ (and so does WMS at the start of Section 2.5). What does this mean?

Consider rolling a standard die and counting the number of throws needed until a 6 is thrown. What is the sample space? (That is, what are the
possible outcomes?) The 6 could appear on the first roll, or the second, or the third. But there is a chance it will take 20 rolls, or 50 rolls. What is the maximum number? There is no maximum number (though there is a point beyond which it becomes very unlikely). In such a situation, we say there are a ‘countable’ number of points even though we can’t list them all.

DGS discusses finite sample spaces at this stage. A finite sample space is a special case of a discrete sample space. The more general discrete sample space definition, and other important definitions, are stated below for convenience, consistency and reference.

### 1.2.1 Some definitions

**Definition 1.1** A random experiment is characterized by the following features:

(i) Each experiment can be repeated indefinitely under essentially unchanged conditions;

(ii) Although we are, in general, not able to state what a particular outcome will be, we are able to list all possible outcomes of the experiment;

(iii) As the experiment is performed repeatedly, the individual outcomes tend to occur in a haphazard manner. However, as the experiment is repeated a large number of times, a definite pattern of regularity appears.

**Definition 1.2** A sample space (or event space or outcome space) for an experiment \( E \) is a set of all possible outcomes and we shall denote this set by \( S \). An element of \( S \) is called a sample point.

**Definition 1.3** A discrete sample space is one that contains a finite or countably infinite number of distinct sample points.

**Definition 1.4** An event is a collection of sample points. That is, an event is a subset of \( S \), and we write \( A \subseteq S \). In the two extreme cases, \( S \) itself is an event called the certain (or sure) event, and \( \emptyset \) is called the impossible (or nonrealisable) event.

**Definition 1.5** An important concept is that of an occurrence of an event. We say that an event \( A \) occurs on a particular trial of an experiment if the outcome of the trial is one of the sample points in \( A \). Also, \( \overline{A} \) (known as the complement of \( A \)) occurs if \( A \) does not occur.
Module 1. Probability basics

Definition 1.6 For two events $A$ and $B$ associated with a random experiment $E$ with sample space $S$,

\begin{align*}
    A \cup B & \text{ occurs iff the outcome is in } A \text{ or in } B \text{ (or in both)} \\
    A \cap B & \text{ occurs iff the outcome is in } A \text{ and in } B \\
    A \cap \overline{B} & \text{ occurs iff the outcome is in } A \text{ but not in } B
\end{align*}

Note that provided there is no ambiguity likely, $A \cup B$ is sometimes denoted $AB$. DGS uses this notation.

Definition 1.7 Events $A$ and $B$ are mutually exclusive (m.e.) if and only if $A \cap B = \emptyset$.

The term disjoint is used for sets, whereas mutually exclusive is used when referring to events.

Definition 1.8 We say that an event $A$ implies an event $B$ iff $A \subseteq B$.

Definition 1.9 An elementary event (or a simple event) is an event that cannot be decomposed. It is an event with only one sample point.

1.3 Assigning probabilities

In order to proceed it is necessary to formalize the notion of the chance of an event occurring. This is done by assigning a number to this chance and calling this number a probability. What values should these numbers take? There are options for assigning numbers to the chance of events occurring.

It’s already been suggested that an impossible event be assigned 0, and a certain event assigned 1. Although this is an accepted standard it’s worth reflecting on other alternatives; eg assigning $\infty$ (infinity) to a certain event and 0 to an event that will happen just as often as not happen, so that positive numbers are assigned to events more likely to happen than not, and negative numbers to events less likely to happen than not?

It should be clear that the 0, 1 method is really quite arbitrary although attractive in that it aligns with the notion of proportions or percentages as numbers between 0 and 1. However, there is one other method in standard use also. In this method, an impossible event is assigned 0, a certain event is assigned $\infty$, and an event that will happen just as often as not happen is assigned 1. These values are known as ‘odds’ and are commonly used in gambling.
1.3. Assigning probabilities

We will only use the system with probabilities between 0 and 1 in this course. It is surprisingly difficult to develop a method of assigning sensible probabilities to events. Four methods are discussed here.

**Reading 1.3** DGS, Section 1.2.

### 1.3.1 Empirical (relative frequency) approach

This is often one of the easiest approaches to understand. In this method, the experiment is repeated a large number of times and the proportion of times the event occurs is noted. Mathematically, if an experiment is repeated \( n \) times, and \( m \) of these favour event \( E \), then the probability is

\[
P(E) = \lim_{n \to \infty} \frac{m}{n}.
\]

To accurately compute the probabilities, it is necessary for \( n \) to be very large. In any case, approximate probabilities can only ever be found.

Note that this method cannot always be used. For example, suppose a company wishes to determine the probability that matches in a given box will light, it is necessary to actually strike some matches; but then they can no longer be used, which may defeat the purpose of the experiment. Or consider the probability that the airbag in a car will prevent a serious injury in the passenger. It is not financially viable to crash thousands of vehicles with passengers in them to see how many break bones! (Fortunately, car manufacturers use dummies to represent people and crash small numbers of cars to get some indications of the probabilities). In these situations, sometimes computer simulations can be used to approximate the probabilities.

**Example 1.1** In 1954, Jonas Salk developed a vaccine against polio (see Williams [35, §1.1.3]). In a study to test the effectiveness of the vaccine, the data in Table 1.1 were collected. The relative frequency approach can be used to estimate the probabilities of developing polio with the vaccine and without the vaccine (the control group):

\[
P(\text{develop polio in control group}) = \frac{115}{201\,229} = 0.00057
\]

\[
P(\text{develop polio in vaccinated group}) = \frac{33}{200\,745} = 0.00016
\]

Since the sample sizes are large, the computed probabilities are reasonable estimates of the true probabilities for the population from which the sample was drawn. Do you think this suggests the vaccine was effective?
Module 1. Probability basics

Table 1.1: The number of paralytic cases for two groups of children: one group of controls and another vaccinated with the Salk polio vaccine.

<table>
<thead>
<tr>
<th>Group</th>
<th>Number treated</th>
<th>Paralytic cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaccinated</td>
<td>200 745</td>
<td>33</td>
</tr>
<tr>
<td>Control</td>
<td>201 229</td>
<td>115</td>
</tr>
</tbody>
</table>

Example 1.2 A German named Wolf [36] (as reported in Hand [16, dataset 131]) rolled a die 20 000 times and counted the number of times each face came up. The results are shown in Table 1.2. Using the classical approach, there are six equally likely outcomes (a 1, 2, 3, 4, 5 and 6) so the chance of rolling each should be $1/6 = 0.166\ldots$. Even after 20 000 throws, the probabilities are only approximate!

<table>
<thead>
<tr>
<th>Face showing</th>
<th>Number of times</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3407</td>
<td>0.1704</td>
</tr>
<tr>
<td>2</td>
<td>3631</td>
<td>0.1816</td>
</tr>
<tr>
<td>3</td>
<td>3176</td>
<td>0.1588</td>
</tr>
<tr>
<td>4</td>
<td>2916</td>
<td>0.1458</td>
</tr>
<tr>
<td>5</td>
<td>3448</td>
<td>0.1724</td>
</tr>
<tr>
<td>6</td>
<td>3422</td>
<td>0.1711</td>
</tr>
</tbody>
</table>

Table 1.2: The number of times each face on a standard, fair die was face up in 20 000 rolls. See Hand [16, dataset 131].

Example 1.3 A computer can be used to simulate large numbers of coin tosses. Using a probability of a head as 0.5, 1000 tosses were simulated. After each toss, the probability of obtaining a head using all the available information was computed. For one such simulation, the running probabilities are shown in Figure 1.1.

Example 1.4 The maximum temperature in Toowoomba is a continuous random variable. The relative frequency approach can also be used
to approximate probabilities for continuous random variables. Using data supplied by the DPI\textsuperscript{1} from 1 Jan 1889 to 21 July 2002, the maximum temperatures recorded for January on 3534 days. Of these, 302 maximum temperatures are greater than 30\degree. So an estimate of the probability that a day in January has a maximum temperature greater than 30\degree is 302/3534 = 0.08545, or about 8.5%.

1.3.2 Classical approach

This method of assigning probabilities requires being able to define a sample space containing a set of equally likely outcomes. In practice, this is only true for experiments involving tossing coins, rolling dice, dealing cards, and the like. The classical approach is underpinned by what is sometimes called the principle of indifference; that is, there’s no reason to believe that any one particular outcome in the sample space should be favoured over any other. Consequently, if there are \( n \) outcomes, all are assigned probability \( 1/n \).

Hence, for example, in a game of Lotto, there’s no reason to argue that the six numbers 1, 2, 3, 4, 5, 6 should be any more or less likely to occur than any other six numbers. The interesting aspect of this approach is usually in the determination of \( n \). We will see that the use of combinatorics is often necessary.

\textsuperscript{1}The Queensland Department of Primary Industries.
Example 1.5 When a standard die is tossed, there are six outcome which are equally likely; the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. So the probability $P$ (an even number is thrown) can be computed by counting those outcomes in the sample space that are even (there are three) and dividing by the total number of outcome in the sample space (six). So the probability is $3/6 = 0.5$.

1.3.3 Subjective approach

This approach is the least scientific of all; it simply refers to any individuals idea of the probability. Surprisingly, sometimes there are no other options! What proportion of the Australian public will vote for the Labor Party at the next election? What is the chance an investment will be a success? While there may be some facts on which to draw, people ultimately have their own individual (and different) opinions.

There are major shortcomings in this approach; most importantly, people may have very different ideas of the probability. However, often people with a lot of experience can estimate these probabilities (often unconsciously) based on their experience. In this case, there is also an element of the relative frequency approach in their method.

1.3.4 Axiomatic approach

Reading 1.4 DGS, Section 1.5; the remainder of WMS, Section 2.4, from after Definition 2.5 to the end of the Section.

For a sample space $S$ for an experiment, for every event $A$ (a subset of $S$) a number $P(A)$ can be assigned which is called the probability of event $A$. The three axioms of probability are

1. $P(A) \geq 0$ (that is, probabilities are never negative);
2. $P(S) = 1$ (that is, the total probability (or the probability of everything in the sample space) is one);
3. If $A_1, A_2, \ldots$ form a sequence of pairwise mutually exclusive events in $S$, then $P(A_1 \cap A_2 \cap A_3 \cap \ldots) = \sum_{i=1}^{\infty} P(A_i)$ (that is, for mutually exclusive events, the probability of the union of events is the sum of the individual probabilities).
1.3. Assigning probabilities

While this approach can sometimes help to assign probabilities, its main purpose is to formally define the rules that apply to probabilities. This means that if any other approach is used to assign probabilities, the probabilities must obey the rules specified in these axioms.

The axioms of probability can also be used to develop probability formulae.

Example 1.6 It is possible to prove that $P(\emptyset) = 0$. (Note that this may appear ‘obvious’ but it is not one of the axioms!) Since, by definition, the empty set $\emptyset$ contain no points, $\emptyset \cup A = A$ for any event $A$. Also, $\emptyset \cap A = \emptyset$ as $\emptyset$ and $A$ are mutually exclusive (that is, they have no elements in common). Hence, by the third axiom

$$P(\emptyset \cup A) = P(\emptyset) + P(A) \quad (1.1)$$

But since $\emptyset \cup A = A$, $P(\emptyset \cup A) = P(A)$, and so $P(A) = P(\emptyset) + P(A)$ from Equation (1.1). Hence $P(\emptyset) = 0$.

This may have seemed obvious, but it is important to note that all probability formulae can be developed just from assuming the three axioms of probability. See Section 1.4.

Example 1.7 What is the likelihood of rain in Charleville (a town in Queensland) during April? Many farmers could give a subjective estimate of the probability based on their experience and the conditions. It is not possible to use a classical approach to determine the probability. There are only two outcomes—it will rain or it will not rain—but these are almost certainly not equally likely. It would be a bad error to say that the probability of rain is $1/2$ as there are only two outcomes, and one is that rain will fall!

A relative frequency approach can be adopted. Using data supplied by the DPI\(^2\) from 1882 to 1994 (113 years), there are 96 years with rain in April. An approximation to the probability is therefore $96/113 = 0.85$, or 85%.

\(^2\)The Queensland Department of Primary Industries.
1.4 Rules of probability

**Reading 1.5** WMS, Section 2.8.

Many results can be deduced from the axioms. The important ones are often referred to as *rules of probability* and are stated here. They allow probabilities to be computed in complex situations.

**Theorem 1.10** For any event $A$, the probability of ‘not $A$’,

$$P(\overline{A}) = 1 - P(A)$$

This is commonly known as the *complementary rule of probability*.

**Theorem 1.11** For any two events $A, B$, the probability of $A$ or $B$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This is commonly known as the *addition rule of probability*.

If the two events are mutually exclusive we have a corollary of Theorem 1.11.

**Corollary 1.12** If the two events $A, B$ are mutually exclusive, then the probability of $A$ or $B$,

$$P(A \cup B) = P(A) + P(B)$$

Some other results that follow from the axioms are the following.

$$P(\emptyset) = 0 \quad (1.2)$$

$$0 \leq P(A) \leq 1 \quad (1.3)$$

For events $A$ and $B$,

$$P(A \cup B) \leq P(A) + P(B) \quad (1.4)$$

For events $A$ and $B$, if $A \subseteq B$ then

$$P(B \cap \overline{A}) = P(B) - P(A) \quad (1.5)$$

For events $A$ and $B$, if $A \subseteq B$ then

$$P(A) \leq P(B) \quad (1.6)$$
1.5 Conditional probability and independence

1.4.1 Equally-likely outcomes

For the calculation of probabilities for events in a finite sample space, we can sometimes describe our sample points in such a way that we have equally-likely outcomes; ie the classical approach to the assignment of probabilities is appropriate. Associating probabilities with events in this situation boils down to counting sample points. Example 1.5 is a simple demonstration of the principle which we now formalise.

**Theorem 1.13** Suppose that a sample space \( S \) consists of \( k \) equally likely elementary events \( E_1, E_2, \ldots, E_k \), and \( A = \{E_1, E_2, \ldots, E_r\} \), \((r \leq k)\), then

\[
P(A) = \frac{n(A)}{n(S)}
\]

This method of calculating the probability of an event is sometimes called the *sample-point method* (WMS, Section 2.5). Caution should be exercised in its application as errors are frequently committed by incorrectly diagnosing the nature of a simple event and by failing to list all the sample points in \( S \).

Methods of counting the points in a sample space are discussed in Section 1.7.

1.5 Conditional probability and independence

1.5.1 Conditional probability

**Reading 1.6** DGS, Section 2.7; WMS, Sections 2.1 and 2.2.

Assume that a sample space \( S \) for the experiment has been constructed, an event \( A \) has been identified, and its probability, \( P(A) \), has been determined. We then receive additional information that some event \( B \) has occurred. We wish to determine the effect, if any, that this new information has on our choice of \( P(A) \). That is, we wish to determine the probability that \( A \) will occur, given that \( B \) has already occurred. We call this probability the *conditional probability of* \( A \) *given* \( B \), and denote it by \( P(A \mid B) \).

**Example 1.8** In Example 1.7, the chance of obtaining rain during April in Charleville was computed to be 0.85 using the empirical approach. However, studies have suggested that the chance of rainfall may depend on the value of the Southern Oscillation Index (SOI). Stone and Auliciems [31] have defined five phases of the SOI. During any one
Module 1. Probability basics

<table>
<thead>
<tr>
<th>SOI Phase for April</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days of rain:</td>
<td>9</td>
<td>24</td>
<td>12</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>Number of days:</td>
<td>16</td>
<td>25</td>
<td>15</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>Prob of rain:</td>
<td>0.56</td>
<td>0.96</td>
<td>0.80</td>
<td>0.96</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 1.3: The (empirical) probabilities of rainfall at Charleville during April, conditional on the SOI phases during that month.

month, the SOI phase can be determined. For the month of April, the SOI phases have been found and the corresponding empirical probabilities of observing rainfall in Charleville have been computed for each SOI phase, and are shown in Table 1.3.

The chance of observing rain is conditional on the SOI phase during that month. If the SOI is in Phase 1, the (historical) chance of rain is $P(\text{rain} \mid \text{Phase 1}) = 0.56$, while if the SOI is in Phase 2, history suggests rain is almost certain (the chance is $P(\text{rain} \mid \text{Phase 2}) = 0.96$).

Example 1.9 Suppose I throw a standard six-sided die and I cover the die so you cannot see what is rolled. However, I tell you that the number that has been rolled is even; what is the probability that the number is a 2?

Solution Without the extra information, the chance of obtaining a 2 is $1/6$ as the sample space here has six (equally likely) outcomes: a 1, 2, 3, 4, 5 or 6. That is, $S = \{1, 2, 3, 4, 5, 6\}$. But since there is extra information—that the number is even—this probability can be refined. If the number is even, there are only three possibly outcomes: a 2, 4 or 6. So the reduced sample space under the condition that the number rolled is even is $S_R = \{2, 4, 6\}$. Then, since each of these outcomes are equally likely, the chance of obtaining a 2 is $1/3$. We could write $P(2 \mid \text{even number rolled}) = 1/3$.

Example 1.10 To help understand conditional probability, consider the difference between these two probabilities:
1.5. Conditional probability and independence

(a) $P$ (a person dies $|$ person falls out of a plane); and
(b) $P$ (person falls out of a plane $|$ a person dies).

In the first, the probability that a person dies is very high given that they fall from an (airborne) plane. In the second, given that a person has died, the probability is very small that the cause was falling from a plane as very few people die in this manner. Thus, the first probability is very close to one, and the second is very close to zero.

There are two methods for handling conditional probability problems: first principles (or common sense) or a formal definition of $P(A \mid B)$. For the first, we simply consider the original sample space $S$, throw out all its sample points not consistent with our new information that $B$ has occurred, form a new sample space, $S^*$ say, and recompute the probability of event $A$ relative to $S^*$. $S^*$ is often called the reduced sample space. This method is appropriate when the number of outcomes is relatively small.

The following definition allows us to handle more general problems.

**Definition 1.14** Let $A$ and $B$ be events in $S$ with $P(B) > 0$. Then

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

The definition automatically takes care of the sample space reduction we noted earlier.

**1.5.2 Multiplication rule**

As a consequence of Definition 1.14, we have the following theorem.

**Theorem 1.15** For any events $A$ and $B$, the probability of $A$ and $B$ is

$$P(A \cap B) = P(A) P(B \mid A)$$

$$= P(B) P(A \mid B)$$

This is known as the multiplicative rule for probabilities. It can be generalised to any number of events. For example, for three events $A, B, C$, we have

$$P(A \cap B \cap C) = P(A) P(B \mid A) P(C \mid A \cap B) \quad (1.7)$$

What say we have $k$ events? (Check the formula in WMS, Section 2.8.)
1.5.3 Independent events

To avoid the conditions $P(A) > 0$ and $P(B > 0)$ we follow the DGS approach and define independence as follow.

**Definition 1.16** Two events $A$ and $B$ are said to be independent if

$$P(A \cap B) = P(A) P(B).$$

Otherwise the events are said to be dependent.

Provided $P(B) > 0$, we see from Definitions 1.14 and 1.16 that $A$ and $B$ are independent if and only if $P(A \mid B) = P(A)$. This statement of independence is sensible in that the first probability is the probability of $A$ occurring if $B$ has already occurred. The second is the probability of $A$ occurring. If these are equal, then the fact that $B$ has occurred has made no difference to the probability that $A$ has occurred, which is precisely what is meant by independence.

The idea of independence can be generalised to more than two events. For three events we have the following definition and this naturally extends to any number of events.

**Definition 1.17** Three events $A$, $B$ and $C$ are mutually independent if and only if

$$P(A \cap B) = P(A) P(B)$$
$$P(A \cap C) = P(A) P(C)$$
$$P(B \cap C) = P(B) P(C)$$
$$P(A \cap B \cap C) = P(A) P(B) P(C)$$

It is worth noting that three events can be pairwise independent in the sense of Definition 1.16 but not mutually independent.

The following theorem concerning independent events is sometimes useful.

**Theorem 1.18** If $A$ and $B$ are independent events then

(i) $A$ and $\overline{B}$ are independent

(ii) $\overline{A}$ and $B$ are independent

(iii) $\overline{A}$ and $\overline{B}$ are independent
1.5. Conditional probability and independence

1.5.4 Independence and mutually exclusive events

It’s important to understand the difference between mutually exclusive and independent events.

Check back at the Definition 1.7 for the concept of mutual exclusiveness. The simple events defined by the outcomes in a sample space are m.e. since only one can occur in any realisation of the experiment. Mutually exclusive events have no common outcomes: you cannot, for example, get both an A grade and a C grade for this course. Obtaining one excludes the possibility of the other.

In contrast, if two events are independent then whether or not the one occurs does not affect the chance of the other happening. If event $A$ can occur, then $B$ happening will not influence the chance of $A$ happening if they are independent, so it certainly does not exclude the possibility of the other occurring.

Example 1.11 Independent events refer to events that have no impact on each other. For example, how I get to work tomorrow (walk or ride my bicycle) has no bearing on whether or not I have a busy day at work. They are completely independent. If I decide to walk, I might have a busy day... or I might not. If I walk, this does not exclude the possibility of having a busy day at work; in fact, it has no bearing on whether my day is busy or not.

Confusion between mutual exclusiveness and independence arises sometimes because the sample space is not clearly identified. For example, consider an experiment involving tossing a coin twice. We have the sample space $S_2 = \{(H,H), (H,T), (T,H), (T,T)\}$ and these outcomes are mutually exclusive with probabilities of $1/4$ assigned to each according to the classical approach. So, for example, $P(\{(H,H)\}) = 1/4$.

An alternative way of looking at this experiment though is to think of repeating the experiment of tossing a coin once. For one toss of a coin, the sample space $S_1 = \{H, T\}$ and $P(H) = 1/2$ is the probability of getting a head on the first toss and also the probability of getting a head on the second toss. The event of getting a head on the first toss and of getting a head on the second toss are not mutually exclusive however because both events can occur together, in the sense that $(H, H)$ is an outcome in $S_2$. Whether or not the outcomes $H, H$ occurred simultaneously, because the two coins were tossed at the one time, or sequentially, in that one coin was tossed...
twice, is irrelevant. The fact that our interest is in the joint outcomes from two tosses is what matters. Really it is best to think of getting a head on the ‘first’ toss as the event \( E_1 = \{(H, H), (H, T)\} \) and of getting a head on the ‘second’ toss as the event \( E_2 = \{(H, H), (T, H)\} \) where \( E_1 \) and \( E_2 \) are events defined on \( S_2 \). This makes it clear that \( E_1 \) and \( E_2 \) are not mutually exclusive because \( E_1 \cap E_2 \neq \emptyset \).

The two events \( E_1 \) and \( E_2 \) are however independent because regardless of whether or not a head occurs on one of the tosses, the probability of a head occurring on the other is still 1/2. If you don’t believe this then you must believe coins have got memories or can talk to each other! And this indeed provides another way of calculating the probability of the two heads occurring ‘together’ as \( 1/2 \times 1/2 = 1/4 \) since the probabilities of independent events multiply.

Example 1.12 (This example is based on one in Williams [35].) Mendell [21] conducted some famous experiments in genetics. In one study, Mendel crossed a pure line of round yellow peas with a pure line of wrinkled green peas. Table 1.4 shows what happened in the second generation. For example, \( P(\text{round peas}) = 0.7608 \). Biologically, we would expect about 75% to be round; the data appear reasonably sound in this respect.

Is the type of pea (round or wrinkled) independent of the colour? That is, if the pea is round, does it have any effect in the colour of the pea? To test independence, one form of the formula is \( P(\text{round} \mid \text{yellow}) = P(\text{round}) \). In other words, the fact that the pea is yellow does not affect that probability that the pea is round. From Table 1.4,

\[
P(\text{round}) = 0.7608, \\
P(\text{round} \mid \text{yellow}) = \frac{0.5665}{0.7482} = 0.757.
\]

These two probabilities are very close. Given that the data in the Table is only a sample (from the population of all peas), it seems reasonable

<table>
<thead>
<tr>
<th></th>
<th>Yellow</th>
<th>Green</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round</td>
<td>0.5665</td>
<td>0.1942</td>
<td>0.7608</td>
</tr>
<tr>
<td>Wrinkled</td>
<td>0.1817</td>
<td>0.0576</td>
<td>0.2392</td>
</tr>
<tr>
<td>Total</td>
<td>0.7482</td>
<td>0.2518</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 1.4: The second generation results from Mendel’s experiment crossing a pure line of round yellow peas with a pure line of wrinkled green peas.
to conclude that there’s little evidence suggesting colour and shape of the peas are not independent.

1.5.5 Partitioning the sample space

The concepts introduced in this section allow us to determine the probability of an event by what is called the event-decomposition approach in WMS. The technique is to express the event of interest, say event $A$, as a composition (unions and/or intersections), then apply the laws of probability to find $P(A)$. We first need some definitions.

**Definition 1.19** The events $B_1, B_2, \ldots, B_k$ are said to represent a partition of the sample space $S$ if

1. $B_i \cap B_j = \emptyset$ for all $i \neq j$
2. $B_1 \cup B_2 \cup \ldots \cup B_k = S$
3. $P(B_i) > 0$ for all $i$

A way to think of a partition is that the events $B_1, B_2, \ldots, B_k$ are mutually exclusive and exhaustive. That is, when the experiment is performed, one and only one of the events $B_i$ ($i = 1, \ldots, k$) occurs.

We use this concept in the following theorem.

**Theorem 1.20** Let $A$ be an event in $S$ and $\{B_1, B_2, \ldots, B_k\}$ a partition of $S$. Then

$$P(A) = P(A \mid B_1) P(B_1) + P(A \mid B_2) P(B_2) + \ldots + P(A \mid B_k) P(B_k)$$

**Proof** The proof follows from writing $A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_k)$, where the events on the RHS are mutually exclusive. The third axiom of probability together with the multiplication rule yield the result. ♠
1.5.6 Bayes’ theorem

**Theorem 1.21** Let $A$ be an event in $S$ such that $P(A) > 0$ and \{$B_1, B_2, \ldots, B_k$\} a partition of $S$. Then

$$P(B_i \mid A) = \frac{P(B_i) P(A \mid B_i)}{\sum_{j=1}^{k} P(B_j) P(A \mid B_j)}$$

for $i = 1, 2, \ldots, k$.

**Proof** This is a direct application of Definition 1.14, the multiplication rule and Theorem 1.20.

Bayes’ theorem has many uses in many disciplines (including finance, medicine and in legal matters) as it uses a conditional probability that can be easy to find or estimate, to compute a probability that is not easy to find or estimate.

The theorem is the basis of a branch of statistics known as Bayesian statistics which involves using pre-existing evidence in drawing conclusions from data. Bayesian statistics is explored later in the statistics major.

**Example 1.13** The following example about cervical cancer is taken from data collected by Graham and Shotz [14]. In the study, women known to have cervical cancer and women known to not have cervical cancer are asked about their age at first pregnancy. By doing so, it is easy to estimate the probability $P$ (age at first pregnancy $\mid$ get cervical cancer), since the presence or absence of cervical cancer is ‘given’ as this condition is known at the time. But a more useful probability might be $P$ (get cervical cancer $\mid$ age at first pregnancy). For convenience, define $CC$ to be the event that a woman gets cervical cancer, and $Y$ be the event that the woman was under 25 at first pregnancy.
From the study, the following probabilities were obtained:

\[
\begin{align*}
P(Y \mid CC) &= 0.857 \\
P(Y \mid CC) &= 0.640 \\
P(CC) &= 0.134
\end{align*}
\]

(Do not be alarmed by these figures; they are gathered only from women aged 50 to 59 in Buffalo, New York who had at least one child.) From this information, we wish to compute, for example \( P(CC \mid Y) \); that is, ‘reverse’ the probability. This is more useful since most of the time the age of a woman at first pregnancy is known (or ‘given’), and the chance of getting cervical cancer is of interest.

To do this, we use Bayes’ theorem. First, every woman in the study either has or does not have cervical cancer. This is a *partition* of the sample space: the sample space has been split into two groups that encompass the *whole* sample space. In each group, some women will have had their first pregnancy under 25, and some when older than 25. See the top figure in Figure 1.2.

The situation may be clearer using a *tree diagram*; see Figure 1.3. In either case, Bayes’ theorem can then be used to determine

\[
P(Y \cap CC) = P(Y \mid CC) \times P(CC) = 0.857 \times 0.134 = 0.1148;
\]

and likewise

\[
P(Y \cap CC) = P(Y \mid CC) \times P(CC) = 0.640 \times 0.866 = 0.5542.
\]

Then the probability of a woman in the study being under 25 at first pregnancy is

\[
P(Y) = P(Y \cap CC) + P(Y \cap CC) = 0.1148 + 0.5542 = 0.6690.
\]

Then, using Bayes’ theorem again,

\[
P(CC \mid Y) = \frac{P(CC \cap Y)}{P(Y)} = \frac{0.1148}{0.6690} = 0.1716.
\]
Module 1. Probability basics

Figure 1.2: The Venn diagram for Example 1.13. The top diagram shows the partitioning of the sample space on the basis of having or not having cervical cancer (CC). The middle diagram shows those women with cervical cancer who were under 25 at first pregnancy; the bottom diagram then adds those women without cervical cancer who were under 25 at first pregnancy. Combining the two pieces gives the total probability that a woman in the study was under 25 at first pregnancy as 0.66904. (Note that the diagram has not been drawn to scale.)

1.6 Computing probabilities

A tree diagram (see Figure 1.3) is just one of several tools that can assist in calculating and manipulating probabilities. Typically there is more than one way of arriving at a correct answer and as experience grows you will become adept at selecting the best method for a particular problem. Some examples are given below to illustrate the various approaches.

1.6.1 Tree diagrams

Tree diagrams are useful when an experiment can be seen or thought of as occurring in steps or stages.
Example 1.14 In an article by Anderson [1], 124 first year Honours mathematics students at a university in the UK were asked the question:

The point $z$ lies in the fourth quadrant (i.e. $-\pi/2 < \arg(z) < 0$) of the complex plane. $z^*$ denotes the complex conjugate of $z$. Prove or disprove $\arg(z + z^*) = 0$.

The expectation was that most should have been able to correctly answer the question. The data from the last table on page 499 of the article has been used to construct a tree diagram, where the first stage or step can be considered the gender of the person to whom the test is given, and the next step is how the given person answers the question; see Figure 1.4.

So, for example, $P(\text{correct} \mid \text{male}) = 0.82$, and $P(\text{not answered} \mid \text{female}) = 0.17$. Note that the probabilities are usually given on the branches and the events at the end of the branches. Apart from the initial probabilities, the probabilities given are conditional, depending on which branch the probabilities are placed.

Now, the probability $P(\text{male} \mid \text{correct})$ can be computed as $(0.76 \times 0.82)/(0.76 \times 0.82 + 0.24 \times 0.70) = 0.788$. 

---

Figure 1.3: A tree diagram for Example 1.13.
Module 1. Probability basics

1.6.2 Tables

When there are two variables of interest, tables are a convenient way of summarizing the information. Although tables can be used to represent experiments using conditional probabilities, it is usual to use the table to represent the whole sample space.

Example 1.15 The following table summarizes the number of men and women in the Senior Executive Service of the Australian Public Service in 1994. The data is from Townsend and McLennan [32]. Band 1 is the lowest of the three Bands.

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 1</td>
<td>231</td>
<td>1031</td>
</tr>
<tr>
<td>Band 2</td>
<td>47</td>
<td>332</td>
</tr>
<tr>
<td>Band 3</td>
<td>7</td>
<td>79</td>
</tr>
<tr>
<td>Total</td>
<td>285</td>
<td>1442</td>
</tr>
</tbody>
</table>

The counts can be turned into probabilities (by dividing by the total number, $285 + 1442 = 1727$) or left as they are.

Then, for example, the probability that a randomly chosen Senior Executive is female is $285/1727 = 0.16$. Likewise, the probability that
a randomly chosen male Senior Executive is higher than Band 1 is 
\((332 + 79)/1442 = 0.29\).

### 1.6.3 Venn diagrams

Venn diagrams can be useful when there are two events, sometimes three, but become unworkable for more than three. Often, tables can be used to better represent situations that have been shown in Venn diagrams.

**Example 1.16** Townsend and McLennan [32, p 84] report on the Australian Labour Force in 1995. They infer that the proportion of the work force that is male is 0.568; the proportion that is under 25 is 0.218; and that 0.113 are males *and* under 25. This information can be used to construct Venn diagrams to summarise the sample space as there are two variables, gender (either Male or Female) and age (either under 25, or 25 and over). See Figure 1.5. The same information can be compiled into a Table; see Table 1.5.

![Venn Diagrams](image)

**Figure 1.5:** The Venn diagram for Example 1.16.
What proportion of the labour force were males over 25 in 1995? This can be answered using numerous methods: The Venn diagram, the table, or formulae. The answer can be taken directly from the Table to give 0.455. Using the Venn diagram, the Male set consists of 0.133 that belong in the Under 25 set, so the remainder must be over 25, so the probability is again 0.455.

To use the formula, define $M$ to be Male, and $Y$ to be Under 25. Then, $P(M) = 0.568$, $P(Y) = 0.218$ and $P(M \cap Y) = 0.113$. Now, $P(M \cap \neg Y) + P(M \cap Y) = P(M)$ (the Theorem of Total Probability), so that $P(M \cap \neg Y) = 0.568 - 0.113 = 0.455$.

Try the three methods to compute the proportion of females over 25.

### 1.6.4 Sample spaces

Sometimes the most convenient way to compute probabilities is to list the sample space. This only works for small discrete sample spaces, and works best when each of the outcomes in the sample space are equally likely.

**Example 1.17** Consider rolling two standard dice and noting the sum of the numbers shown. The sample space can be listed in a table:
1.7 Counting techniques

It is then easy to see that $P(\text{sum is } 7) = \frac{6}{36}$ by counting the equally likely outcomes that sum to seven.

When the number of elements in the sample space is large, principles and formulae exist which facilitate counting. Some of these are described in the next section.

### 1.7 Counting techniques

#### Reading 1.9

DGS, Sections 1.7 and 1.8; WMS, Section 2.6.

The number of elements in a set is often large and a complete enumeration of the set is unnecessary if all we want to do is to count the number of elements in it.

#### 1.7.1 Multiplication principle

If we have $m$ ways to perform act $A$ and if, for each of these, there are $n$ ways to perform act $B$, then there are $mn$ ways to perform the acts $A$ and $B$. This is referred to as the multiplication principle and to state it differently,

With $m$ elements $a_1, a_2, \ldots, a_m$ and $n$ elements $b_1, b_2, \ldots, b_n$ it is possible to form $mn$ pairs containing one element from each group.

The principle can be extended to any number of sets. Given 3 sets of elements $a_1, a_2, \ldots, a_m$, $b_1, b_2, \ldots, b_n$, and $c_1, c_2, \ldots, c_p$, the number of distinct triplets containing one element from each set is equal to $mnp$. 

<table>
<thead>
<tr>
<th>Die B</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>12</td>
</tr>
</tbody>
</table>
1.7.2 Permutations

Another important counting problem deals with different permutations of a finite set. The word *permutation* in this context is simply an ordering of the set.

Suppose a set has $n$ elements. If we wish to order these, we must first choose the first element of the ordering. This can be done in $n$ different ways. The second element must then be chosen and this can be done in $(n-1)$ ways (since it cannot be the same as the first). There are then $(n-2)$ ways for the third, and so on. We find that (by the multiplication principle), there are $n(n-1)(n-2)\ldots 2 \times 1$ different ways to order the $n$ elements. This is denoted by $n!$ and called ‘$n$ factorial’. So we have

$$n! = n(n-1)(n-2)\ldots(2)(1)$$

where $n \geq 1$ and we define $0! = 1$.

In some cases, it is necessary to know, not all possible orderings of a set, but rather, the number of ways in which the first $r$ elements of an ordering may be chosen. For a problem such as this, the first element in the ordering may be chosen in $n$ different ways, the second in $(n-1)$ ways, the third in $(n-2)$ ways, and so on, down to the $r$th, which may be chosen in $(n-r+1)$ ways. This number is denoted by $^nP_r$, so we may write

$$^nP_r = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!} \quad (1.8)$$

and this expression is referred to as the number of permutations of $r$ elements from a set with $n$ elements.

1.7.3 Combinations

In some cases, when choosing a subset, order is not important. A *combination* of $r$ elements from a set $S$ is a subset of $S$ with exactly $r$ elements.

Now we need to find the number of combinations of $r$ elements from a set with $n$ distinct elements, and we will denote it by $^nC_r$. The number of permutations is given by 1.8, but of course the number of combinations must be smaller, since each combination gives rise to several permutations. In fact, each combination of $r$ elements can be ordered in $r!$ different ways, and so gives rise to $r!$ permutations. This being so, the number of permutations must be $r!$ times the number of combinations. We have then,

$$^nC_r = \frac{n(n-1)\ldots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!} \quad (1.9)$$
This is often written as \( \binom{n}{r} \).

It is important to understand the difference between permutations and combinations. With permutations, the order is important, whereas order is not important with combinations.

In addition, both formulae apply when there is no replacement of the items. That is, after an item is selected, it cannot be selected again.

Notation for combinations and permutations is varied. Some notation for permutations include \( nPr \), \( nP_r \), \( P_r^n \) or \( P(n, r) \). Notation for combinations include \( nCr \), \( nC_r \), \( C_r^n \), \( \binom{n}{r} \) or \( C(n, r) \).

**Example 1.18** Consider the set of integers 1, 3, 5, 7. Choose two numbers, without replacement, and call the first \( a \) and the second \( b \). If we then compute \( a \times b \), there are \( \binom{4}{2} = 6 \) possible answers since the order is not important (that is, \( 3 \times 7 \) gives the same answer as \( 7 \times 3 \)).

However, if we compute \( a \div b \), there are \( 4P_2 = 12 \) possible answers, since the order is important (that is, \( 3 \div 7 \) gives a different answer than \( 7 \div 3 \)).

---

**Example 1.19** Consider a biologist planning an experiment using \( mdx \) mice (mice with a strain of Duchene muscular dystrophy). Measurements of heart weights need to be taken at 3 weeks, 6 weeks and 12 weeks. There are twenty mice in the experiment. To take the measurements, the mice need to be sacrificed. If the biologist needs to select six mice at 3 weeks, how many ways are there to select the mice?

Since the mice cannot be replaced (they are sacrificed), and the order is not important, there are

\[
\binom{20}{3} = \frac{20!}{3!17!} = 1140
\]

ways of selecting the sample.
1.7.4 Sampling with or without replacement

The definitions of \( ^nP_r \) and \(^nC_r\) above, refer to the situation where the choices at successive stages are made without replacement or without repetition. That is, after the first element is chosen, it is not available for re-selection when choosing the second element.

We will now consider

(i) the number of ordered selections,

(ii) the number of unordered selections,

when \( r \) elements are chosen from \( n \) with replacement or with repetition.

The number of ordered selections of \( r \) objects chosen from \( n \) with replacement is \( n^r \). The first object can be selected in \( n \) ways, the second can be selected in \( n \) ways, etc., and by the multiplication principle, we have \( n^r \).

The remaining case of the number of ways \( r \) objects can be chosen from \( n \) with replacement and without regard to order is somewhat more difficult and won’t be attempted here. (The number of ways is actually \( (^{n+r-1}) \).)

1.7.5 Binomial theorem

The binomial expansion

\[
(a + b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}
\]

for \( n \) a positive integer, is often referred to as the Binomial Theorem and hence \( \binom{n}{r} \) is referred to as a binomial coefficient. Knowledge of this series and associated properties is sometimes useful in counting. Some of these properties are stated below.

1. \( \binom{n}{r} = \binom{n}{n-r} \), \( r = 0, 1, \ldots, n \).

2. As a special case of 1, \( \binom{n}{0} = 1 = \binom{n}{n} \).

3. \( \sum_{r=0}^{n} \binom{n}{r} = 2^n \).
1.8 Introducing statistical computing

Computers and computer packages are essential tools in the application of statistics to real problems. In this course you will be introduced to the statistical package R. It will be used to illustrate various concepts and should help you understand the theory.

In particular, R will be used to verify theoretical results obtained. To do this a technique called computer simulation will be used. Simulation can also be used to solve problems for which it may be difficult (or impossible) to obtain a theoretical result. Sometimes these numerical solutions to intractable analytical problems is termed Monte Carlo simulation. That term was coined at Los Alamos during construction of the bomb in 1943 when computers were in their infancy.

1.8.1 The gameshow problem

A gameshow contestant is told there is a car behind one of three doors. There is a goat behind each of the other doors. The contestant is asked to select a door. The host of the show (who of course knows where the car is) now opens one of the doors not selected by the contestant to reveal a goat. The host now gives the contestant the choice of retaining the door chosen first or switching and choosing the the other (unopened) door. What is the contestants best strategy?

1. Always retain the first choice.
2. Always change and select the other door.
3. Choose either unopened door (chances of success are equal anyway??).

A computing solution Only a brief outline of the method is given here. A more complete solution is given later (see Exercise 1.11).

Step 1: Generate a random sequence of length 100 consisting of the integers 1,2 and 3. (These numbers represent the door the car is behind on each of 100 nights.

Step 2: Generate another such sequence to represent the contestants first choice on each of the 100 nights.

Step 3: The number of times the numbers in the two columns agree represents the number of times the contestant will win if the contestant doesn’t change doors. If the numbers in the two columns don’t agree then the contestant will win only if the contestant decides to change doors.
My result  For the 100 nights simulated, contestants would have won the car 30 times if they retained their first choice which means they would have won 70 times if they had changed. (Does this agree with your intuition?) This implies (correctly!) that the best strategy is to change. You might like to try doing the simulation for yourself. (The correct theoretical probability of winning if you retain the original door is 1/3 but increases to 2/3 if you change. (This is not a trivial theoretical problem!) We obtained a reasonable estimate of these probabilities from the simulation, 30/100 and 70/100. These estimates could be expected to improve for larger sample sizes.)

1.9 Looking ahead

The study of probability is basic to the study of modern statistics. Consider the following example, the data coming from the Journal of Applied Physiology (1984), 1020–1023. (Source: Devore, Jay and Peck, Roxy (1993) Statistics: The Exploration and Analysis of Data,(2nd ed.).)

The experiment

Chronic airflow obstruction (CAO) severely limits the exercise capability of sufferers. Maximum exercise ventilation for each of 21 CAO patients was determined under two different experimental conditions. Fifteen patients recorded their best ventilation under experimental condition 1.

The question  Of interest is whether the evidence allows us to conclude that either experimental condition gives better results than the other.

A computing solution  We might be tempted to say that clearly experimental condition 1 is better since 15 of 21 did better. What is wrong with this reasoning? Let us suppose that a patient actually performs equally well under either experimental condition. Then it will be only a matter of chance under which experimental condition a patient is recorded as performing best. The original question can now be restated as: Is it likely that 15 (or more) CAO patients will appear to do better under experimental condition 1 (or 2) if in fact there is no difference between the effect of the two?

Let us tackle the problem experimentally. The problem above is similar to that of tossing a coin. Here we could let a head represent a patient doing better under condition 1, a tail as better under condition 2. Tossing the
1.9. Looking ahead

Coin 21 times then gives us one possible outcome for our experiment. If we do this a large number of times then we can observe how many times we get 15 or more patients doing better on condition 1 (or condition 2). (NOTE: If there is really no difference the results could go in either direction, not just the direction observed in the ONE experiment described above.) Suppose we carried out the experiment 1000 times, we should be able to make some definite conclusion. However, it is of course not really feasible to actually throw a coin this number of times. We get around this problem by simulating the experiment on a computer, in our case utilising the statistical package, R. (The method is described below, full technical details are to be found in the Computer Guide.)

We let 1 represent a patient doing better on experimental condition 1, and by a 0 a patient doing better on condition 2. We can then generate a sequence of 1’s and 0’s representing the results of 21 patients. This can be repeated 1000 times. The results of one such a simulation experiment are given below. Let \( X \) represent the number of times (out of 21) that CAO patients under condition 1 do better than under condition 2,. The Count column then gives the number of trials (out of 1000) in which there were exactly \( x \) CAO patients doing better under condition 1. This is given as a percentage in the third column.

### Tabulated statistics

<table>
<thead>
<tr>
<th>( x )</th>
<th>Count</th>
<th>Percentage</th>
<th>( x )</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>0.20</td>
<td>11</td>
<td>173</td>
<td>17.30</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.10</td>
<td>12</td>
<td>120</td>
<td>12.00</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>1.00</td>
<td>13</td>
<td>109</td>
<td>10.90</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>2.20</td>
<td>14</td>
<td>65</td>
<td>6.50</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>6.50</td>
<td>15</td>
<td>23</td>
<td>2.30</td>
</tr>
<tr>
<td>8</td>
<td>104</td>
<td>10.40</td>
<td>16</td>
<td>14</td>
<td>1.40</td>
</tr>
<tr>
<td>9</td>
<td>132</td>
<td>13.20</td>
<td>17</td>
<td>3</td>
<td>0.30</td>
</tr>
<tr>
<td>10</td>
<td>157</td>
<td>15.70</td>
<td>Total</td>
<td>1000</td>
<td>100.00</td>
</tr>
</tbody>
</table>

(Note: In the 1000 trials although the values of \( x = 0,1,2,18,19,20 \) or 21 were possible, they did not occur.)

We now have the information to make an informed answer to the original question. Notice that 40 (=23+14+3) of the experiments result in 15 or more CAO patients doing better on condition 1 while 35 (=2+1+10+22) of the experiments result in 6 or less of the CAO patients doing better on condition 1 (the latter corresponding of course to 15 or more doing better on condition 2.) That is 75 out of the 1000 experiments resulted in an outcome as extreme
as the one observed in the actual experiment, or an estimated probability of 0.075 that a result this extreme would occur purely by chance if there is in fact NO DIFFERENCE in effect of the two experimental conditions. If the probability is above 0.05 (that is 5%) we usually say the case of a difference in effect has not been shown on the grounds that we can’t be sure it has not occurred by chance alone. (5% would mean we expect it to occur by chance 1 out of 20 times, a rate which is not unreasonable to expect to happen.) Our conclusion here would be that there is insufficient evidence to conclude a difference in effect of the two experimental conditions.

**Comments**

1. The above conclusion relies on estimating an appropriate probability.
2. The calculations involved using a computer to simulate the problem.
3. Very little theoretical knowledge of probability was necessary although we did assume that the probability of getting a head on each toss of a fair coin was 0.5.
4. In this case if we had known more about probability modelling simulation would not have been necessary, as a satisfactory theoretical model can be found for this problem. (This could then be applied in many other similar situations.) A knowledge of probability theory would have made our task much simpler here.
5. One of the aims in this course is to provide the tools for you to model different experimental situations.
6. *Final note:* Simulation is a very powerful tool. In many complex situations it may either provide the only method of proceeding or may be used to gain insight into the problem that allows some progress to be made by more conventional means. It is also used to verify theoretical results.

We will return to this example later when we have the tools to find the probability model.

### 1.10 Some useful series

For convenience and reference we state some results here involving series that occur quite frequently in probability and statistical theory.
1.10. Some useful series

1. Some finite series

(a) Sum of natural numbers

\[ 1 + 2 + 3 + \ldots + n = \frac{n(n + 1)}{2} \]  \hspace{1cm} (1.11)

(b) Sum of squares of natural numbers

\[ 1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n + 1)(2n + 1)}{6} \]  \hspace{1cm} (1.12)

(c) Geometric series

\[ a + ar + ar^2 + \ldots + ar^{n-1} = \frac{a(1 - r^n)}{1 - r} \]  \hspace{1cm} (1.13)

\[ \text{as } n \to \infty \]  \hspace{1cm} (1.14)

where the infinite series only converges if \(|r| < 1\).

(d) Binomial expansion

\[(a + b)^n = b^n + \binom{n}{1} ab^{n-1} + \ldots + \binom{n}{r} a^r b^{n-r} + \ldots a^n \]  \hspace{1cm} (1.15)

2. Power series

A power series is a series of the form \( \sum_{n=0}^{\infty} a_n z^n \).

\[ e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots \]  \hspace{1cm} (1.16)

\[ e^{-z} = 1 - z + \frac{z^2}{2!} - \frac{z^3}{3!} + \ldots \]  \hspace{1cm} (1.17)

\[ \log(1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \ldots \]  \hspace{1cm} (1.18)

\[ \log(1 - z) = -z - \frac{z^2}{2} - \frac{z^3}{3} - \ldots \]  \hspace{1cm} (1.19)

\[ (1 - z)^{-1} = 1 + z + z^2 + z^3 + \ldots \]  \hspace{1cm} (1.20)

\[ (1 - z)^{-r} = 1 + rz + \frac{r(r + 1)z^2}{2!} + \frac{r(r + 1)(r + 2)z^3}{3!} + \ldots \]  \hspace{1cm} (1.21)

Note that the expressions on the RHS above do not converge to the LHS for all values of \( z \).
1.11 Self-assessment exercises

The following exercises are designed to provide practice at problem-solving based on the material in this module. Solutions are provided at the end of the module. Additional exercises are available in the next section and in the textbook. In the R exercises, remember your simulated answers will not be exact. Increasing the sample size will in general improve the result. (You can increase the sample size indefinitely by repeating the simulation a number of times.)

**Ex. 1.1** Given \( P(A) = 0.3, P(B = 0.5) \) and \( P(A \cap B) = 0.24 \), find
(a) \( P(A \cup B) \)
(b) \( P(\overline{A} \cap B) \)
(c) \( P(\overline{A} \cup \overline{B}) \)
(d) \( P(A | B) \)
(e) Are events \( A \) and \( B \) independent?

**Ex. 1.2** A computer programmer is running a subroutine to solve the general quadratic equation \( y = ax^2 + bx + c = 0 \). This is an ‘experiment’ of choosing values for the constants \( a, b \) and \( c \) satisfying the quadratic equation.
(a) Define the sample space \( S \) for the experiment.
(b) Define the event \( A \): ‘the equation has two equal real roots’.

**Ex. 1.3** Senie et al. [28] studied the relationship between breast self-examination and the age of diagnosis of breast cancer; the data are shown in Table 1.6.
(a) Compute \( P(\text{age}<45) \).
(b) Compute \( P(\text{conduct self-examinations monthly}) \).
(c) Compute \( P(\text{age under 45} \cap \text{never use self-examination}) \).
(d) Compute \( P(\text{age between 45 and 59} \cup \text{never use self-examination}) \).
(e) Compute \( P(\text{age between 45 and 59} | \text{never use self-examination}) \).
(f) Is the frequency of breast self-examination independent of age at diagnosis? Justify.
Ex. 1.4 A courier company is interested in the length of time a certain set of traffic lights is on green. The lights have been set so that the longest possible time between green lights in any one direction is 2 minutes, and the shortest possible time is 15 seconds. An employee is instructed to observe the lights and record the length of time between consecutive green lights.

(a) What is the variable of interest?
(b) What is the sample space?
(c) Can the classical approach to probability be used to determine \( P(\text{less than } 60 \text{ seconds between green lights}) \)? Why or why not?
(d) What about the relative frequency approach? If so, how? If not, why not?

Ex. 1.5 On October 13, the American television programme *Nightline* interviewed Dr Richard Andrews, director of the California Office of Emergency Services. They were discussing various natural disasters that were being predicted as a result of the El Niño. This is part of their interview:

*Ted Koppel:* Dr Andrews, I’m sure you have heard such cautionary advice before, so on what basis is the assumption being made that this is the one that’s going to have the kind of impact on southern California... that’s being predicted?

*Richard Andrews:* Well, in the business that I’m in and that local government and state government is in, which is to protect lives and property, we have to take these forecasts very seriously... I listen to earth scientists talk about earthquake probabilities a lot and in my mind every probability is 50–50, either it will happen or it won’t happen...

Explain *why* Dr Andrews is incorrect. Give an example to show why he *must* be incorrect. (Chance News 6.12.)

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency of self-examination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monthly</td>
</tr>
<tr>
<td>&lt; 45</td>
<td>91</td>
</tr>
<tr>
<td>45–59</td>
<td>150</td>
</tr>
<tr>
<td>60+</td>
<td>109</td>
</tr>
</tbody>
</table>

Table 1.6: The frequency of breast self-examination according to age at diagnosis of breast cancer.
Module 1. Probability basics

Table 1.7: Aviation deaths of private pilots in Australia from 1997 to 1999 according to the pilot’s age.

<table>
<thead>
<tr>
<th>Age</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 30</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>30 – 49</td>
<td>5</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>≥ 50</td>
<td>9</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

Ex. 1.6 The data is Table 1.7 was part of the data collected by Dr Peter Ruscoe in an investigation into aviation deaths of private pilots in Australia.

(a) What is the probability that a randomly chosen death in 1997 was of a pilot 50 or older?

(b) What proportion of deaths from 1997 to 1999 were of pilots aged under 30?

(c) What other information may be useful in studying the effect of age on pilot deaths?

Ex. 1.7 A bag contains 21 balls, which are either blue or red. Two balls are drawn, without replacement, from the bag.

(a) Define the sample space for the experiment.

(b) If it is known that the probability of a blue ball being drawn followed by a red ball is 7/30, determine the number of red balls in the bag.

(c) What would happen if the answer to the previous part was determined assuming that the balls were drawn with replacement?

Ex. 1.8 International practice is for electronic mail (or e-mail) addresses to conclude with a two letter code indicating the country of origin. E-mail addresses in Australia, for example, end with au, and so e-mail addressed to Queensland government departments will often have the form persons-name@department-name.qld.gov.au.

(a) How many countries can be accommodated in such a scheme?

(b) Do you think this is a sufficient number?

(c) To avoid potential confusion between similar country codes, suppose it was decided that countries could not use the same two
letters in their code as any other country. (That is, no country would be allowed to use the suffix `ua`, as it uses the same two letters as Australia’s code.) How many countries can be accommodated in this scheme?

**Ex. 1.9** At the Cricket World Cups, each cricket squad consist of fifteen players from which a team of eleven must be chosen for each game. The squad consists of five bowlers, seven batsmen, two all-rounders and one wicketkeeper.

(a) Find the number of teams possible if the team consists of five batters, four bowlers, one all-rounder and one wicketkeeper.

(b) After a game concludes, each member of one team shakes hands with each member of the opposing team, and each member of both teams shakes hands with the two umpires. How many handshakes are there in total at the conclusion of a game?

**Ex. 1.10** Suppose I am dealt five cards from a pack of cards (as in poker). Explain why the number of possible hands is \( \binom{52}{5} \) and not \( 52 \times 51 \times 50 \times 49 \times 48 = 50P_5 \).

**Ex. 1.11** (Computer exercise) Use R to simulate the gameshow described in Section 1.8.1.

**Ex. 1.12** (Computer exercise) Use R to simulate the experiment described in Section 1.9.

## 1.12 Exercises

**Ex. 1.13** For any two events \( A \) and \( B \) in the sample space \( S \) prove (a), (b) and (c), stating condition(s) for the equality sign to hold.

(a) \( P(A \cup B) \leq P(A) + P(B) \)

(b) \( P(A \cap B) \leq P(A) \)

(c) \( P(A \cup B) \geq P(A) \)

(d) If \( P(A) = \frac{2}{3} \) and \( P(A \cup B) = \frac{3}{4} \) what are the upper and lower bounds on the value of \( P(B) \).

**Ex. 1.14** Given an experiment such that \( P(A) = \frac{2}{3} \), \( P(B) = \frac{1}{4} \), \( P(A \cup B) = \frac{5}{6} \), evaluate the following algebraically using set identities and the rules of probability.
Module 1. Probability basics

(a) \( P(\overline{B}) \)
(b) \( P(A \cap B) \)
(c) \( P(A \cap \overline{B}) \)
(d) \( P(\overline{A} \cap \overline{B}) \)
(e) \( P(A \cup B) \)
(f) \( P(A \cup \overline{B}) \)

Ex. 1.15 If a box contains 100 tickets numbered from 1 to 100 inclusive and four tickets are drawn from the box one at a time (without replacement), find the probability that
(a) all four numbers drawn are even,
(b) exactly two even numbers are drawn,
(c) at least two odd numbers are drawn before drawing the first even number, (at least one even number to be drawn).
(d) the sum of the numbers drawn is odd.

Ex. 1.16 Prove that if \( A \) and \( B \) are events with non-zero probabilities then if \( A \) and \( B \) are independent they are not mutually exclusive.

Ex. 1.17 An ordinary die is rolled twice and the number of of spots on the upper face is observed after each roll. The events \( A, B \) and \( C \) are defined as
\( A \): The number of spots showing on the first roll is 3.
\( B \): The sum of the number of spots in the two rolls is 6.
\( C \): The sum of the number of spots in the two rolls is 7.
Show that \( A \) and \( B \) are not independent but that \( A \) and \( C \) are independent. Are \( B \) and \( C \) independent?

Ex. 1.18 In a two person game a fair die is thrown in turn by each player. The player who throws a six first wins.
(a) Find the probability that the first player to throw the die wins.
(b) Show that if the player to throw first is selected by the toss of a fair coin then each player has an equal chance of winning.

Ex. 1.19 In order to get honest answers to sensitive questions the randomised response technique is sometimes used. For example, suppose the aim is to discover the proportion of students who have used illegal drugs in the past twelve months. To do this we prepare \( N \) cards. On \( m \) of these is the statement ‘I have used an illegal drug in the past twelve months’. On the remaining \( N - m \) cards is the statement ‘I have not used an illegal drug in the past twelve months’.
A number of students are now chosen at random from all the students available. Each student so chosen then selects one card at random from the prepared pile of \( N \) cards and answers yes or no to the question ‘Is the statement on the card you have selected true?’ without
Table 1.8: The results in the second generation of Bateson’s experiment crossing a pure strain of peas with purple flowers and long pollen, with a pure strain of peas with red flowers and round pollen.

<table>
<thead>
<tr>
<th></th>
<th>Purple</th>
<th>Red</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long pollen</td>
<td>1528</td>
<td>117</td>
<td>1645</td>
</tr>
<tr>
<td>Round pollen</td>
<td>106</td>
<td>381</td>
<td>487</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1634</td>
<td>498</td>
<td>2132</td>
</tr>
</tbody>
</table>

divulging which statement is actually on the card. If \( P(Y) \) is the probability that a student answers yes, and \( P(D) \) is the probability that a student chosen at random has used an illegal drug and assuming that each student answers truthfully,

(a) find \( P(Y) \) in terms of \( P(D) \), \( m \) and \( N \) stating any conditions required on \( m \) and \( N \).

(b) If in a sample of 400 students, 175 answer yes so that the estimate of \( P(Y) \) is \( \frac{175}{400} \), estimate \( P(D) \) from the expression found above, given that \( N = 100 \) and \( m = 25 \).

**Ex. 1.20** A multiple choice question contains \( m \) possible choices. There is a probability of \( p \) that a candidate chosen at random will know the correct answer (and probability \( q = 1 - p \) that this candidate will not know the answer). If a candidate does not know the answer, the candidate guesses and is equally likely to select any of the \( m \) choices. If a candidate is randomly selected what is the probability of the question being answered correctly?

**Ex. 1.21** Bateson [5] repeated many of the famous experiments of Mendel (see Mendel [21]). One of these experiments involved crossing a pure strain of peas with purple flowers and long pollen, with a pure strain of peas with red flowers and round pollen. In the second generation, the data in Table 1.8 were collected.

(a) Determine \( P(\text{red flower}) \) from the data.

(b) Determine \( P(\text{round pollen}) \) from the data.

(c) Do the data suggest the colour of the pollen is independent of the shape? Explain.

(d) Determine \( P(\text{red} \cap \text{long}) \).

(e) Determine \( P(\text{red} \cup \text{long}) \).
A computer system requires all users to have passwords, which must be made up entirely of the letters a to z. Capital letters are considered to be different than lower case letters. (For example, the passwords aaaaaa, AAAAAA, aaaaAA and AaaAAa are all considered different passwords.)

(a) Determine the total number of passwords that are possible if every password must contain exactly six letters.

(b) Determine the total number of passwords possible if every password must be at least four letters long, and at most six letters long.

(c) Determine the total number of passwords if no letter can be repeated in any one password. (The passwords must still be at least four and at most six letters long, and lower case and upper case letters are still considered different.)

Ex. 1.23 Table 1.9 displays data from a cross-classification of coronary heart disease (CHD) by coffee drinking and smoking habits in a case-control study of 66 CHD cases and 85 unmatched controls. The results are all for men aged 40 to 55. Heavy smokers are defined as those individuals who smoke one pack per day or more. The data are from Paul et al. [25].

(a) What is the probability that a randomly selected individual is a heavy coffee drinker?

(b) What is the probability that a randomly chosen individual is a heavy coffee drinker and a heavy smoker?

(c) Suppose you meet one of the men in the study whom you know does not smoke. What is the probability this man will have had a CHD problem?

(d) Could the data be used to conclude that coffee use causes CHD?

(e) Determine if coffee use and CHD cases appear to be independent, justifying your answers with appropriate probability calculations.
Ex. 1.24 Consider the following argument.

When I toss a coin twice, there are only three outcomes: a Head and a Head; a Tail and a Tail; or one of each. So the probability of obtaining two Tails must be one-third.

Explain *why* the reasoning is incorrect, and then determine the correct answer.

Ex. 1.25 Consider the following experiment: A die is loaded so that the chance of obtaining a 1 is three times the chance of obtaining a 2. The chance of obtaining a 2 is the same as obtaining a 3, the same as obtaining a 4, the same as obtaining a 5, the same as obtaining a 6. The experiment consists of rolling the die once and observing the uppermost face.

(a) List the sample space for the experiment.
(b) Give the definition for the classical approach to probability. Is it possible to use the classical approach to probability to determine \( P(\text{a 1 is rolled}) \)? If so, determine the probability. If not, why not?
(c) Is it possible to use the relative frequency approach to probability to determine \( P(\text{a 1 is rolled}) \)? If so, how? If not, why not?
(d) Is it possible to use the three axioms of probability to determine \( P(\text{a 1 is rolled}) \)? If so, how? If not, why not?
(e) A class of students is given the die without being told it is loaded and is asked to determine \( P(\text{a 2 is rolled}) \) by repeated rolls of the die. One group of students repeats the experiment 100 times and records a 2 on 15 occasions. Do you think they have evidence that the die is loaded? Why or why not?

Ex. 1.26 Prove, using only the three axioms of probability, that \( P(A) \leq 1 \) for any event \( A \).

Ex. 1.27 Partridge [24, p 21] gives the data in Table 1.10 showing the relationship between the southern oscillation index (SOI) and rainfall in eastern Australia. The proportions can be considered as historical probabilities.

(a) What method has been used to compute the probabilities given in the table?
(b) Can the data be used to make predictions about the future? Explain.
Module 1. Probability basics

Table 1.10: The effect of the southern oscillation index (SOI) on rainfall in eastern Australia. The table shows the proportion of years that historically give significantly above and below average rainfall in years with the given average SOI.

<table>
<thead>
<tr>
<th>Average Annual SOI</th>
<th>Proportion of Eastern Australia receiving annual rainfall:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>below average</td>
</tr>
<tr>
<td>below −10</td>
<td>0.63</td>
</tr>
<tr>
<td>−10 to 10</td>
<td>0.25</td>
</tr>
<tr>
<td>above 10</td>
<td>0.03</td>
</tr>
</tbody>
</table>

(c) Are the SOI and rainfall in eastern Australia independent? Justify your answer. What does this mean in this context?

(d) Construct a tree diagram that summarises the data.

(e) The text in Partridge [24, p 20] implies that for 8 years, the average annual SOI was below −10, and that for 8 years the average annual SOI was above 10; for the remaining 53 years the average annual SOI was between −10 and 10.

Use this information to determine the following historical probabilities.

i. $P(\text{SOI below } -10 \cap \text{ below average rainfall in eastern Australia})$;

ii. $P(\text{SOI above } -10 \mid \text{ below average rainfall in eastern Australia})$;

iii. $P(\text{receiving below average rainfall in eastern Australia})$.

Ex. 1.28 The Brisbane newspaper, The Sunday Mail, gave details about the numbers of people who, after strip-searching, were subsequently charged. The data are from the twelve months from April 1998 and May 1999 for various districts in Queensland. The following data were deduced from the article for the Cairns region: the probability that a person was charged in Cairns after being strip-searched is 0.077 (which seems frighteningly low); the probability that the person was male and also charged after being strip-searched is 0.074; the probability that a person was male is 0.903.

Construct a Venn diagram to answer the following questions.

(a) Determine the probability that a person who was strip-searched was female.

(b) Determine the probability that a female who was strip-searched was then charged.
1.12. Exercises

(c) Determine the probability of selecting a male who was charged.
(d) Determine the probability of either a female or someone who was not charged.

Repeat the questions using a table rather than a Venn diagram.

Ex. 1.29 Here is an amazing true story about the Arizona Lottery (Chance News, 27 June 1998 to 8 August 1998 edition). The Arizona Lottery started a new game “Pick 3” on May 3, 1998. In Pick 3, three numbers are chosen from zero to nine (repetitions are allowed). Three ordered winning numbers are picked and prizes are awarded depending on how the numbers match the winning numbers. New winning numbers are picked every day of the week except Sunday.

After 32 games, no nines had been drawn in any of the 96 winning numbers. One woman, who always chose 9-0-7, called to complain that there must be a problem. It was later discovered that the company supplying the random number generator provided a program that did not select nines.

About 1.2 million tickets with nines had been sold during this period. To compensate, lottery officials ran the same game for the period July 15 to July 31 but with all the prizes doubled. They also went back to the old-fashioned method of using numbered balls to determine the winning numbers.

(a) Determine the probability that 96 numbers are selected from zero to nine without selecting a nine. When do you think the error could reasonably have been detected?

(b) When you play the game for $1, three of the options available to play are Exact Order, Front Pair and Back Pair. Three ordered winning numbers are then drawn. If you choose “Exact Order” and your numbers agree with the winning numbers in their chosen order, you win $500. If you chose “Front Pair” and your first two numbers match the first two of the winning numbers in their chosen order, or you chose “Back Pair” and your last two numbers match the last two of the winning numbers in their chosen order order, you win $50. Find the expected winnings under each choice. Were any of the choices favourable games when the prizes were doubled?

(c) How much do you think the doubling of prize money affected the number of people who played the game?

Ex. 1.30 The data in Table 1.11 were collected by L.A. Goodman [13]. They concern eye and hair colour of 5387 children in the Scottish town of Caithness.
Module 1. Probability basics

<table>
<thead>
<tr>
<th>Hair Colour</th>
<th>Fair (HF)</th>
<th>Red (HR)</th>
<th>Medium (HM)</th>
<th>Dark (HD)</th>
<th>Black (HB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue (EB)</td>
<td>326</td>
<td>38</td>
<td>241</td>
<td>110</td>
<td>3</td>
</tr>
<tr>
<td>Light (EL)</td>
<td>688</td>
<td>116</td>
<td>584</td>
<td>188</td>
<td>4</td>
</tr>
<tr>
<td>Medium (EM)</td>
<td>343</td>
<td>84</td>
<td>909</td>
<td>412</td>
<td>26</td>
</tr>
<tr>
<td>Dark (ED)</td>
<td>98</td>
<td>48</td>
<td>403</td>
<td>681</td>
<td>85</td>
</tr>
</tbody>
</table>

Table 1.11: Data for Exercise 1.30

(a) Determine the following probabilities.
   i. P (Blue eyes);
   ii. P (Dark or black hair);
   iii. P (HF ∩ ED);
   iv. P (HM | EB).

(b) Explain the meaning of each of the following in words and in the context of the question (you do not have to determine probabilities).
   i. HF ∩ EB;
   ii. EM ∪ EL;
   iii. (HF ∪ HR) | ED.

(c) Does it appear that hair colour and eye colour are independent? Justify your answer.

Ex. 1.31 A rugby league team of 13 consists of five backs, six forwards and two halves.

(a) How many teams are possible from a squad of 17, consisting of six backs, eight forwards and three halves?

(b) How many teams are possible from a squad of 17, consisting of six backs, eight forwards and two halves, and one player who could play as a back or a half?

Ex. 1.32 In Gold Lotto, players pick any six numbers from the numbers one through to 45 (inclusive). Six winning numbers are then drawn (without replacement), and players who have selected the same six numbers win the first prize. (Two “supplementary numbers” are also drawn, but we will not be concerned with them.) To play, players need to play a minimum of four games, costing $1.75 (at the time of writing). Suppose the McMillan family play the minimum four games every week.
(a) What is the probability of the McMillans correctly choosing the six winning numbers?

(b) The expected payoff is the amount one can expect to win or lose for each game played in the long term. It is calculated as

\[
\text{Expected Payoff} = P(\text{win}) \times \text{(Net amount won)} + P(\text{lose}) \times \text{(Net amount lost)}.
\]

Suppose the first prize is one million dollars ($1,000,000). (This may not be true—the first prize changes weekly.) What is the expected payoff from playing Gold Lotto?

Ex. 1.33 In the game of Yahtzee, players roll five six-sided dice and are awarded points for various combinations that appear. Assuming the five dice are fair, determine the following.

(a) The probability that all five dice show the same number (for example, 2, 2, 2, 2, 2).

(b) The probability of obtaining three dice with the same number, and the remaining two dice as the same number, but different to what is on the other three (a “full house”; for example, 3, 3, 3, 6, 6.)

(c) The probability of obtaining one pair. (That is, two dice have the same number, but the other three are all different than each other, and also different than the pair; for example 1, 1, 3, 5, 4).

(d) The probability of obtaining a “large (or long) straight”. (That is, five consecutive numbers, like 1, 2, 3, 4, 5.)

Ex. 1.34 In Orleans Parish, Louisiana, the law states juries must be selected by jury commissioners at random from a list compiled from the general population. In 1958, the list for one case consisted of 30 black people and 40 white people. Twelve people were chosen for the jury, none of whom were black.

(a) The commissioners claimed that the jury was selected according to the law (that is, at random). Given that this is true, what is the probability that such a jury was selected?

(b) Do you believe the jury really was selected at random? (Justify your answer.)

Ex. 1.35 In drug testing, there are two keys aspects of the test:

- the sensitivity of the test, defined as:
  \[ P(\text{Positive test result} \mid \text{person actually has the disease}); \]
• the specificity of the test, defined as:
  \[ P(\text{Negative test result } | \text{ person actually doesn’t have the disease}) \].

In testing for HIV, the standard test is the Wellcome Elisa test, commonly called the Elisa test. For the Elisa test, the sensitivity is approximately 0.993 and the specificity approximately 0.9999, according to the US Department of Food and Drug Administration. In a research paper (Johnson and Gastwirth [17]), the incidence of HIV-positive in the general population is estimated to be about 0.000025.

(a) Explain why the specificity is important. What does this tell you about the Elisa test?
(b) Explain why the sensitivity is important. What does this tell you about the Elisa test?
(c) Explain how the probabilities for specificity and sensitivity have most likely been found.
(d) Explain why \( P(\text{a person has HIV } | \text{ the test is positive}) \) is important.
(e) Explain why \( P(\text{a person has HIV } | \text{ the test is negative}) \) is important.
(f) Determine \( P(\text{a person has HIV } | \text{ the test is positive}) \).
(g) Determine \( P(\text{a person has HIV } | \text{ the test is negative}) \).
(h) Discuss the practical consequences of your answers to parts (f) and (g) for someone who has been given the Elisa test.
(i) Suppose you work as a statistician for a local hospital. The administrator is considering compulsory HIV testing for all new employees, by using the Elisa test. The administrator, Dr Tam, has asked you to write a one page report on the proposal from a statistical point-of-view. Your report should be easily understood by Dr Tam (who is a medical expert, but not an expert on statistics), and should focus only on your conclusion from a statistical point-of-view. (That is, she will get others’ opinions on costs, medical aspects, privacy concerns, and so on; your contribution is the statistical component only.) You must write this report like a real report: addressed to Dr Tam, typed wherever possible, using proper language and grammar. This question will be marked as if you were submitting in a real job. While you are asked to submit a one-page report, you can include one extra page of calculations if you wish.

Before answering the question, consider what issues are important to Dr Tau and to the individual staff members who have been given the test, and what probabilities help address these issues. These issues should be listed somewhere in the report.
1.12. Exercises

Ex. 1.36 Consider the following events concerning the type of computer a customer purchases: A: orders a 17" monitor; B: orders a computer with a ergonomic keyboard; C: orders a computer with a modem. Write in words what is meant by the following (don’t determine probabilities).

(a) $A \cap B$
(b) $A' \cup C$
(c) $A' \cap B \cap C$
(d) $C | A$
(e) $(B \cap C) \cup A'$

Ex. 1.37 Given two independent events $C$ and $D$, prove that $\overline{D}$ and $C$ are not mutually exclusive.

Ex. 1.38 An experiment involves tossing a coin and rolling a standard die.

(a) List the sample space.
(b) Determine the probability of tossing a Head and rolling a 2.
(c) Determine the probability of rolling a 5 given that a Head has been tossed. What is the reduced sample space in this situation?

Ex. 1.39 The following problem appeared in an American newspaper.

A woman and a man (unrelated) each have two children. At least one of the women’s children is a boy, and the man’s older child is a boy. Do the chances that woman has two boys equal the chances that the man has two boys?

One newspaper reader wrote in

I will send $1000 to your favourite charity if you can prove me wrong. The chances of both the woman and the man having two boys are equal.

What is the answer to this problem? If the reply is incorrect, what was the misunderstanding? (From Chance News 6.12, October 10, 1997 to November 9, 1997.)

Ex. 1.40 The following item is taken from Shen [29]. It is known that oceans cover more than one half of the earth’s surface. The following probabilistic argument has been used to prove that there are at least two symmetric points covered by water. (‘Two symmetric points’ are two points on directly opposite sides of the globe.)
(a) Choose a point $X$ on the surface of the earth at random. Let $X'$ the point antipodal to $X$ (on the opposite side of the globe). Determine $P(X \text{ is in the ocean})$ and $P(X' \text{ is in the ocean})$.

(b) Write an addition formula for computing $P(X \text{ or } X' \text{ is in the ocean})$.

(c) Use this formula to deduce that $P(X \text{ and } X') > 0$ for the addition formula to make sense.

**Ex. 1.41** In a given situation, the sample space is given as $S = \{1, 2, 3, 4, 5\}$. The probabilities of obtaining these numbers from the sample space are only known to be:

$$
P(1) = a, \quad P(2) = 0.2 - a, \quad P(3) = b, \quad P(4) = 2b, \quad P(5) = 0.8 - 3b,
$$

where $a > 0$ and $b > 0$. It is also known that the events “obtaining 1, 2 or 3” and “obtaining a 3 or 4” are independent.

(a) Confirm that the sample space consists of the set $S$ as given above.

(b) Use the information above to solve for $b$.

(c) If it is also known that the events “obtain a 1 or 5” and “obtain a 1 or 3” are also independent, show that $a$ can take two possible values.

**Ex. 1.42** Sleigh et al. [30] compared two methods for detecting *Schistosoma mansoni* eggs in faeces: Bell’s method and the Kato–Katz method. Let $B$ refer to the event “Bell’s test returned positive” and $K$ refer to the event “the Kato–Katz test returned positive”. Then, the following results were obtained from the study.

$$
P(B) = 0.756; \quad P(K) = 0.629; \quad P(B \cap K) = 0.584.
$$

(a) Explain in simple terms what the probability $P(B \cap K)$ actually means.

(b) Determine the probability that neither test was positive.

(c) Compute $P(B \mid K)$.

(d) Compute $P(B \cup K)$.

(e) Would you expect the two test to be independent? Explain. Perform some calculations to confirm.

**Ex. 1.43** The following information is taken from the abstract of Baron et al. [3].
We performed a randomized, double-blind trial of aspirin as a chemopreventive agent against colorectal adenomas. We randomly assigned 1121 patients with a recent history of histologically documented adenomas to receive placebo (372 patients), 81 mg of aspirin (377 patients), or 325 mg of aspirin (372 patients) daily. Follow-up colonoscopy was performed at least one year after randomization in 1084 patients (97 percent). The incidence of one or more adenomas was 47 percent in the placebo group, 38 percent in the group given 81 mg of aspirin per day, and 45 percent in the group given 325 mg of aspirin per day.

Of the original 1121 patients, 1084 remain in the study. Assume that the $1121 - 1084 = 37$ people who pulled out left 360 in the placebo group, 364 in the 81 mg aspirin group, and 360 in the 325 mg aspirin group (that is, left the study randomly).

(a) Determine the proportion of people in the study who were found to have an adenoma.

(b) Given that a person has an adenoma, what is the probability that the person was in the placebo group?

(c) What is the probability that a person had no adenomas and was in the 325 mg of aspirin group.

(d) Do you think the aspirin had any effect on the occurrence of adenomas? Why or why not?

Ex. 1.44 (Computer exercise) Simulate the situation in Exercise 1.19 when the proportion of students taking an illegal drug is 30%. (Note: The advantage of simulation is that we know what the answer should be.)

(a) The R script is listed below. The first few lines allocate numerical values to parameters prior to generating the random samples.

```R
# lines like this are comments - no need to type them
# set the parameter values
p <- 0.3 # assumed prob of a person using illegal drugs
sample.size <- 400
Ncards <- 100 # size of card deck
m <- 25 # no of cards which state that ‘I have used ..’
# generate the card deck
card.deck <- c(rep(0,(Ncards-m)),rep(1,m))
# Draw a sample of students
incidence.sample <- sample(c(0,1),size= sample.size, replace = TRUE,prob=c((1-p),p))
```
# Draw a sample from the card deck
answer.sample <- sample(card.deck, size=sample.size, replace=T)

# match the students with their answers
confusion.tab <- table(incidence.sample, answer.sample)

print(confusion.tab)

(b) Estimate the proportion of students using drugs by using the result of 1.19. Is the result reasonable?

(c) Try the simulation again but vary the proportion of students taking drugs and both $N$ and $m$.

Ex. 1.45 (Computer exercise) Simulate the situation in Exercise 1.20 when 60% of students know the correct answer.

(a) First generate a sample of 100 students (1 = student knows answer, 0 = student doesn’t know answer) randomly chosen from a population in which 60% of students know the correct answer. (Use the sample() function as in the previous question to generate a sample from a discrete distribution. Count the number of one’s in your sample. (These know the answer and will answer correctly.) Use the number of zeros as the sample size for generating the random sample in the next step.

(b) Now generate a random sample for the students who did not know the answer. Since these students select the answers 1 to 4 at random they have a probability of 25% of getting the correct answer. That is to sample 1 and 0 with the probabilities 0.25 and 0.75 respectively. Count the number of 1’s in the random sample (the number of correct guesses)

(c) Find the total number of correct answers obtained (‘Know’ plus ‘guessed!’) to get your estimate.

(d) How did it compare with the theoretical answer?

1.13 Some answers and hints

1.1 You can use Venn diagrams to make the formulae clearer.

(a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.56$.

(b) $P(\overline{A} \cap B) = P(B) - P(A \cap B) = 0.26$.

(c) $P(\overline{A} \cup \overline{B}) = P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cap \overline{B}) = 0.76$ or $1 - P(A \cap B) = 0.76$. 

1.13. Some answers and hints

(d) \( P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.24}{0.50} = 0.48 \).

(e) Since \( P(A) \times P(B) = 0.3 \times 0.5 = 0.15 \neq P(A \cap B) \), \( A \) and \( B \) are not independent events.

1.2 (a) All real numbers are possible values for \( a \), \( b \) and \( c \); hence \( S = \{(a, b, c) \mid -\infty < a < \infty, -\infty < b < \infty, -\infty < c < \infty\} \).

(b) The solutions to a quadratic equation are given by

\[
x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.
\]

For two equal roots, \( b^2 - 4ac = 0 \). So event \( A \) is defined as \( A = \{(a, b, c) \mid b^2 - 4ac = 0\} \).

1.3 \( \frac{232}{1216}; \frac{350}{1216}; \frac{51}{1216}; \frac{(150 + 200 + 155 + 51 + 172)}{1216} = \frac{728}{1216} \approx 0.599; \frac{155}{378}; \) The variables are not independent (e.g. \( P(\text{age between 45 and 59} \mid \text{never use self-examination}) = 0.599 \) but \( P(\text{age between 45 and 59}) = \frac{505}{1216} \approx 0.415 \).

1.4 (a) The length of time (in seconds) between green lights at the intersection, say \( G \)

(b) \( S = \{G \mid 15 \leq G \leq 120\} \)

(c) No—no equally likely events are defined

(d) The probability can be approximated—observe the lights a large number of times and count how often there is less than 60 seconds between green lights.

1.5 He is erroneously assuming that the events ‘an earthquake occurs’ and ‘an earthquake does not occur’ are equally likely events.

1.6 \( \frac{9}{16}; \frac{6}{57}; \) pilot experience, type of flight (solo, etc.), type of aircraft, etc.

1.7 \( S = \{(B, B); (B, R); (R, B); (R, R)\} \) where \( B \) refers to drawing a blue ball and \( R \) to drawing a red ball; \( 7 \) or \( 14 \) red balls; \( 13.2111 \) or \( 7.7889 \), neither of which are possible.

1.8 \( 26^2 \); There are under 200 sovereign countries, so should be sufficient (but takes no account of whether there are enough sensible codes for each country); \( C_2^6 \times 25 \).

1.9 \( C_2^4 \times C_5^4 \times C_7^2 \times C_1^1 \); \( 11^2 + (2 \times 22) = 165 \).

1.10 Because the order is not important (after getting the cards, they can be arranged in any order in your hand).
Module 1. Probability basics

1.11 Step 1: Generate a random sequence of length 100 consisting of the integers 1, 2 and 3, putting the sequence in column C1 to represent the position of the car.

Step 2: Generate another such sequence (and put in column C2) to represent the contestants first choice on each of the 100 nights.

Step 3: Count the number of times the numbers in the two column agree representing the number of times the contestant will win if the contestant doesn’t change doors. If the numbers in the two columns don’t agree then the contestant will win only if the contestant decides to change doors.

\[
\text{car.door<-sample(1:3,size=100,replace=T)}
\]
\[
\text{first.choice<-sample(1:3,size=100,replace=T)}
\]
\[
\text{confusion.table<-table(car.door,first.choice)}
\]
\[
\text{agree<-sum(diag(confusion.table))}
\]

One simulation produced the confusion table

<table>
<thead>
<tr>
<th>first.choice</th>
<th>car.door</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>10</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Notice contestants would have won the car 30 times if they retained their first choice which means they would have won 70 times if they had changed.

The apparent paradox is explained by the fact that the contestant did not know about the second chance at the start.

1.12 sim.results<-numeric(0) # names a NULL vector ready to collect results
for ( i in 1:1000){
    # number of 1’s in sample of 21 (0,1)
    Ncond1 <- sum(sample(0:1,size=21,replace=T))
    sim.results<-c(sim.results,Ncond1) } # accumulate incidences of 1’s
print(table(sim.results)) # frequency of counts out of 21

1.13 (a) Use Theorem 1.11 noting \( P(A \cap B) \geq 0 \). (What is the condition on \( A \) and \( B \) for equality?)

(b) First write \( A \) as union of two mutually exclusive events; that is, \( A = (A \cap B) \cup (A \cap \overline{B}) \).

(c) Note \( A \cup B = A \cup (B \cap \overline{A}) \).

(d) \( 1/12 \leq P(B) \leq 3/4 \).
1.13. Some answers and hints

1.14 Note using Venn diagrams alone is not acceptable.
(a) 3/4  (b) 1/12  (c) 7/12  (d) 1/6  (e) 11/12  (f) 5/6

1.15  
(a) \[ \frac{188}{3201} \]
(b) \[ \frac{1225}{3201} \]
(c) \[ P(2 \text{ odd before 1st even}) = P(\text{odd,odd,even,odd or even}) + P(\text{odd,odd,odd,even}) = \frac{3625}{19206} \]
(d) \[ P(\text{sum odd}) = P(\text{1 odd number OR 3 odd numbers drawn}) = \frac{1600}{3201} \]

1.16 If \( A \) and \( B \) are independent then \( P(A \cap B) = P(A)P(B) > 0 \). Hence show condition for \( A \) and \( B \) mutually exclusive not satisfied.

1.17 Apply Definition 1.16.

1.18 \( P(\text{Player throwing first wins}) = P(\text{First six on throw 1 or 3 or 5 or ...}) = P(\text{First six on throw 1}) + P(\text{First six on throw 3}) + \cdots \) and sum the geometric progression obtained.

Use Theorem 1.20 for the second part. Define the events \( A = \{\text{Player 1 wins}\} \), \( B_1 = \{\text{Player 1 throws first}\} \), \( B_2 = \{\text{Player 1 throws second}\} \).

1.19  
(a) Use Theorem 1.20 to find \( P(Y) \).
\[ Y = \{\text{Yes}\}, \ D = \{\text{Student has used illegal drugs in last 12 months}\}, \ D' = \{\text{student hasn’t used illegal drugs in last 12 months}\}. \]

(b) \[ P(D) = \frac{NP(Y) - \left[N - m\right]}{2m - N} \]

1.20 Use Theorem 1.20 to find \( P(C) \) where \( C = \{\text{correct answer}\} \), \( K = \{\text{student knows answer}\} \). Ans: \( P(C) = \frac{mp + q}{m} \)
Module 2

Distribution of random variables

Contents

2.1 Introduction .......................... 58
2.2 Random variables .......................... 60
2.3 Univariate distributions .......................... 63
   2.3.1 The probability function .......................... 63
   2.3.2 The probability density function .......................... 67
   2.3.3 The distribution function .......................... 70
   2.3.4 Comparing discrete and continuous distributions .......................... 73
2.4 Probabilities of compound events for discrete r.v.’s .......................... 74
   2.4.1 Mixed random variables .......................... 76
2.5 Self-assessment exercises .......................... 76
2.6 Exercises .......................... 78
2.7 Some answers and hints .......................... 83

Module objectives

Upon completion of this module students should be able to:
• understand the concept of both univariate and bivariate random variables

• distinguish between discrete, continuous and mixed random variables

• define the probability function of a random variable and different random variables for a given experiment, both for discrete and continuous cases

• determine the probability function of random variables defined for an experiment

• determine the distribution function of a random variable from its probability function and apply it to the computation of probabilities of defined events

• plot the probability function and distribution function of a random variable

• understand the concept of a mixed random variable and manipulate its probability function

\[ \text{2.1 Introduction} \]

In Module 1, the tools of probability were used to assist in answering questions about phenomena involving uncertainty. The sample space was defined as the set of all possible outcomes and an event was defined as a subset of the sample space. Probabilities were associated with events. The outcomes themselves could be numeric, as in tossing a die, or non-numeric, as in tossing a coin.

Interestingly it turns out that it is advantageous to restrict ourselves to numerical outcomes from experiments. This may sound unnecessarily restrictive, but is not. After all, we are quite at liberty to number all the outcomes in the sample space, thereby converting non-numeric elements to numeric ones. So, for example, heads could be denoted 1 and tails 2 in the sample space associated with tossing a coin.

This might sound a bit unnecessary, and in the case of coin tossing it probably is. However by restricting ourselves to numeric outcomes we can make use of the powerful concepts that are associated with numbers. In particular the concepts of discreteness, continuousness, types of infinity, limits, etc are available. So what might appear to be a restrictive step is in reality the key to unlocking some powerful ideas and results that make probability theory and in turn, statistical theory, a force to be reckoned.
Of course, in most scientific, engineering and business applications, our interests are in assigning probabilities to numeric outcomes. For example,

- Probabilities can be assigned to different amounts of money made by an investment
- Probabilities can be assigned to the proportion of people that contract a given disease
- Probabilities can be assigned to observing different water levels at a certain river crossing
- Probabilities can be assigned to matching DNA sequences when their genome sequences are almost identical
- Probabilities can be assigned to recording certain rainfall amounts.

The point however is that some experiments give rise to just a limited number of outcomes (eg tossing a die), some give rise to an unlimited number of discrete outcomes (eg the number of attempts needed to win a lottery) and some give rise to a range of possible outcomes over some interval (eg water levels, rainfall, tomorrow’s maximum temperature). The differences amongst these types of sample spaces is what drives us to thinking in terms of discrete numbers (such as integers) and continuous numbers (such as the reals).

To make the connection we define something called a \textit{random variable} on the sample space. This is simply a way of converting the elements of the sample space into numbers. Depending on the type of sample space, this random variable can be discrete, continuous, or a mixture of both. Regardless, it is useful and important to be able to describe the possible values of the random variable and the probability associated with these values. This is called the \textit{distribution of the random variable} and is the concept underlying the name of this course.

As an example, when a fair die is rolled, the classical approach to probability can be used to determine the theoretical probabilities \(P(\text{roll a 1}) = P(\text{roll a 2}) = \cdots = P(\text{roll a 6}) = 1/6\). If we define a random variable \(X\) in this experiment as the number of pips showing on the die then we see, for example, that \(X = 1\) for the outcome 1. Consequently we have \(P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6\) and this describes the distribution of the random variable \(X\) for this experiment.

This random variable \(X\) is discrete and the distribution of \(X\) is an example of a \textit{discrete uniform distribution} for obvious reasons. We learn more about this distribution in Section 4.1.
When a die is rolled repeatedly, and the relative frequency is determined for each value of $X$, we don’t of course get $1/6$ for each. We recognise the discrete uniform distribution is a theoretical distribution based on the principle of indifference as described in section 1.3.2. The observed distribution of relative frequencies based on actual dice tossing will only approximate the theoretical distribution. How well the two match we would be justified in thinking depends on how many times the die is tossed and how balanced it is. This is an investigation we pursue later.

The theoretical distribution is said to model the experiment. Hence the discrete uniform distribution is a model for dice tossing. The classical approach to assignment of probabilities gives us a model in this case. For other experiments other models will be appropriate.

This module discusses the concept of the random variable and the distribution of probabilities amongst the values of a random variable. Further modules describe specific distributions useful in modelling experiments.

## 2.2 Random variables

Colloquially, a random variable is a numerical-valued variable. A random variable may be continuous, discrete or a combination of the two. Some examples follow the more formal definition below.

**Definition 2.1** A random variable $X$ is a real-valued function mapping each element of the sample space into a real number. The domain of a random variable is the sample space, and the range space or value set is the set of real numbers taken by function.

The random variable $X$ can be thought of as a rule assigning a number to each and every outcome in the sample space.

Using functional notation we see that a function $X$ assigning to every sample point $s \in S$ a real number $X(s)$ is a random variable. Further, since $X$ is a function, to every $s \in S$ there corresponds exactly one value $X(s)$. The domain of $X$ is the set $S$ and the range space is the set $\{X(s) \mid s \in S\}$.

Commonly, “random variable” is abbreviated to rv.

A discrete random variable (rv) is defined on a discrete sample space. Examples of discrete random variables include:
2.2. Random variables

- The number of mice in a cohort of twenty that develop an illness
- The number of rain events occurring during a week
- The amount of money in a person’s pocket
- The shoe size of a person
- The number of attempts needed to win Lotto
- The number of offspring successfully reared by an endangered species of bird
- The gender (after coding) of the next customer.

The last example might need explaining. The sample space for the associated experiment contains two points ‘female’ and ‘male’; ie \( S = \{ \text{female}, \text{male} \} \). A random variable though must be numeric so if ‘Gender’ is to be a random variable we must associate the outcomes in the sample space with numerical values; eg an arbitrary coding such as 0 for ‘female’ and 1 for ‘male’ achieves this. We can then write \( X(\text{female}) = 0 \) and \( X(\text{male}) = 1 \) where \( X \) denotes the random variable ‘Gender’. Notice the range space or value set of \( X \) is \( \{0, 1\} \) which we might denote as \( R_X \).

Often we dispense with formal functional notation when defining random variables and write, for example,

\[
X = \begin{cases} 
0 & \text{if female} \\
1 & \text{if male} 
\end{cases}
\]

Incidentally don’t confuse the definition of a random variable with the values of the random variable. In the last example, the rv ‘Gender’ can have two values, female or male coded as 0 or 1. ‘Male’ or 0 is not a rv; it is one value of the rv ‘Gender’.

Loosely speaking, a continuous random variable is one that can never be measured exactly. If a random variable can assume any value or a set of values in an interval, then it is called continuous. For example, consider your height. Your height could be recorded as 174 cm, but with better measuring instruments it may be 174.023451006 cm, and even more decimals are possible with even better instruments! Even though height may be written down to the nearest centimetre or even millimetre, height itself is still a continuous measure. In addition, your height does not change by a distinct amount. My baby daughter grows continually; she doesn’t grow in jumps of 1 cm, for example! Examples of continuous random variables include:
• The volume of waste water treated at a sewage plant per day
• The amount of money made (or lost) by an investment over 12 months
• The weight of hearts in normal rats
• The lengths of the wings of butterflies
• The lifetime of diseased mice after taking an experimental drug
• The yield of barley from a large paddock
• The amount of rainfall recorded each year
• The time taken to perform a psychological test.

Mixed random variables are less common. Typically, they consist of a quantity that can be measured exactly (discrete) sometimes and inexacty (continuous) at other times. Consider, for example, the time spent waiting at a set of traffic lights before proceeding. If the light is already green when you arrive, you can drive straight through and have to wait exactly zero seconds (the discrete part). If the light is red when you arrive, you have to wait a continuous amount of time before it turns green and you go. Hence, the time spent waiting is a mixed random variable.

Example 2.1 The *Sunday Mail* newspaper (Brisbane, Australia) on 21 December 1997 gave data giving the birth weight, gender, and time of birth of 44 babies born in the 24-hour period of 18 December 1997 at the Mater Mother’s Hospital in Brisbane, Australia. The gender of the babies is a discrete random variable; the birth weight and time of birth are continuous. The data can be used to determine the number of births each hour (discrete) and the time between births (continuous).

Example 2.2 Consider tossing a coin twice and observing the outcome of the two tosses. Since a random variable is a *real-valued function*, simply observing the outcome as \{H, T\}, for example, is not a rv. We could define the rv of interest, say \(H\), as the *number* of heads on the two tosses of the coin. The *sample space* for the experiment is \(S = \{(TT), (TH), (HT), (HH)\}\). The connection between the sample space and \(H\) is shown in Table 2.1.

In this case, the range of \(H\) is \(R_H = \{0, 1, 2\}\).
2.3. Univariate distributions

<table>
<thead>
<tr>
<th>Element of $S$</th>
<th>Function $H$</th>
<th>Value of $H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(TT) = $s_1$</td>
<td>$H(s_1)$, the number of heads in $s_1$</td>
<td>0</td>
</tr>
<tr>
<td>(TH) = $s_2$</td>
<td>$H(s_2)$, the number of heads in $s_2$</td>
<td>1</td>
</tr>
<tr>
<td>(HT) = $s_3$</td>
<td>$H(s_3)$, the number of heads in $s_3$</td>
<td>1</td>
</tr>
<tr>
<td>(HH) = $s_4$</td>
<td>$H(s_4)$, the number of heads in $s_4$</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2.1: The random variable $H$ maps each element in the sample space $S$ to a real number.

Methods for describing random variables and how probabilities are assigned or distributed amongst the values of a random variable occupies much of this course.

One of the reasons for distinguishing between discrete and continuous random variables is that the ways we can describe the distribution of probabilities depend to some extent on this distinction.

2.3 Univariate distributions

Reading 2.2 WMS, Section 3.2.

This section focuses on methods of describing the distribution of probabilities amongst the values of a random variable. We need to consider separately discrete and continuous rvs.

2.3.1 The probability function

The probability function is a function that indicates how probabilities are assigned to the values of a discrete rv; that is, a rv with range space that is finite or countably infinite.

Definition 2.2 Let the range space of the discrete rv $X$ be $R_X$. With each $x \in R_X$ we associate a number

$$p_X(x) = P(X = x)$$

The function $p_X$ is called the probability function of $X$.

Some texts call the probability function, the probability mass function.
Definition 2.3 If \( R_X = \{x_1, x_2, \ldots, x_n, \ldots\} \), the set \( \{(x_i, p_X(x_i); \ i = 1, 2, \ldots\} \) is called the probability distribution of the discrete rv \( X \).

WMS and DGS denote the probability function as \( p(x) \) rather than \( p_X(x) \). Using the subscript is recommended to avoid confusion in situations where a number of random variables are on the go. So if we are dealing with the discrete rv’s \( X \) and \( Y \) we can denote their respective probability distributions by \( p_X \) and \( p_Y \) respectively. We use this subscript idea where appropriate throughout this course.

Essentially the probability distribution of a random variable is a description of the range space or value set of the variable and the associated assignment of probabilities. The above definition only applies to a discrete rv.

The probability distribution of a discrete rv \( X \) can be represented by a formula, a table or a graph which displays the probabilities \( p(x) \) corresponding to each \( x \in R_X \).

Example 2.3 Five balls numbered 1, 2, 3, 4, 5 are in a box. Two balls are selected at random. Find the probability distribution of the larger of the two numbers.

\[
S = \{(1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5)\}
\]

and all 10 points are equally likely.

Let \( X \) be the larger of the 2 numbers chosen. Then \( R_X = \{2, 3, 4, 5\} \) and

\[
\begin{align*}
P(X = 2) &= P((1, 2)) = 1/10 \\
P(X = 3) &= P((1, 3) \text{ or } (2, 3)) = 2/10 \\
P(X = 4) &= P((1, 4) \text{ or } (2, 4) \text{ or } (3, 4)) = 3/10 \\
P(X = 5) &= P((1, 5) \text{ or } (2, 5) \text{ or } (3, 5) \text{ or } (4, 5)) = 4/10.
\end{align*}
\]

This is the probability distribution of \( X \) but it can be expressed more neatly as a ‘formula’,

\[
P(X = x) = (x - 1)/10, \quad x = 2, 3, 4, 5,
\]

or shown in a table:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( P(X = x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.1</td>
</tr>
<tr>
<td>3</td>
<td>.2</td>
</tr>
<tr>
<td>4</td>
<td>.3</td>
</tr>
<tr>
<td>5</td>
<td>.4</td>
</tr>
</tbody>
</table>

or as a graph as in Figure 2.1.

Computer exercise Write an R function to allow you to simulate the distribution in this example.
2.3. Univariate distributions

Example 2.4 Suppose a fair coin is tossed twice. Then the *sample space* is \( S = \{ \text{HH, HT, TH, TT} \} \). Let \( H \) be the *number* of heads observed. \( H \) is a (discrete) random variable, and the range of \( H \) is \( R_H = \{0, 1, 2\} \), representing the values that \( H \) can take. The probability function maps each of these values to the associated probability. Using techniques from Module 1, the probabilities can be computed and given as a list:

\[
\begin{align*}
P(H = 0) &= P(\text{no heads}) = 0.25 \\
P(H = 1) &= P(\text{one head}) = 0.5 \\
P(H = 2) &= P(\text{two heads}) = 0.25.
\end{align*}
\]

This information can be presented as a table:

\[
\begin{array}{c|c|c|c}
  h & 0 & 1 & 2 \\
\hline
  P(H = h) & 0.25 & 0.5 & 0.25
\end{array}
\]

As a function, the probability function could be written as

\[
p_H(h) = P(H = h) = \begin{cases} 
0.25 & \text{if } h = 0 \\
0.5 & \text{if } h = 1 \\
0.25 & \text{if } h = 2 \\
0 & \text{otherwise}
\end{cases}
\]

(Recall that the upper case \( H \) refers to the name of the rv.) For the more adventurous, it could also be given as

\[
p_H(h) = P(H = h) = \begin{cases} 
(0.5)0.5^{|h-1|} & \text{for } h = 0, 1 \text{ or } 2 \\
0 & \text{otherwise}
\end{cases}
\]

Graphically, the probability function could be shown as in Figure 2.2. For discrete distributions, the pf can be displayed using the ‘stick’ approach in this figure, or a histogram.

Note that \( \sum_{t \in \{0,1,2\}} p_H(t) = 1 \) and \( p_H(h) \geq 0 \) for all \( h \) as required of a pf.
Module 2. Distribution of random variables

The following properties of the probability function are implied by the definition.

1. \( p_X(t) \geq 0 \) for all values of \( t \); that is, probabilities are never negative;

2. \( \sum_{t \in R_X} p_X(t) = 1 \) where \( R_X \) is the range of \( X \); that is, the probability function accounts for all possible sample points in the sample space;

3. \( p_X(t) = 0 \) if \( t \notin R_X \);

4. For an event \( A \) defined on a sample space \( S \), the probability of event \( A \) is computed using
   \[
   P(A) = \sum_{t \in A} p_X(t).
   \]

Don’t be put off by the use of \( t \) in these properties; \( x, y \) or any other letter could also have been used. The letter simply represents a particular value of the rv.

The distinction between upper and lower case letters though when discussing rv’s is worth noting. Consider the notation \( p_X(x) \) used to define the probability function. The upper case \( X \) is the name of the random variable; the lower case \( x \) is a particular value that the random variable \( X \) may take. For example, if I roll a standard die, the random variable of interest may be “the
2.3. **Univariate distributions**

number on the top face”, which may be called $X$. Then, the *values* that this random variable can take are $x = 1, x = 2$ through to $x = 6$. So the notation $P(X = 1)$ means “the probability that the number of the top face (that is, $X$) will be equal (that is, the equal to sign $=$) to the particular value 1”. In general, upper case (usually Roman, not Greek) letters refer to the *name* of a random variable; lower case Roman letters refer to the values the variable can take.

There are times when it’s difficult to know whether it’s the rv itself or a value of the rv that is being referred to. Don’t lose any sleep over it. You won’t be penalised if you get it wrong.

**2.3.2 The probability density function**

We run into a major difficulty when attempting to use a probability function to describe the distribution of a continuous rv; that is, a rv that can take on every value in a bounded or unbounded interval.

The problem arises because probability behaves like mass. In the discrete case we imagine mass can be distributed over a number (possibly countably infinite) of distinct points where each point has non-zero mass. This is sufficient to tell us all we need to know about the mass or probability associated with any set of points.

In the continuous case, mass cannot be thought of as an attribute of a point but rather of a region surrounding a point in that as an object is shrunk to a single point its mass also shrinks to zero. The only way we can retain information about how ‘massive’ an object is at a point is to consider its mass per unit volume in the neighbourhood of that point and consider what happens as the volume of the neighbourhood shrinks to zero. This measure does not go to zero. And it is familiar to all of us as (mass) density.

Not only does density have meaning at a point, it allows us to determine the mass of an object by integration if we know the density at every point throughout the mass.

Similarly, to describe the probability distribution for a continuous rv requires us to know the probability density at every point in the range space. The function describing this density is naturally called the *probability density function* or pdf. Once the pdf is known we can determine by integration the probability associated with any event defined on the range space.
Many texts (eg DGS) either don’t explicitly define the pdf or define it indirectly (eg WMS). The definition given below relates directly to the idea of a density as described above but doesn’t tell us much about how to actually find the function.

**Definition 2.4** The probability density function (pdf) of the continuous rv $X$ is a function $f_X$ such that

$$P(a < X \leq b) = \int_a^b f_X(x) \, dx$$

for any interval $(a, b]$ ($a < b$) on the real line.

We are usually only concerned with $a, b \in R_X$ but it makes sense to think of the pdf as defined for all $x$, insisting that $f_X(x) = 0$ for $x \notin R_x$. This definition focuses on the idea that areas under the graph of the pdf represent probabilities and leads to the following properties.

The following properties of the probability density function are implied by the definition.

1. $f_X(x) \geq 0$ for all $-\infty < x < \infty$
2. $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$
3. For an event $E$ defined on a sample space $S$, the probability of event $E$ is computed using

   $$P(E) = \int_E f_X(x) \, dx$$

4. $P(a < X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a \leq X \leq b) = \int_a^b f_X(x) \, dx$
5. From the mean value theorem in calculus it follows that

   $$P(x < X < x + \Delta x) = \int_x^{x+\Delta x} f(t) \, dt = \Delta x f(\xi), \ x < \xi < x + \Delta x$$

6. For $\Delta x$ small,

   $$P(x < X < x + \Delta x) \simeq f(x) \Delta x \quad (2.1)$$
2.3. Univariate distributions

Properties 1 and 2 are sufficient to prove that a function is a pdf; ie if we’re asked to show that some function \( g(x) \) is a pdf all we need do is show that \( g(x) \geq 0 \) for all \(-\infty < x < \infty \) and that \( \int_{-\infty}^{\infty} g(x) \, dx = 1 \).

Property 4 results from noting that if \( X \) is a continuous rv, \( P(X = a) = 0 \) for any and every value \( a \) for the same reason that a point has mass zero.

Properties 5 and 6 are useful in theoretical work and in approximations.

It’s important to remember that the value of a pdf at some point \( x \) does not represent a probability, but rather a probability density, and as such can have any non-negative value of arbitrary size. Properties 5 and 6 directly relate the idea of a density to that of a probability. These properties often find use in making approximations and in theoretical work.

**Example 2.5** Consider the continuous rv \( W \) with the pdf

\[
f_W(w) = 2w \quad \text{for } 0 < w < 1
\]

There are two ways to compute the probability \( P(0 < W < 0.5) \). One is to use the pdf.

\[
P(0 < W < 0.5) = \int_0^{0.5} 2w \, dw
\]

\[
= w^2 \bigg|_0^{0.5}
\]

\[
= 0.25
\]

Alternately, the probability can be computed geometrically. The pdf is shown in Figure 2.3. The region corresponding to \( P(0 < W < 0.5) \) is triangular; integration simply finds the area of this region. The area can also be found using the area of a triangle: the length of the base times the height, divided by two. So the area shaded is

\[
\frac{\text{height}}{\text{base}} = \frac{0.5 \times 1}{2} = 0.25,
\]

and the answer is the same as before.
2.3.3 The distribution function

**Reading 2.4** DGS, Section 3.3.

Another way of describing rvs is using a *distribution function* (df), also called a *cumulative distribution function* (cdf). We will use the df description in this course.

The df gives the probability that a random variable $X$ is less than or equal to a given value $t$.

**Definition 2.5** For any rv $X$ the distribution function, $F_X(x)$, is given by

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

Note that the distribution function applies to discrete or continuous or mixed (see below) rvs.

Two important points to note:

1. The definition includes a less than or equal to sign.
2. The distribution function is defined for all values of $x$. 

Figure 2.3: The pdf of the rv $W$ from Example 2.5. The shaded region is $P(0 < W < 0.5)$. 
2.3. Univariate distributions

If $X$ is a discrete rv with range space $R_x$, the df,

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) \text{ for } x_i \in R_x, \text{ and } -\infty < x < \infty$$

If $X$ is a continuous rv, the df,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t) \, dt \text{ for } -\infty < x < \infty$$

Example 2.6 Consider the simple example in Example 2.4. The probability function for $H$ is given in that Example in numerous forms. To determine the df, first note that when $t < 0$, the accumulated probability is zero; hence, $F_H(t) = 0$ when $t < 0$. At $t = 0$, the probability of 0.25 is accumulated, and no more probability is accumulated until $t = 1$. Thus, $F_H(t) = 0.25$ for $0 \leq t < 1$. Continuing, the df is

$$F_H(t) = \begin{cases} 
0 & \text{for } t < 0 \\
0.25 & \text{for } 0 \leq t < 1 \\
0.75 & \text{for } 1 \leq t < 2 \\
1 & \text{for } t \geq 2 
\end{cases}$$

Note that the variable $t$ is used where $-\infty < t < \infty$. The df can be produced graphically, being careful to clarify what happens at $X = 1$, $X = 2$ and $X = 3$ using open or filled circles (see Figure 2.4).

Example 2.7 Consider a continuous rv $V$ with pdf

$$f_V(v) = \begin{cases} 
v/2 & \text{for } 0 < v < 2 \\
0 & \text{otherwise} 
\end{cases}$$

The df is zero whenever $v \leq 0$. For $0 < v < 2$,

$$F_V(v) = \int_{0}^{v} t/2 \, dt = v^2/4.$$ 

Whenever $v \geq 2$, the df is one. So the df is

$$F_V(v) = \begin{cases} 
0 & \text{if } v \leq 0 \\
v^2/4 & \text{if } 0 < v < 2 \\
1 & \text{if } v \geq 2 
\end{cases}$$
Figure 2.4: A graphical representation of the distribution function for Example 2.6. The filled circles contain the given point, while the empty circles omit the given point.

**Note:** For the integral, *do not write*

\[ \int_0^v \frac{v}{2} \, dv \]

It makes no sense to have the variable of integration as a limit on the integral and also in the function to be integrated. Either write the integral as given in the example, or write \( \int_0^t \frac{v}{2} \, dv = t^2/4 \) and then change the variable from \( t \) to \( v \).

Note that if a random variable \( X \) is discrete, then the df will have points of discontinuity. If the variable is continuous then the df will be continuous.

These and other properties of the df are stated below.

(a) \( 0 \leq F_X(x) \leq 1 \) because \( F_X(x) \) is a probability.

(b) \( F_X(x) \) is a non-decreasing function of \( x \).
   If \( x_1 < x_2 \) then \( \{ x : x \leq x_1 \} \subset \{ x : x \leq x_2 \} \)
   So \( F_X(x_1) = P(X \leq x_1) \leq P(X \leq x_2) = F_X(x_2) \)

(c) Denoting \( \lim_{x \to \infty} F_X(x) \) by \( F_X(\infty) \) and \( \lim_{x \to -\infty} F_X(x) \) by \( F_X(-\infty) \) we have
   \( F_X(\infty) = 1 \) and \( F_X(-\infty) = 0 \).

(d) \( P(a < X \leq b) = F_X(b) - F_X(a) \)
(e) If $X$ is discrete then $F_X(x)$ is a step-function but if $X$ is continuous $F_X$ will be a continuous function for all $x$.

(f) We can use (2.2) to find $F_X(x)$ given $P(X = x)$ or (3.4) to find $F_X(x)$ given $f_X(x)$ but we need to be able to proceed in the other direction as well. That is, given $F_X(x)$ how do we find $P(X = x)$ for $X$ discrete or $f_X(x)$ for $X$ continuous?

(i) It can be seen from the graph of the df in Example 2.6 that the values of $x$ where a ‘jump’ in $F_X(x)$ occurs are the points in the range space and the probability associated with a particular point in $R_X$ is the ‘height’ of the jump there. That is,

$$p_X(x_j) = P(X = x_j) = F_X(x_j) - F_X(x_{j-1})$$  \hspace{1cm} (2.2)

(ii) For $X$ continuous, from the Fundamental Theorem of Calculus,

$$f_X(x) = \frac{dF_X(x)}{dx} \quad \text{where the derivative exists.} \hspace{1cm} (2.3)$$

2.3.4 Comparing discrete and continuous distributions

Figure 2.5 provides a qualitative comparison of the information in discrete and continuous distributions.
Figure 2.5: Probability function and distribution function for a discrete random variable (left side) and density function and distribution function for a continuous random variable (right side)

\[ P(X = x) \quad f_X(x) \]

\[ P(X \leq x) \quad F_X(x) \]

2.4 Probabilities of compound events for discrete r.v.’s

In Figure 2.6, the X axis represents a discrete random variable 0, 1, 2, …, (i.e. integers) and the Y axis the cumulative probability for X = 0, 1, 2, ….

The case (A) indicates the event \( X \leq x_1 \) and its complement \( X > x_1 \).

\( P(X > x_1) = 1 - P(X \leq x_1) \). The event \( X \leq x_1 \) is termed the lower tail of the distribution with respect to \( x_1 \) and \( X > x_1 \) is the upper tail of the distribution.

It is important to recognize that the cumulative distribution function (cdf)
2.4. Probabilities of compound events for discrete r.v.'s

for discrete random variables refers to the probability of values up to and including the particular value. For example, \( P(X \leq x_1) = P(X = 0) + P(X = 1) + \ldots + P(X = x_1) \).

If \( P(X < x_1) \) was required, it would be determined from the cumulative distribution function by taking the integer below \( x_1 \), \( P(X < x_1) = P(X \leq (x_1 - 1)) \)

The upper tail refers to events like \( P(X > x_1) \), i.e. where the event is strictly greater than i.e. \( X > x_1 \) and NOT \( \geq x_1 \). Since \( X \geq x_1 \) is the same event as \( X > (x_1 + 1) \), \( P(X \geq x_1) = P(X > (x_1 + 1)) \) and we would find \( P(X \geq x_1) \) by the upper tail of the cdf for \( X = (x_1 + 1) \).

Case (B) shows the mutually independent events \( P(X \leq x_1), (x_1 < X \leq x_2), (X > x_2) \).

The event \((x_1 < X \leq x_2)\) has probability \( P(X \leq x_2) - P(X \leq x_1) \) or the interval between the dotted lines on the Y-axis.

If we wanted \( P[x_1 \leq X \leq x_2] \) it could be obtained by summing elementary events, \( P[x_1 \leq X \leq x_2] = P[X = x_1] + P(X = (x_1 + 1)] + P(X = (x_1 + 2)] + \ldots + P[X = x_2] \) but this may be tedious if there are very many integer values between \( x_1 \) and \( x_2 \).

Recognising that \( P[X \geq x_1] \) is the same as \( P[X > (x_1 - 1)] \),
\( P[x_1 \leq X \leq x_2] = P[(x_1 - 1) < X \leq x_2] = P[X \leq x_2] - P[X \leq (x_1 - 1)] \)

Figure 2.6: Probabilities of Events.
2.4.1 Mixed random variables

Some rvs are neither continuous nor discrete, but have parts of both. These rvs are called \textit{mixed random variables}.

Example 2.8 In a factory producing diodes, a fraction of the diodes \( p \) fail immediately. The distribution of the lifetime (in hundred of days), say \( Y \), of the diodes is given by a discrete part at \( y = 0 \) for which \( P(Y = 0) = p \) and a continuous part for \( y > 0 \) described say by the pdf

\[
 f_Y(y) = (1 - p) \exp(-y) \quad \text{if} \quad y > 0.
\]

(Strictly speaking \( f_Y(y) \) is not a proper pdf because it doesn’t integrate to one but we can see that the total probability is

\[
 0.4 + \int_0^\infty 0.6 \exp(-y) \, dy = 0.4 + 0.6 = 1
\]

as required.)

Consider a diode for which \( p = 0.4 \). The probability distribution is displayed in Figure 2.7 (a) where a solid dot is included to show the discrete part.

We see there are difficulties representing the probability distribution in this mixed case because we need to combine a probability distribution and a pdf. These difficulties are circumvented by using the distribution function. The df of \( Y \) is

\[
 F_Y(y) = \begin{cases} 
 0 & \text{if } y < 0 \\
 0.4 & \text{if } y = 0 \\
 0.4 + 0.6(1 - \exp(-y)) & \text{if } y > 0
\end{cases}
\]

2.5 Self-assessment exercises

The following exercises are designed to provide practice at problem-solving based on the material in this module. Solutions are provided at the end of the module. Additional exercises are available in the next section and in the textbook.

Ex. 2.1 Find a formula for the pf of the number of heads observed in five tosses of a fair coin. Compute the df also.
2.5. Self-assessment exercises

Ex. 2.2 If the pdf of $X$ is

$$f_X(x) = \begin{cases} k \exp(-3x) & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

compute

(a) the value of $k$.
(b) the probability $P(0.5 < X < 1)$.
(c) the df of $X$.

Ex. 2.3 Consider the rv $W$ with pf

$$p_W(w) = 0.6(0.4)^{w-1}$$

when $w = 1, 2, \ldots$, and is zero elsewhere.

(a) Find $P(W < 5)$.
(b) Find $P(2 \leq W < 4 \mid W < 5)$. 

---

Figure 2.7: A mixed random variable. In (a), the pdf and probability function; in (b), the distribution function (see Section 2.3.3). In (a), the empty circle implies the probability does not occur at this point.
Ex. 2.4 The rv $X$ has the pf

$$p_X(x) = \begin{cases} 
0.3 & \text{for } x = 10 \\
0.2 & \text{for } x = 15 \\
0.5 & \text{for } x = 20 \\
0 & \text{elsewhere}
\end{cases}$$

(a) Find the distribution function for $X$.
(b) Compute $P(X > 13)$.
(c) Compute $P(X \leq 10 \mid X \leq 15)$.

Ex. 2.5 Consider the continuous rv $Z$ with pdf

$$f_Z(z) = \begin{cases} 
2(3 - z)/15 & \text{for } -1 < z < 2 \\
0 & \text{elsewhere}
\end{cases}$$

(a) Find the df of $Z$.
(b) Find $P(z < 0)$.

Ex. 2.6 The random variable $A$ has the distribution

$$p_A(a) = \frac{c}{a} \quad \text{for } a = 1, 2, 4,$$

and is zero otherwise.

(a) Determine the value for $c$.
(b) Determine and plot the distribution function for $A$.
(c) Calculate $P(A < 2 \mid A < 3)$.

Ex. 2.7 The random variable $Q$ has a probability density function given by Figure 2.8.

(a) Determine the distribution function of $Q$.
(b) Using the distribution function or otherwise, find the median of $Q$ (ie the point that divides the area under the pdf into two equal halves).

2.6 Exercises

Ex. 2.8 Suppose $X$ is a continuous rv and

$$f_X(x) = ke^{-x/4}, \quad 0 \leq x < \infty$$
Figure 2.8: Probability density function in Question 2.7.

(a) Find $k$ so that $f_X(x)$ is a pdf.
(b) Find $F_X(x)$.
(c) Find $P(X \leq 2)$.
(d) Find $P(2 < X < 4)$.

**Ex. 2.9** The pdf of a random variable $X$ is defined by

$$f(x) = \begin{cases} 
0 & x \leq 1 \\
\frac{1}{4} & 1 < x \leq 2 \\
0 & 2 < x \leq 3 \\
\frac{1}{4} & 3 < x \leq 4 \\
0 & 4 < x \leq 5 \\
\frac{1}{4} & 5 < x \leq 6 \\
0 & x > 6 
\end{cases}$$

(a) Find $F(x)$ and draw a graph of both $f(x)$ and $F(x)$.
(b) Find $P(X < 4.3)$.
(c) Find $P(1.5 < X < 5.3)$
(d) Find $P(X \geq 3.5)$.

**Ex. 2.10** For the following experiments, determine the sample space and carefully write down the random variable of interest stating whether it is a discrete, continuous or mixed random variable.
(a) The number of heads in two throws of a fair coin.
(b) The number of throws of a fair coin until a head is observed.
(c) A quality audit team at USQ is interested in the time taken for students to download the USQ homepage.
(d) The Toowoomba City Council is currently undertaking a study into alternative means of transport, and are gathering data on the time it would take people to walk to work.
(e) Traffic engineers collect data to improve traffic control. They count number of cars that pass through an intersection during a day.
(f) Hospital administrators collect data on the number of procedures they perform to ensure adequate service. They record the number of X-rays performed at the local hospital per day.
(g) Weather and climate forecasting requires knowledge of the barometric pressure in Toowoomba at 5pm each afternoon.

Ex. 2.11 Consider an experiment where a fair coin is tossed twice. Let \( X \) be the number of heads observed in the two tosses, and \( Y \) be the number of heads on the first toss of the coin.

(a) Construct the table of the joint probability function for \( X \) and \( Y \).
(b) Determine the marginal probability function for \( X \).
(c) Determine the conditional distribution of \( X \) given one head appeared on the first toss.
(d) Determine if the variables \( X \) and \( Y \) are independent or not, justifying your answer with necessary calculation or argument.

Ex. 2.12 Consider the density function defined as follows:

\[
    f_W(w) = \begin{cases} 
        kw & \text{for } 0 < w < 1 \\
        k & \text{for } 1 \leq w \leq 2 \\
        0 & \text{otherwise}
    \end{cases}
\]

for some constant \( k \).

(a) For what value of \( k \) is \( f_W(w) \) a valid pdf?
(b) Plot the probability density function of \( W \).
(c) Determine and plot the distribution function of \( W \).
2.6. Exercises

Ex. 2.13 Very little is known about the random variable $U$. It is known, however, that $U$ is a continuous random variable such that the probability density function is zero for $U < -2$ and $U > 4$, and is non-zero for $-2 < U < 4$.

(a) Sketch a possible distribution function for $U$. Label any important details of your sketch.

(b) It is later found the median of $U$ is 2. On a fresh set of axes, sketch a possible distribution function for $U$. Label any important details of your sketch.

Ex. 2.14 Suppose a fair coin is tossed three times, and a head (H) or a tail (T) is recorded on each toss. The two events $A$ and $B$ are defined as:

- $A$: The number of heads in the three tosses
- $B$: The number of heads in the first two tosses

(a) Construct a joint probability table for the events $A$ and $B$.

(b) Determine the correlation coefficient between $A$ and $B$.

(c) Explain in the context of the question what the correlation coefficient means.

Ex. 2.15 Consider the diagram in Figure 2.9. Answer the following questions.

(a) Is the random variable, $X$, a discrete, continuous or mixed random variable? Justify your answer.

(b) Determine and sketch the distribution function for $X$.

(c) Find $P(12 < X \leq 15)$.

Ex. 2.16 In production lines, it is important to ensure that the amount of product placed in containers is not less than the amount advertised. However, it is also important that there is not too much product in the container as this reduces profits. Balancing the two criteria requires careful management.

A company selling cordial advertises that the bottles contain two litres of cordial. In practice, the machine actually fills bottles with an amount that varies randomly between a minimum of 1.97 litres and a maximum of 2.07 litres.

(a) Carefully write down the variable of interest.

(b) Write down the sample space for the amount of cordial placed in the bottles.
Could the relative frequency approach to probability be used to determine the probability that a bottle is underfilled? Explain your answer.

It is found that the actual amount of cordial placed in the bottles, $C$, follows the probability density function pictured in Figure 2.10. Determine a value for $k$ (preferably without using calculus).

Determine the probability that the bottles are underfilled.

Calculate the median amount of cordial placed in the containers.

**Ex. 2.17** The rv $V$ has the probability function

$$p_V(v) = \frac{c^2}{v}$$

for $v = 1, 2, 3$ and is zero elsewhere.

(a) Show there are two values of $c$ such that $\sum_{v=1}^{3} p_V(v) = 1$.

(b) What is the only possible value for $c$? Explain.

**Ex. 2.18** A random variable $X$ has probability function given by

$$P(X = x) = kx, \quad x = 1, 2, 3, \ldots, 2n$$

(a) Find the value of $k$ that makes this a probability function.
2.7 Some answers and hints

(a) Let $X$ be the number of heads in five tosses. The range of $X$ is $R_X = \{0, 1, 2, 3, 4, 5\}$. There are a total of $2^5 = 32$ possible outcomes from the five tosses. Zero heads can occur in $C_0^5 = 1$ way; one head in $C_1^5 = 5$ ways; and so on. So the probability of zero heads is $C_0^5 / 32$; and the probability of one head is $C_1^5 / 32$; and so on. In general then,

$$p_X(x) = \frac{C_x^5}{32}$$

for $x = 0, 1, 2, 3, 4, 5$

(b) Find the probability that $X \leq n$.

(c) Find the probability that $X$ is even.

Ex. 2.19 A random variable $X$ has probability function

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X = x)$</td>
<td>$\frac{1 + 3\theta}{4}$</td>
<td>$\frac{1 - \theta}{4}$</td>
<td>$\frac{1 + 2\theta}{4}$</td>
<td>$\frac{1 - 4\theta}{4}$</td>
</tr>
</tbody>
</table>

For what range of values of $\theta$ is this a probability distribution?

Ex. 2.20 (Computer exercise) Use simulation to estimate the value of $\pi$.
(b) The df is

\[
F_X(x) = \begin{cases} 
0 & \text{if } x < 0 \\
1/32 & \text{if } 0 \leq x < 1 \\
6/32 & \text{if } 1 \leq x < 2 \\
16/32 & \text{if } 2 \leq x < 3 \\
26/32 & \text{if } 3 \leq x < 4 \\
31/32 & \text{if } 4 \leq x < 5 \\
1 & \text{if } x \geq 5.
\end{cases}
\]

2.3 (a) \(0.9744\) (b) \(0.3744/0.9744 = 0.3842\).

2.3 (a) First, \(f_X(x) \geq 0\) for all values of \(x\) provided \(k > 0\). Secondly,

\[
\int_0^\infty k \exp(-3x) \, dx = k \frac{\exp(-3x)}{-3} \bigg|_{x=0}^{x=\infty} = k/3 = 1,
\]

so \(k = 3\).

(b) \(P(0.5 < X < 1) = \int_{x=0.5}^1 3 \exp(-3x) \, dx \approx 0.173\).

(c)

\[
F_X(x) = \int_{u=0}^{u=x} 3 \exp(-3u) \, du = 3 \frac{\exp(-3u)}{-3} \bigg|_{u=0}^{u=x} = 1 - \exp(-3x).
\]

Therefore

\[
F_X(x) = \begin{cases} 
0 & \text{if } x \leq 0 \\
1 - \exp(-3x) & \text{if } x > 0
\end{cases}
\]

2.4 \(F_X(x) = 0\) for \(x < 10\), \(F_X(x) = 0.3\) for \(10 \leq x < 15\), \(F_X(x) = 0.5\) for \(15 \leq x < 20\), \(F_X(x) = 1\) for \(x \geq 20\); \(P(X > 13) = 1 - F_X(13) = 0.7\);

\(P(X \leq 10 \mid X \leq 15) = P(X \leq 10) / P(X \leq 15) = F_X(10) / F_X(15) = 0.3/0.5 = 0.6\).

2.5 \(F_Z(z) = 0\) for \(z \leq -1\); \(F_Z(z) = 6z/15 - z^2/15 + 7/15\) for \(-1 < z < 2\);

\(F_Z(z) = 1\) for \(z \geq 2\). \(P(Z < 0) = F_Z(0) = 7/15\).

2.5 \(\kappa = 21.623732\); (no answer); \(P(0 < Z < 1) = 0.0547\).

2.6 \(c = 4/7\); \(F_A(a) = 0\) for \(a < 1\), \(F_A(a) = 4/7\) for \(1 \leq a < 2\), \(F_A(a) = 6/7\) for \(2 \leq a < 4\), \(F_A(a) = 1\) for \(a \geq 4\); \((4/7)/(6/7) = 2/3\).

2.7 \(F_Q(q) = 0\) for \(q < -1\), \(F_Q(q) = (q_1)/2\) for \(-1 \leq q < 0\), \(F_Q(q) = (1/2) + (q/2) - (q^2/8)\) for \(0 \leq q < 2\), \(F_Q(q) = 1\) for \(q > 2\); solve for \(1/2 = (1/2) + (q/2) - (q^2/8)\), median is \(q = 0\) (ignore solution \(q = 4\)).

2.8 (a) \(k = 0.25\) (b) \(F_X(x) = 1 - e^{-x/4}\) (c) Use the result of (b) \(0.3935\) (d) \(0.2387\).
2.7. Some answers and hints

2.9 Remember $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$ which will become a sum of integrals over the appropriate ranges in this problem. Use the distribution function to answer parts (b)–(d).

(b) 0.75 (c) 0.7 (d) 0.5

2.18 All three parts involve summation of an arithmetic progression.

(a) $k = \frac{1}{n(1 + 2n)}$ (b) $\frac{1 + n}{2(1 + 2n)}$ (c) $\frac{(1 + n)}{(1 + 2n)}$

2.19 Probabilities must lie in range $[0,1]$. Thus $-1/3 \leq \theta \leq 1/4$. (Check probabilities sum to 1 for all $\theta$.)

2.19 See Section 3.5, Definition 3.15 for a discrete rv. $M_X(t) = \frac{1}{2}(e^t + e^{-t})$

(a) Expand the mgf and find the coefficient of $\frac{t^r}{r!}$.

(b) Note that since $E(X) = 0$ that $E(X - \mu)^r = E(X)^r$. 
Module 2. Distribution of random variables
Module 3

Mathematical expectation

Contents

3.1 The expected value ......................... 88
   3.1.1 Expectation of a function of a random variable .... 90
3.2 The variance and standard deviation .......... 91
3.3 Features of distributions ..................... 94
   3.3.1 Measures of location ..................... 94
   3.3.2 Measures of dispersion .................... 96
   3.3.3 Symmetry ............................... 97
3.4 Higher moments ................................ 97
3.5 Moment generating function .................. 99
   3.5.1 Using the mgf to generate moments .......... 101
   3.5.2 Some useful results ....................... 102
   3.5.3 Determining the distribution from the mgf ..... 104
3.6 Tchebysheff’s inequality ..................... 105
3.7 Self-assessment exercises ..................... 106
3.8 Exercises .................................. 108
3.9 Some answers and hints ....................... 114
Module objectives

Upon completion of this module students should be able to:

- understand the concept and definition of mathematical expectation
- compute the expectations of a rv, functions of a rv and linear functions of rv’s
- know how to compute the variance and other higher moments of a rv
- derive the moment generating function of a rv and linear functions of a rv
- find the moments of a rv from the moment generating function
- interpret and compute the covariance and the coefficient of correlation between two random variables
- compute the conditional mean and conditional variance of a random variable for some given value of another random variable
- state and make use of Tchebysheff’s inequality

3.1 The expected value

In discussing random variables, it is important to be able to identify a typical (or ‘average’) value of a random variable. To do this, we need to understand expectation.

**Reading 3.1** DGS, Sections 4.1, 4.2 and 4.3; WMS, Sections 3.3 and 4.3.

The expectation is a mathematical expression of the mean of the probability distribution of the rv. Note that the expectation is a mathematical operator. The definition looks different in detail for discrete and continuous random variables, but the intention is the same.

**Definition 3.1** The expectation or expected value or mean of a random variable $X$ is defined as

$$E(X) = \begin{cases} \sum_{x \in \mathbb{R}_X} xp_X(x) & \text{if } X \text{ is discrete with pf } p_X(x) \\ \int_{-\infty}^{\infty} xf_X(x) & \text{if } X \text{ is continuous with pdf } f_X(x) \end{cases}$$
3.1. The expected value

Effectively \( E(X) \) is a weighted average of the points in \( R_X \), the weights being the probabilities in the discrete case and probability densities in the continuous case.

Typically the expectation is denoted by \( \mu \) or \( \mu_X \) if there is a need to distinguish between rvs.

**Example 3.1** Consider a continuous random variable \( X \) with pdf

\[
f_X(x) = \begin{cases} 
  \frac{x}{4} & \text{for } 1 < x < 3 \\
  0 & \text{elsewhere}.
\end{cases}
\]

The expected value of \( X \) is, by definition,

\[
E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx \\
= \int_{1}^{3} x \left( \frac{x}{4} \right) \, dx \\
= \frac{1}{12} x^3 \bigg|_{1}^{3} = \frac{13}{6}.
\]

The expected value of \( X \) is \( \frac{13}{6} \).

**Example 3.2** Consider the discrete random variable \( U \) with probability function

\[
p_U(u) = \begin{cases} 
  \frac{(u^2 + 1)}{5} & \text{for } u = -1, 0, 1 \\
  0 & \text{elsewhere}.
\end{cases}
\]

The expected value of \( U \) is, by definition,

\[
E(U) = \sum_{u=-1,0,1} u p_U(u) \\
= \left( -1 \times \frac{(-1)^2 + 1}{5} \right) + \left( 0 \times \frac{(0)^2 + 1}{5} \right) + \left( 1 \times \frac{(1)^2 + 1}{5} \right) \\
= -2/15 + 0 + 2/15 = 0.
\]

The expected value of \( U \) is 0.
**Example 3.3** Consider tossing a coin once and counting the number of tails. Let this random variable be $T$. The probability function is

$$p_T(t) = \begin{cases} 0.5 & \text{for } x = 0 \text{ or } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of $T$ is, by definition

$$E(T) = \sum_{i=1}^{2} t p_T(t)$$

$$= P(T = 0) \times 0 + P(T = 1) \times 1$$

$$= (0.5 \times 0) + (0.5 \times 1) = 0.5.$$

Note that 0.5 tails can never actually be observed in practice. But it would be silly to round up (or down) and say that the expected number of tails in one toss of a coin is one (or zero). The expected value of 0.5 simply means that over a large number of repeats of this experiment, we expect a tail to occur in half of those repeats.

### 3.1.1 Expectation of a function of a random variable

Let $X$ be a discrete random variable with a probability function $p_X(x)$, or a continuous random variable with pdf $f_X(x)$. Then if $g(X)$ is a real-valued function of $X$, the expected value of $g(X)$ is defined as follows.

**Definition 3.2**

$$E(g(X)) = \begin{cases} \sum_{x \in \mathbb{R}^X} g(x)p_X(x) & \text{if } X \text{ is discrete with pf } p_X(x) \\ \int_{-\infty}^{\infty} g(x)f_X(x) \, dx & \text{if } X \text{ is continuous with pdf } f_X(x) \end{cases}$$

If $Y = g(X)$ we would naturally write $\mu_Y = E(Y) = E(g(X))$.

Before considering a particular function of interest we prove an important property of the expectation operator.

**Theorem 3.3** For any rv $X$ and constants $a$ and $b$

$$E(aX + b) = aE(X) + b$$
3.2. The variance and standard deviation

Proof We assume \( X \) is a discrete rv with probability function \( p_X(x) \). (An analogous proof applies in the continuous case by replacing the probability function with a pdf and summations with integrals.) Then by Definition 3.2 with \( g(X) = aX + b \)

\[
E(aX + b) = \sum_x (ax + b)p_X(x) = a \sum_x p_X(x) + \sum_x bp_X(x) = aE(X) + b
\]
where we have used the fact that \( \sum_x p_X(x) = 1 \).

Example 3.4 Consider the random variable \( Y = 2x \) where \( X \) is defined in Example 3.1. Using Theorem 3.3 with \( a = 2 \) and \( b = 0 \), the value of \( E(Y) \) can be written

\[
E(2X) = 2E(X) = 2 \times 13/6 = 13/3.
\]

3.2 The variance and standard deviation

An important function of a random variable gives rise to the variance of a random variable (or, if you like, of the distribution of the rv).

The variance is a measure of how spread out the values of a random variable are. A small variance means the observations are nearly the same; a large variance means they are quite different. The variance of a random variable \( X \) can be defined using the expectation operator.

Definition 3.4 The variance of a rv \( X \) (or of the distribution of \( X \))

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

where \( \mu = E(X) \).

The variance of \( X \) is commonly denoted by \( \sigma^2 \), or \( \sigma_X^2 \) if it’s necessary to distinguish amongst variables.

The unit of measurement for variance is the original \((x)\) unit squared. It is often more meaningful to describe the ‘spread’ in terms of the original units by taking the square root of the variance.
**Definition 3.5** The standard deviation of a rv $X$ is defined as the positive square root of the variance (and is thus denoted by $\sigma$); ie

$$sd(X) = +\sqrt{\text{var}(X)}$$

**Example 3.5** Suppose a fair die is tossed and $X$ denotes the number of points showing. Then $P(X = x) = 1/6$ for $x = 1, 2, 3, 4, 5, 6$. and $\mu = E(X) = \sum xP(X = x) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 7/2$.

The variance of $X$ is then

$$\sigma^2 = \text{var}(X) = \sum (X - \mu)^2 P(X = x)$$

$$\quad = \frac{1}{6} \left[ (1 - \frac{7}{2})^2 + (2 - \frac{7}{2})^2 + \cdots + (6 - \frac{7}{2})^2 \right]$$

$$\quad = \frac{70}{24}.$$ 

The standard deviation is then $\sigma = \sqrt{\frac{70}{24}} = 1.71$.

An important result is known as the *computational formula for variance*.

**Theorem 3.6** For any rv $X$,

$$\text{var}(X) = E(X^2) - [E(X)]^2$$

**Proof** Let $E(X) = \mu$, then

$$\text{var}(X) = E((X - \mu)^2)$$

$$\quad = E(X^2 - 2X\mu + \mu^2)$$

$$\quad = E(X^2) - E(2X\mu) + E(\mu^2)$$

$$\quad = E(X^2) - 2\mu E(X) + \mu^2$$

$$\quad = E(X^2) - 2\mu^2 + \mu^2$$

$$\quad = E(X^2) - \mu^2$$

$$\quad = E(X^2) - (E(X))^2.$$
3.2. The variance and standard deviation

This formula is often easier to use to compute \( \text{var}(X) \) than using the definition.

**Example 3.6** Consider the continuous random variable \( X \) with pdf

\[
f_X(x) = \begin{cases} 
-x^2 + 2x - 1/6 & \text{for } 0 < x < 2 \\
0 & \text{elsewhere}
\end{cases}
\]

The variance of \( X \) can be computed in two ways, directly using \( \text{var}(X) = E((X - \mu)^2) \) or using the computational formula.

The expected value of \( X \) is

\[
E(X) = \int_0^2 x(-x^2 + 2x - 1/6) \, dx
\]

\[
= -\frac{x^4}{4} + \frac{2x^3}{3} - \frac{x^2}{12} \bigg|_0^2 = 1.
\]

To use the computational formula,

\[
E(X^2) = \int_0^2 x^2(-x^2 + 2x - 1/6) \, dx
\]

\[
= -\frac{x^5}{5} + \frac{x^4}{2} - \frac{x^3}{18} \bigg|_0^2 = \frac{52}{45},
\]

and so

\[
\text{var}(X) = E(X^2) - (E(X))^2 = \frac{7}{45}.
\]

Using the definition,

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

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

\[
= \int_0^2 (x - 1)^2(-x^2 + 2x - 1/6)^2 \, dx
\]

\[
= \int_0^2 x^5 - 31x^3 + 7x^2 - \frac{1}{6}x \, dx
\]

\[
= -\frac{x^6}{5} + \frac{x^4}{18} - \frac{7x^2}{6} - \frac{x}{6} \bigg|_0^2 = \frac{7}{45}.
\]

Both methods give the same answer of course, and both methods require initial computation of \( E(X) \). The computational formula requires less work.
The variance represents the mean squared distance of the values of the rv from the mean.

The variance can never be negative. It is only ever zero in the uninteresting case where all the values of the random variable are identical (that is, there is no variation).

The units of the variance are the square of the units of $X$. That is, if $X$ is measured in metres the variance of $X$ is in metres$^2$. Because of this, the variance is less popular than the standard deviation, which has the same units as the rv, in describing spread numerically. In theoretical work however, because of the square root function, variance is easier to work with than standard deviation and you will therefore find it rather than standard deviation featuring in many results throughout the study of statistics.

### 3.3 Features of distributions

Instead of describing the distribution of a rv in terms of a probability function of pdf or df, it is often sufficient to describe a distribution in terms of certain quantities that can be obtained from the distribution function.

Two general features of a distribution are its

(i) location or centre

(ii) dispersion or spread

#### 3.3.1 Measures of location

The most important measure of location is the mean $\mu = \text{E}(X)$, which we have already defined. In physical terms this measure is the balance point of the distribution in the sense that the mean deviation of the rv from $\mu$ is zero; is $\text{E}(X - \mu) = 0$. (Why is this?)

The mean is not always a suitable measure of centre since the balance point can be unduly affected by distributions that have a relatively higher concentration of probability at one end of the distribution than the other. Other measure of centre may then be more suitable. The mode and median are two commonly-used alternatives.
3.3. Features of distributions

**Definition 3.7** The mode(s) is/are the value(s) \( x \in \mathbb{R}_X \) for the discrete case at which \( p_X(x) \) attains its maximum or the value(s) of \( x \) for the continuous case at which \( f_X(x) \) attains its maximum.

**Definition 3.8** The median of a random variable \( X \) (or of a distribution) is a number \( \nu \) such that

\[
P(X \leq \nu) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq \nu) \geq \frac{1}{2}.
\]

(3.1)

For \( X \) discrete, the median is not necessarily unique.

**Example 3.7** If \( X \) is a discrete random variable with probability function,

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = x) )</td>
<td>1/128</td>
<td>7/128</td>
<td>21/128</td>
<td>35/128</td>
<td>35/128</td>
<td>21/128</td>
<td>7/128</td>
<td>1/128</td>
</tr>
</tbody>
</table>

We see this distribution is bimodal with modes at 3 and 4.

The median, \( \nu \), is any value of \( X \) in the range \( 3 < x < 4 \) since in this range \( P(X \leq x) = 0.5 \) and \( P(X \geq x) = 0.5 \). That is the median is not unique. The probability function and its df are shown in Figure 3.1.

**Example 3.8** Let \( X \) be a discrete random variable with probability function

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X = x) )</td>
<td>1/64</td>
<td>6/64</td>
<td>15/64</td>
<td>20/64</td>
<td>15/64</td>
<td>6/64</td>
<td>1/64</td>
</tr>
</tbody>
</table>

Figure 3.1: Probability function and df for \( X \) in Example 3.7.
The mode of this distribution is 3.
The median is $\nu = 3$ since this is the only value of $X$ that will satisfy Definition 3.8. The probability function and it’s df are shown in 3.2.

![Figure 3.2: Probability function and df for $X$ in Example 3.8.](image)

### 3.3.2 Measures of dispersion

If most of the probability lies near the mean, the dispersion will be small; if the probability is spread out over a considerable range the dispersion will be large. We want a measure of the ‘spread’ irrespective of the location or centre.

We have already seen variance and standard deviation are suitable measures of dispersion. Another alternative is the following.

**Definition 3.9** The mean absolute difference (MAD) is defined as

$$E(|X - \mu|) = \begin{cases} \sum_x |x - \mu|p_X(x) & \text{for discrete } X \\ \int_{-\infty}^{\infty} |x - \mu|f_X(x) \, dx & \text{for continuous } X \end{cases}$$

Like the standard deviation, the MAD is difficult to use in theoretical work.

The most-used measure of location is the mean while for dispersion it is the variance (or standard deviation).

**Example 3.9** Consider the fair die described in Example 3.5.
3.4. Higher moments

Then \( \mu = E(X) = 7/2 \) and thus

\[
\sigma^2 = \text{var}(X) = \sum (x - \mu)^2 P(X = x) \\
= \frac{1}{6} \left[ \left( 1 - \frac{7}{2} \right)^2 + \left( 2 - \frac{7}{2} \right)^2 + \cdots + \left( 6 - \frac{7}{2} \right)^2 \right] \\
= \frac{70}{24}.
\]

The standard deviation is then given by \( \sigma = \sqrt{\frac{70}{24}} = 1.71 \).

The MAD of \( X \) is

\[
E(|X - \mu|) = \sum |x - \mu| P(X = x) \\
= \frac{1}{6} \left[ \left| 1 - \frac{7}{2} \right| + \left| 2 - \frac{7}{2} \right| + \cdots + \left| 6 - \frac{7}{2} \right| \right] \\
= 1.5
\]

3.3.3 Symmetry

Definition 3.10 The distribution of \( X \) is said to be symmetric if for all \( x \in \mathbb{R}^X \),

\[
P(X = \mu + x) = P(X = \mu - x) \quad \text{for all } x \in \mathbb{R}^X \text{ in the discrete case} \\
f_X(\mu + x) = f_X(\mu - x) \quad \text{for all } x \in \text{the continuous case}
\]

It can easily be seen that for a symmetric distribution the mean is also a median of the distribution.

3.4 Higher moments

The idea of a mean and variance are generalised in the following definitions.

Definition 3.11 The \( r \)th moment about the origin or \( r \)th raw moment of a random variable \( X \) is defined as

\[
\mu'_r = E(X^r) = \begin{cases} 
\sum_{X} x^r P_X(x) & \text{for } X \text{ discrete with pf } P_X(x) \\
\int_{-\infty}^{\infty} x^r f_X(x) & \text{for } X \text{ continuous with pdf } f_X(x)
\end{cases}
\]

where \( r \) is a positive integer.
Definition 3.12  The \( r \)th central moment or \( r \)th moment about the mean is defined as

\[
\mu_r = E((X - \mu)^r) = \begin{cases} 
\sum_{x} (x - \mu)^r p_X(x) & \text{for } X \text{ discrete with } p_X(x) \\
\int_{-\infty}^{\infty} (x - \mu)^r f_X(x) & \text{for } X \text{ continuous with } f_X(x) 
\end{cases}
\]

where \( r \) is a positive integer.

From these definitions we see the mean \( \mu'_1 = \mu \) is the first moment about the origin, or the first raw moment and the variance \( \mu_2 = \sigma^2 \) is the second moment about the mean or the second central moment. Higher moments also exist that describe other features of a rv.

It can be shown that for a symmetric distribution the odd central moments are zero. This suggests that odd central moments can be used to measure the asymmetry of a distribution. It is convenient to use, instead of central moments themselves, expressions that are unaffected by a linear transformation of the type \( Y = AX + b \). Now the ratio \( (\mu_r)^p/\mu_r^p \) is such an expression and the simplest form of it is \( \mu_3^2/\mu_2^3 \).

Definition 3.13  The skewness of a distribution is defined as

\[
\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}
\]

If \( \gamma_1 > 0 (\leq 0) \) we say the distribution is positively (negatively) skewed, and it is ‘drawn out’ in the positive (negative) direction.

Another ratio gives a measure of kurtosis (or peakedness).

Definition 3.14  The kurtosis of a distribution is defined as

\[
\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3
\]

Example 3.10  Monypenny and Middleton [22, 23] use the skewness and kurtosis to analyse wind gusts at Sydney airport.

Example 3.11  Galagedera, Henry and Silvapulle [12] used higher moments in a capital analysis pricing model for Australian stock returns.
Example 3.12 Consider the discrete random variable $U$ from Example 3.2. The raw moments are

\[
\mu'_r = E(U^r) = \sum_{u=-1,0,1} u^r \frac{u^2 + 1}{5} = (-1)^r \frac{(-1)^2 + 1}{5} + (0)^r \frac{(0)^2 + 1}{5} + (1)^r \frac{(1)^2 + 1}{5} = -\frac{2(-1)^r}{5} + 0 + \frac{2}{5} = \frac{2}{5} \left[(-1)^r + 1\right]
\]

for the $r$th raw moment. Then,

\[
E(X) = \mu'_1 = \frac{2}{5} \left[(-1)^1 + 1\right] = 0,
E(X^2) = \mu'_2 = \frac{2}{5} \left[(-1)^2 + 1\right] = 4/5,
\]

so that var$(X) = E(X^2) - E(X)^2 = (4/5) - 0^2 = 4/5$. Notice that once the initial computations to find $\mu'_r$ have been done, the evaluation of any raw moment is simple.

By themselves, the mean, variance and skewness do not completely describe a distribution; many different distributions can be found having a given mean, variance and skewness. However, it turns out that in general all the moments of a distribution together define the distribution.

### 3.5 Moment generating function

So far, the distribution of a rv has been described using a probability function, pdf or a distribution function. Sometimes, however, it is convenient to work with a different representation. In this section, the moment generating function is used to represent the distribution of the probabilities of a random variable. In addition, this function can be used to generate any moment of a distribution. There are other uses of the moment generating function that are seen later (see Section 7.4).
Definition 3.15  The moment generating function \( M_X(t) \) of the \( X \) is defined as

\[
M_X(t) = E(\exp(tX)) = \begin{cases} 
\sum_{x} \exp(tx)p_X(x) & \text{for } X \text{ discrete with pf } p_Y(y) \\
\int_{-\infty}^{\infty} \exp(tx)f_X(x) & \text{for } X \text{ continuous with pdf } f_X(x)
\end{cases}
\]

The mgf is defined as an infinite series or an infinite integral. Such an expression may not always exist (that is, converge to a finite value) for all values of \( t \). Hence it may happen that the mgf is not defined for all values of \( t \). Note that the mgf always exists for \( t = 0 \); in fact \( M_X(0) = 1 \).

It turns out that provided the mgf is defined for some values of \( t \) other than zero it uniquely defines a probability distribution and we can use it to generate the moments of the distribution as described in Theorem 3.16 below.

Example 3.13  Consider the random variable \( Y \) with pdf

\[
f_Y(y) = \begin{cases} 
\exp(-y) & \text{for } y > 0 \\
0 & \text{elsewhere.}
\end{cases}
\]

The mgf is given by

\[
M_Y(t) = E[\exp(tY)] = \int_{0}^{\infty} \exp(ty) \exp(-y) \, dy = \int_{0}^{\infty} \exp[y(t-1)] \, dy = \frac{1}{t-1} \exp\{y(t-1)\}\bigg|_{y=\infty}^{y=0} = (1-t)^{-1}
\]

provided \( t - 1 < 0 \); that is, \( t < 1 \). (If \( t > 1 \), the integral does not converge. For example, if \( t = 2 \), we would have

\[
\frac{1}{2-1} \exp(y)\bigg|_{y=\infty}^{y=0} = \exp(0) - \exp(\infty)
\]

which does not converge.)
3.5. Moment generating function

Example 3.14 Consider the pf of $X$, the outcome in tossing a fair die (see Example 3.5). The mgf of $X$ can be expressed

$$M_X(t) = E(\exp(tX)) = \sum_x \exp(tx)p_X(x)$$

$$= \frac{1}{6} \left(1 + e^{2t} + e^{3t} + e^{4t} + e^{5t} + e^{6t}\right)$$

which exists for all values of $t$.

3.5.1 Using the mgf to generate moments

Replacing $e^{xt}$ by its series expansion (see Section 1.10) in the definition of the mgf gives

$$M_X(t) = \sum_x \left(1 + xt + \frac{x^2t^2}{2!} + \ldots\right)P(X = x)$$

$$= 1 + \mu'_1 t + \mu'_2 \frac{t^2}{2!} + \mu'_3 \frac{t^3}{3!} + \ldots$$

Note that to obtain the result the order of summations has been interchanged, and in cases where this interchange is permitted, the $r$th moment of a distribution about the origin is the coefficient of $t^r/r!$ in the series expansion of $M_X(t)$.

$$M'_X(t) = \sum_x xe^{xt}P(X = x)$$

$$M''_X(t) = \sum_x x^2 e^{xt}P(X = x),$$

and for each positive integer $r$,

$$M^{(r)}_X(t) = \sum_x x^r e^{xt}P(X = x).$$

On setting $t = 0$ we find that $M'_X(0) = E(X)$, $M''_X(0) = E(X^2)$, and for each positive integer $r$,

$$M^{(r)}_X(0) = E(X^r). \quad (3.2)$$

This result is summarised in the following theorem.

Theorem 3.16 The $r$th moment $\mu'_r$ of the distribution of the rv $X$ about the origin is given by either
(i) the coefficient of $t^r/r!, r = 1, 2, 3, \ldots$ in the power series expansion of $M_X(t)$, or

(ii) $\mu'_r = M^{(r)}(0) = \left[ \frac{d^r M(t)}{dt^r} \right]_{t=0}$ where $M_X(t)$ is the mgf of $X$.

Example 3.15 Continuing Example 3.13, the mean and variance of $Y$ can be found from the mgf. To find the mean, first find

$$\frac{d}{dt}M_Y(t) = (1 - t)^{-2}.$$ Setting $t = 0$ gives the mean as $E(Y) = 1$. Likewise,

$$\frac{d^2}{dt^2}M_Y(t) = 2(1 - t)^{-3}.$$ Setting $t = 0$ gives $E(Y^2) = 2$. The variance is therefore $\text{var}[Y] = 2 - 1^2 = 1$.

Once the moment generating function has been computed, raw moments can be computed using

$$E(Y^r) = \mu'_r = \left. \frac{d^r}{dt^r}M_Y(t) \right|_{t=0}.$$  

3.5.2 Some useful results

The moment generating function can be used to derive the distribution of a function of a rv (see Section 7.4). The following theorems are valuable for this task.

**Theorem 3.17** If the rv $X$ has mgf $M_X(t)$ and $Y = aX + b$ where $a$ and $b$ are constants, then the mgf of $Y$ is

$$M_Y(t) = E(\exp\{t(aX + b)\}) = \exp(bt)M_X(at)$$

**Theorem 3.18** If $X_1, X_2, \ldots, X_n$ are $n$ independent random variables, where $X_i$ has mgf $M_{X_i}(t)$, then the mgf of $Y = X_1 + X_2 + \cdots + X_n$ is

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t).$$
3.5. Moment generating function

The proofs are left as an exercise.

Note that in the special case when all the rvs are independently and identically distributed in Theorem 3.18,

\[ M_Y(t) = [M_{X_i}(t)]^n. \]

**Example 3.16** Consider the random variable \( X \) with pf

\[ p_X(x) = 2(1/3)^x \quad \text{for } x = 1, 2, 3, \ldots \]

and zero elsewhere. Determine

(a) the mgf of \( X \)
(b) the mgf of \( Y = (X - 2)/3 \).

**Solution**

(a) The mgf of \( X \) is

\[
M_X(t) = \sum_{x:p(x)>0} \exp(tx)p_X(x) \\
= \sum_{x=1}^{\infty} \exp(tx)2(1/3)^x \\
= 2 \sum_{x=1}^{\infty} (\exp(t)/3)^x \\
= 2\left\{ \frac{\exp(t)}{3} + \left(\frac{\exp(t)}{3}\right)^2 + \left(\frac{\exp(t)}{3}\right)^3 + \ldots \right\} \\
= 2\frac{\exp(t)/3}{1 - \exp(t)/3} \\
= \frac{2\exp(t)}{3 - \exp(t)}
\]

where \( \sum_{y=1}^{\infty} a^y = \frac{a}{1-a} \) for \( a < 1 \) has been used (see (1.14)); here \( a = \exp(t)/3 \).

(b) From Theorem 3.17 with \( a = 1/3 \) and \( b = -2/3 \) we have

\[
M_Y(t) = \exp(-2t/3)M_X(t/3) = \frac{2\exp((-t)/3)}{3 - \exp(t/3)}
\]

In practice, rather than identify \( a \) and \( b \) and remember Theorem 3.17, problems such as this one are best solved directly from
the definition of the mgf; viz

\[ M_Y(t) = E(\exp(tY)) = E(\exp(t(X - 2)/3)) \]
\[ = E(\exp(tX/3 - 2t/3)) \]
\[ = \exp(-2t/3)M_X(t/3) \]
\[ = \frac{2\exp((-t)/3)}{3 - \exp(t/3)} \]

### 3.5.3 Determining the distribution from the mgf

It can be proven that the mgf (if it exists) completely determines the distribution of a rv. Hence it should be possible, given a mgf, to deduce the probability function. In the discrete case, this is reasonably easy; it is very difficult in the continuous case which is not presented here.

In the discrete case, the mgf is defined as

\[ M_X(t) = E(\exp(tX)) = \sum_X e^{tx}p_X(x) \]

for \( X \) discrete with pf \( p_X(x) \). This can be expressed as

\[ M_X(t) = \exp(tx_1)p_X(x_1) + \exp(tx_2)p_X(x_2) + \ldots \]
\[ = \exp(tx_1)P(X = x_1) + \exp(tx_2)P(X = x_2) + \ldots \]

and so the probability function of \( Y \) can be deduced from the mgf.

**Example 3.17** Suppose a discrete random variable \( D \) has the mgf

\[ M_D(t) = \frac{1}{3} \exp(2t) + \frac{1}{6} \exp(3t) + \frac{1}{12} \exp(6t) + \frac{5}{12} \exp(7t). \]

Then, by the definition of the mgf in the discrete case given above, the coefficient of \( t \) in the exponential indicates values of \( D \), and the coefficient indicates the probability of that value of \( y \).

\[
M_D(t) = \begin{cases} 
    \frac{1}{3} \exp(2t) & \text{if } D=2 \\
    \frac{1}{6} \exp(3t) & \text{if } D=3 \\
    \frac{1}{12} \exp(6t) & \text{if } D=6 \\
    \frac{5}{12} \exp(7t) & \text{if } D=7 
\end{cases}
\]

\[ = P(D=2)\exp(2t) + P(D=3)\exp(3t) + P(D=6)\exp(6t) + P(D=7)\exp(7t), \]
3.6 Tchebysheff’s inequality

so the pf is

\[
p_D(d) = \begin{cases} 
  \frac{1}{3} & \text{for } d = 2 \\
  \frac{1}{6} & \text{for } d = 3 \\
  \frac{1}{12} & \text{for } d = 6 \\
  \frac{5}{12} & \text{for } d = 7 \\
  0 & \text{otherwise} 
\end{cases} 
\]

(Of course, it is easy to check by computing the mgf for \( D \) from the
pf found above; you should get the original mgf.)

3.6 Tchebysheff’s inequality

Reading 3.4 DGS, Section 4.8 (only the material on
Tchebysheff’s inequality is relevant at the moment); WMS,
Sections 3.11 and 4.10.

An inequality that applies to any probability distribution is sometimes useful
in theoretical work or to provide bounds on probabilities.

Theorem 3.19 Let \( X \) be a rv with finite mean \( \mu \) and variance \( \sigma^2 \). Then
for any positive \( k \) we have

\[
P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad (3.3)
\]

or, equivalently

\[
P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \quad (3.4)
\]

Proof The proof for the continuous case only is given here.
Let \( X \) be continuous with pdf \( f(x) \). For some \( c > 0 \), consider

\[
\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx \\
= \int_{-\infty}^{\mu - \sqrt{c}} (x - \mu)^2 f(x) \, dx + \int_{\mu - \sqrt{c}}^{\mu + \sqrt{c}} (x - \mu)^2 f(x) \, dx + \int_{\mu + \sqrt{c}}^{\infty} (x - \mu)^2 f(x) \, dx \\
\geq \int_{-\infty}^{\mu - \sqrt{c}} (x - \mu)^2 f(x) \, dx + \int_{\mu + \sqrt{c}}^{\infty} (x - \mu)^2 f(x) \, dx
\]
since the second integral is non-negative. Now \((x - \mu)^2 \geq c\) if \(x \leq \mu - \sqrt{c}\)
or \(x \geq \mu + \sqrt{c}\). So in both the above remaining integrals we may replace
\((x - \mu)^2\) by \(c\) without altering the direction of the inequality. Thus,
\[
\sigma^2 \geq c \int_{-\infty}^{\mu - \sqrt{c}} f(x) \, dx + c \int_{\mu + \sqrt{c}}^{\infty} f(x) \, dx
\]
\[= cP(X \leq \mu - \sqrt{c}) + cP(X \geq \mu + \sqrt{c})\]
\[= cP(|X - \mu| \geq \sqrt{c}).\]
Putting \(\sqrt{c} = k\sigma\), we obtain (3.19).

Note that if we have the probability function or pdf of a random variable \(X\) we can find \(E(X)\) and \(\text{var}(X)\) but the converse is not true. That is, from a knowledge of \(E(X)\) and \(\text{var}(X)\) we cannot reconstruct the probability distribution of \(X\) and hence cannot compute probabilities such as \(P(|X - \mu| \geq k\sigma)\). But using Tchebysheff’s inequality we can find a useful bound to either the probability outside or inside of \(\mu \pm k\sigma\).

### 3.7 Self-assessment exercises

The following exercises are designed to provide practice at problem-solving based on the material in this module. Solutions are provided at the end of the module. Additional exercises are available in the next section and in the textbook.

**Ex. 3.1** The rv \(Y\) is defined as

\[
f_Y(y) = \begin{cases} 
2y + k & \text{for } 0 \leq y \leq 1 \\
0 & \text{elsewhere}
\end{cases}
\]

(a) Find a value for \(k\).
(b) Plot the pdf of \(Y\).
(c) Compute \(E(Y)\).
(d) Compute \(\text{var}(Y)\).
(e) Compute \(P(X > 0.15)\).

**Ex. 3.2** The rv \(N\) is defined as

\[
p_N(n) = \begin{cases} 
1/3 & \text{for } n = 1 \\
1/6 & \text{for } n = 2 \\
\alpha & \text{for } n = 3 \\
0 & \text{otherwise}
\end{cases}
\]
3.7. Self-assessment exercises

(a) Find the value of $\alpha$.
(b) Compute the mean and variance of $M$.
(c) Find the mgf for $M$.
(d) Compute the mean and variance of $M$ from the mgf.

Ex. 3.3 The rv $A$ has mean 3 and variance 2. The rv $B$ has mean 4 and variance 3. Assume $A$ and $B$ are independent. Find

(a) $E(A + B)$
(b) $\text{var}(A + B)$
(c) $E(2A - 3B)$
(d) $\text{var}(2A - 3B)$

Ex. 3.4 The mgf of the discrete rv $Z$ is

$$M_Z(t) = [0.3 \exp(t) + 0.7]^2$$

(a) Compute the mean and variance of $Z$.
(b) Find the pf of $Z$.

Ex. 3.5 The mgf of $G$ is $M_G(t) = (1 - \beta t)^{-\alpha}$. Find the mean and variance of $G$.

Ex. 3.6 Suppose $X_1, X_2, \ldots, X_n$ are independently distributed rvs, each with mean $\mu$ and variance $\sigma^2$. Define the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(a) Prove that $E(\bar{X}) = \mu$.
(b) Find the variance of $\bar{X}$.

Ex. 3.7 Suppose that the pdf of $X$ is

$$f_X(x) = \begin{cases} 2(1 - x) & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the $r$th raw moment of $X$.
(b) Find $E((X + 3)^2)$ using the previous answer.
(c) Find the variance of $X$.

Ex. 3.8 (Computer exercise) Generate a random sample of size 120 from the distribution in Example 3.5. Find the mean and standard deviation of the sample. Compare the results with those found in Example 3.5. Draw a histogram for the sample. Are they as you would expect?
3.8 Exercises

Ex. 3.9 For a discrete random variable $X$ with probability function

$$P(X = x) = \frac{1}{5}, \quad x = -1, 0, 1, 2, 3,$$

find (a) $E(X)$, (b) $E(X^2)$, (c) $\text{var}(X)$, (d) $E(|X|)$ and (e) the median of $X$.

Ex. 3.10 $X$, $Y$ and $Z$ are uncorrelated random variables with expected values $\mu_x$, $\mu_y$ and $\mu_z$ and standard deviations $\sigma_x$, $\sigma_y$ and $\sigma_z$. $U$ and $V$ are defined by

$$U = X - Z$$
$$V = X - 2Y + Z$$

(a) Find the expected values of $U$ and $V$.
(b) Find the variance of $U$ and $V$.
(c) Find the correlation coefficient of $U$ and $V$.
(d) Under what conditions on $\sigma_x$, $\sigma_y$ and $\sigma_z$ are $U$ and $V$ uncorrelated?

Ex. 3.11 If $(X, Y)$ has joint probability function given by

$$P(X = x, Y = y) = \frac{|x - y|}{11}, \quad \text{for } x = 0, 1, 2; y = 1, 2, 3$$

find (a) $E(X \mid Y = 2)$ (b) $E(Y \mid X \geq 1)$

Ex. 3.12 A rv $X$ is defined as

$$f_X(x) = \begin{cases} 
\frac{k}{x} & \text{for } 1 < x < 2 \\
0 & \text{otherwise}
\end{cases}$$

(a) Determine a value for $k$.
(b) Compute the mean and variance of $X$.
(c) Determine the distribution function for $X$.

Ex. 3.13 The rv $X$ has the mgf

$$M_X(t) = \exp\left(\mu t + \frac{t^2\sigma^2}{2}\right)$$

(a) Find the mean and variance of $X$.
(b) Find the mgf of $X + Y$ if $Y$ has the same mgf as $X$. 
Ex. 3.14 Consider a random variable $W$ for which $P(W = c) = k$, $P(W = -c) = 2k$ and $P(W = 0) = 3k$, and is zero elsewhere. Find the mean and variance of $W$.

Ex. 3.15 Consider the random variable $X$ which has the density function

$$f_X(x) = \frac{k}{x^2}$$

for $3 < x < \infty$.

(a) Find the value of $k$.
(b) Sketch the density function over the range $0 \leq x \leq 6$.
(c) Using the integral definition of the mean, show that the mean cannot be computed.
(d) Determine $P(X > 3 \mid X < 4)$.
(e) Find the (cumulative) distribution function for $X$.
(f) Using the distribution function, determine the median of $X$.

Ex. 3.16 The random variable $W$ has the probability function

$$p_W(w) = \frac{5w}{6(1 + w^2)}$$

for $w = 1, 2, 3$, and is zero for other values of $W$.

(a) Find the moment generating function for $W$.
(b) Using the moment generating function, calculate $E(W)$ and $\text{var}(W)$.
(c) Determine and sketch the distribution function of $W$.

Ex. 3.17 The distribution

$$f_W(w) = \frac{1}{105}w^2(6 - w)$$

is defined for $w = 1, 2, 3, 4, 5$, and is zero for other values of $W$.

(a) Find the moment generating function for $W$.
(b) Using the moment generating function, calculate $E(W)$ and $\text{var}(W)$.
(c) Determine and sketch the distribution function of $W$.

Ex. 3.18 Consider a discrete random variable $X$ defined on $x = 0, 1, 2, \ldots$

(a) Show that $P(X = x) = P(X \geq x) - P(X \geq x + 1)$. (HINT: Write out the right-hand side for the first few values of $x$.)
(b) Hence show that \( E(X) = \sum_{x=0}^{\infty} x[P(X \geq x) - P(X \geq x + 1)]. \)

(c) Use the previous result to show that \( E(X) = \sum_{x=0}^{\infty} P(X > x). \)

(HINT: Write out the summation from (b) for the first few values of \( x \).)

**Ex. 3.19** The probability density function of a random variable \( X \) is given by

\[
f(x) = \begin{cases} 
0 & x \leq 1 \\
\frac{1}{4} & 1 < x \leq 2 \\
0 & 2 < x \leq 3 \\
\frac{1}{2} & 3 < x \leq 4 \\
0 & 4 < x \leq 5 \\
\frac{1}{4} & 5 < x \leq 6 \\
0 & x > 6 
\end{cases}
\]

Find (a) \( E(X) \) (b) \( E(X^2) \) (c) \( \text{var}(X) \) (d) the median of \( X \).

**Ex. 3.20** (a) Find the mgf of the distribution with pdf

\[
f(x) = \begin{cases} 
\frac{1}{2}e^{x} & -\infty < x < 0 \\
\frac{1}{2}e^{-x} & 0 < x < \infty 
\end{cases}
\]

What restrictions must be placed on the value of \( t \)?

(b) Hence find \( E(X^r) \), \( r = 1, 2, 3, \ldots \).

(c) Find the coefficients of skewness and kurtosis.

**Ex. 3.21** (a) For the distribution in Question 3.20 find \( P(|X| \geq 3) \).

(b) Use Tchebychev’s inequality to get an upper bound for this probability. Comment on the bound.

**Ex. 3.22** Prove that for a continuous random variable \( X \) which has a distribution that is symmetric about 0 then \( M_X(t) = M_{-X}(t) \). Hence prove that for such a rv all odd moments about the origin are zero.

**Ex. 3.23** If \( X \) has mgf given by \( M_X(t) = \frac{1}{1-t} \),

(a) Find the moment generating function of \( Y = 4X - 3 \).

(b) Using the moment generating functions find \( E(X), E(Y), \text{var}(X) \) and \( \text{var}(Y) \).
3.8. Exercises

(c) Show the relationship between the means and variances in (d) is as you would expect by finding $E(Y)$ and $\text{var}(Y)$ directly from the definition of $Y$.

(d) Hence find the coefficients of skewness and kurtosis for this distribution.

Ex. 3.24 The random variables $X_1, X_2,$ and $X_3$ have means $\mu_1 = 5, \mu_2 = 3$ and $\mu_3 = 6$, standard deviations $\sigma_1 = 2, \sigma_2 = 3$ and $\sigma_3 = 4$ and correlations $\rho_{12} = -\frac{1}{6}, \rho_{13} = \frac{1}{6}$ and $\rho_{23} = \frac{1}{2}$. If the random variables $U$ and $V$ are defined by $U = 2X_1 + X_2 - X_3$ and $V = X_1 - 2X_2 - X_3$ find

(a) $E(U)$
(b) $\text{var}(U)$
(c) $\text{cov}(U,V)$

Ex. 3.25 Consider the distribution

$$f_Y(y) = \frac{2}{y^2} \quad y \geq 2.$$  

(a) Show that the mean of the distribution does not exist.
(b) Show that the variance does not exist.
(c) Plot the probability density function over a suitable range.
(d) Plot the distribution function over a suitable range.
(e) Determine the median of the distribution.
(f) Determine the interquartile range of the distribution. (The interquartile range is a measure of spread, and is calculated as the difference between the third quartile and the first quartile. The first quartile is the value below which 25% of the data lie; the third quartile is the value below which 75% of the data lie.)
(g) Find $P(Y > 4 \mid Y > 3)$.

Ex. 3.26 Consider the discrete distribution

$$p_X(x) = k \frac{x}{x + 1}, \quad \text{for } x = 1, 2, 3, 5.$$  

(a) Find $k$.
(b) Find the moment generating function of $X$.
(c) Use the mgf to determine the mean and variance of $X$.
(d) Plot the distribution function of $X$. 
Module 3. Mathematical expectation

Ex. 3.27 Consider the random variable $X$ which has a probability function

$$p_X(x) = \kappa \frac{x}{x^2 + 1}$$

for $x = 1, 2, 3$.

where $\kappa$ is a positive constant.

(a) Determine the value of the constant $\kappa$.
(b) Calculate the moment generating function for $X$.
(c) Find the mean of $X$.
(d) Find $F_X(x)$, the distribution function of $X$.

Ex. 3.28 Let $Z$ be a random variable such that

$$P(Z = c) = P(Z = -c) = \frac{1}{2}$$

for some constant $c > 0$.

(a) Determine the expected value and the variance of $Z$.
(b) Determine the probability that $Z$ is within two standard deviations of the mean.

Ex. 3.29 A random variable $Y$ has the pdf

$$f_Y(y) = \begin{cases} 
    y^2 & \text{for } 0 < y < 1 \\
    1 - ky & \text{for } 1 < y < 2
\end{cases}$$

(a) Find a value for $k$.
(b) Plot the probability density function for $Y$.
(c) Find $E[Y]$.
(d) Find the (cumulative) distribution function for $Y$.

Ex. 3.30 The discrete rv $W$ has the pf

$$p_W(w) = \begin{cases} 
    k \log \left( \frac{w+1}{w} \right) & \text{for } w = 2, 3, 4, 5 \\
    0 & \text{otherwise}
\end{cases}$$

(a) Show $k = 1/ \log(3)$.
(b) Deduce the mgf of $W$.
(c) Calculate $E(W)$ and $\text{var}(W)$ using the mgf.
(d) Determine and sketch the distribution function of $W$.

Ex. 3.31 In a particular application, the distribution

$$f_Z(z) = \kappa \exp\{1 - \cos(z)\} \quad 0 \leq z \leq 2\pi$$

is needed, where $\kappa$ is a constant.
3.8. Exercises

(a) Use a numerical integration technique to determine the value of \( \kappa \) accurate to at least one decimal place.

(b) Plot the density function of \( Z \), and deduce (but do not calculate) its mean.

(c) Determine \( P(0 < Z < 1) \) accurate to at least two decimal places.

Ex. 3.32 Find the mgf for the random variable with probability function

\[
P(X = 1) = P(X = -1) = \frac{1}{2}.
\]

(a) Hence show that

\[
E(X^r) = \begin{cases} 
 1 & \text{if } r \text{ even} \\
 0 & \text{if } r \text{ odd}
\end{cases}
\]

(b) Find the mean and variance of \( X \).

Ex. 3.33 The Cauchy distribution has the pdf

\[
f_X(x) = \frac{1}{\pi(1 + x^2)} \quad \text{for } -\infty < x < \infty
\]

(a) By integrating \( xf_X(x) \), show that the mean does not exist for the Cauchy distribution.

(b) Hence show that the variance does not exist for the Cauchy distribution.

Ex. 3.34 (Computer exercise)

(a) A rv \( X \) is uniformly distributed on the interval \([0, 1]\) with density function, \( f_X(x) = 1 \) for \( x \in [0, 1] \). If

\[
g(x) = e^x
\]

find \( E[g(X)] \).

(b) Generate a random sample of size 1000 from a uniform distribution on \([0, 1]\) (\texttt{runif(n=1000,min=0,max=1)} and estimate \( \int_0^1 e^x \, dx \).

Hence obtain an estimate of \( e \).

(c) Let \( h(x) = 1/(1 + x^2) \). Find \( E[h(X)] \).

(d) Estimate \( \int_0^1 (1 + x^2)^{-1} \, dx \) by Monte-Carlo integration and hence obtain an estimate for \( \pi \).
3.9 Some answers and hints

3.1 (a) $k = 0$ (c) $E(Y) = 2/3$ (d) $E(Y^2) = 1/2$ so $\text{var}(Y) = 1/18$ (e) $0.9775$

3.2 (a) $\alpha = 1/2$ (b) $E(N) = 13/6$, $\text{var}(N) \approx 0.8056$ (c) $M_N(t) = \exp(t)/3 + \exp(2t)/6 + \exp(3t)/2$

3.3 (a) $7$ (b) $5$ (c) $-6$ (d) $35$

3.4 $M'_Z(t) = 0.6 \exp(t)[0.3 \exp(t)+0.7]$ so $E(Z) = 0.6$; $M''_Z(t) = 0.18 \exp(2t) + 0.6 \exp(t)[0.3 \exp(t)+0.7]$ so $E(Z^2) = 0.78$, hence so $\text{var}(Z) = 0.42$ (be careful with the derivatives here!); expand the quadratic and find $P(Z = 0) = 0.49$, $P(Z = 1) = 0.42$, $P(Z = 2) = 0.09$.

3.5 $M'_G(t) = \alpha\beta(1-\beta t)^{-\alpha-1}$ so $E(G) = \alpha\beta$; $M''_G(t) = \alpha\beta^2(\alpha+1)(1-\beta t)^{-\alpha-2}$ so $E(G^2) = \alpha\beta^2(\alpha+1)$ and $\text{var}(G) = \alpha\beta^2$.

3.6 (a) Apply Theorem 6.14. (b) $\text{var}(X) = \sigma^2/n$.

3.7 (a) $\mu'_1 = E(X^r) = \int_{x=0}^1 x^r 2(1-x) \, dx = 2 \left[ \left( \frac{x^{r+1}}{r+1} - \frac{x^{r+2}}{r+2} \right) \bigg|_0^1 \right] = 2 \left[ \frac{1}{r+1} - \frac{1}{r+2} \right]$.

(b) Expanding, $E((X+3)^2) = E(X^2) + 6E(X) + 9$. Now, $E(X) = \mu'_1 = 1/3$ from above, and $E(X^2) = \mu'_2 = 1/6$ from above. Hence $E((X+3)^2) = 67/6$.

(c) $\text{var}(X) = E(X^2) - E(X)^2 = 1/18$.

3.8 face<-sample(1:6,size=120,replace=T)
hist(face)

3.9 (a) $E(X) = 1$ (b) $E(X^2) = 3$ (c) $\text{var}(X) = 2$ (d) $E(|X|) = 7/5$ (e) median $= 1$

3.10 (a) and (b) Use results from Section 6.5 remembering $X$, $Y$ and $Z$ are independent.

(c) Find the covariance by rewriting $E(U - E(U)[V - E(V)]$ in terms of $(X - \mu_x)$, $(Y - \mu_y)$ and $(Z - \mu_z)$. Then, for example, $E(X - \mu_x)^2 = \sigma_x^2$ and $E(X - \mu_x)(Y - \mu_y) = \text{cov}(X,Y) = 0$ (independence condition).

3.11 (a) First find $P(X = x \mid Y = 2)$, $x = 0, 1, 2$, then $E(X \mid Y = 2) = 1/3$.
(b) Find $P(Y = y \mid X \geq 1)$, $y = 1, 2, 3$, then $E(Y \mid X \geq 1) = 12/5$

3.19 (a) $E(X) = 3.5$ (b) $E(X^2) = 43/4$ (c) $\text{var}(X) = 2.08333$ (d) median $= 3.5$

3.20 (a) $M_X(t) = \frac{1}{1-t^2}$, $|t| < 1$ (b) $E(X^r) = r!$ for $r$ even, $0$ otherwise.
(d) $\gamma_1 = 0$, $\gamma_2 = 3$
3.9. Some answers and hints

3.21 (a) 0.04979 (b) \( \sigma^2 = 2 \), thus \( k = 3/\sqrt{2} \) and upper bound = 2/9.

3.22 Begin with Definition 3.15 for \( M_X(t) \) and use fact that if a distribution is symmetric about 0 then \( f_X(x) = f_X(-x) \) using symmetry. Transform the resulting integral.

3.23 (a) Using Theorem 3.17, \( M_Y(t) = \frac{e^{3t}}{1 - 4t} \)

(b) Expand \( M_X(t) \) and \( M_Y(t) \) in powers of \( t \) and find the coefficients of \( t/1! \) and \( t^2/2! \). (Only coefficients of \( t \) and \( t^2 \) are needed.)
\[ E(X) = 1, \var(X) = 1, E(Y) = 1, \var(Y) = 16 \]
\( \gamma_1 = 2, \gamma_2 = 6. \)

3.24 (a) 7 (b) \( 19 \frac{2}{3} \) (c) 11

If you are R minded and understand matrices, try

```r
r12 <- -1/6; r13 <- 1/6; r23 <- 1/2
R <- diag(1, nrow = 3)
R[1, 2] <- R[2, 1] <- r12
R[1, 3] <- R[3, 1] <- r13
sigma <- diag(c(2, 3, 4))
VX <- sigma %*% R %*% sigma
a <- c(2, 1, -1)
b <- c(1, -2, -1)
VU <- t(a) %*% VX %*% a
cUV <- t(a) %*% VX %*% b
print(c(VU, cUV))
```

3.33 For the mean, \( E(X) = (1/\pi) \int_{-\infty}^{\infty} \frac{x}{1+x^2} \, dx = (2/\pi) \log(1+x^2) \bigg|_{-\infty}^{\infty} \).
Now, \( \log(\infty) \) is not defined, so the mean is not defined.

3.34 (a) \( E(X) = e - 1 \) (b) See R Exercise 5.35. (c) \( \int (1 + x^2)^{-1} \, dx = \tan^{-1}(x) \), \( E[h(X)] = \pi/4 \) (Don’t expect accuracy to more than 1 dec pl with \( N = 1000. \))
Module 3. Mathematical expectation
Module 4

Standard discrete distributions

Contents

4.1 Discrete uniform distribution ......................................... 118
4.2 Binomial distribution ...................................................... 121
4.3 Poisson distribution ........................................................ 126
  4.3.1 Relationship to the binomial distribution ...................... 130
4.4 Geometric distribution ..................................................... 131
4.5 Negative binomial distribution ......................................... 134
  4.5.1 Alternative parameterization ..................................... 138
  4.5.2 The gamma function ................................................. 139
4.6 Hypergeometric distribution ............................................ 142
4.7 Self-assessment exercises ................................................. 145
4.8 Exercises ................................................................. 146
4.9 Some answers and hints .................................................. 150

Module objectives

On completion of this module, students should be able to:
Module 4. Standard discrete distributions

- be familiar with the probability functions and underlying parameters of uniform, binomial, geometric, negative binomial, Poisson, hypergeometric and multinomial random variables
- know the basic properties of the above discrete distributions covered in this module
- apply these discrete distributions as appropriate to problem solving

In this module, some popular discrete distributions are discussed. Properties such as definitions and applications are considered.

4.1 Discrete uniform distribution

If a discrete rv $Y$ can assume $k$ different and distinct values with equal probability, then $Y$ is said to have a discrete uniform distribution. This is one of the simplest discrete distributions.

**Definition 4.1** If a rv $X$ with range space $\{a, a + 1, a + 2, \ldots, b\}$ where $a$ and $b$ ($a < b$) are integers has the pf

$$p_X(x) = \frac{1}{k} \text{ for } x = a, a + 1, \ldots, b$$

then $X$ has a discrete uniform distribution. We write $X \sim U(a, b)$ or $X \sim \text{unif}(a, b)$.

This distribution is also called the rectangular distribution.

A plot of the pf for a discrete uniform distribution is shown in Figure 4.1.

**Example 4.1**

(i) Let $X$ be the number of spots showing after a single throw of a fair die.
Then $X \sim \text{unif}(1, 6)$.

(ii) For an experiment involving selection of a single-digit number from a table of random digits, the number chosen, $X$, has probability distribution unif(0, 9).
4.1. Discrete uniform distribution

The following are the basic properties of the discrete uniform distribution.

**Theorem 4.2** If $X \sim \text{unif}(a, b)$ then

1. $E(X) = (a + b)/2$
2. $\text{var}(X) = \{(b - a)(b - a + 2)\}/12$

**Proof** A smart approach here is to consider $Y = X - a$ rather than $X$ itself. The point is that $E(X) = E(Y) + a$ and $\text{var}(Y) = \text{var}(X)$ but it’s easier to find $E(Y)$ and $\text{var}(Y)$ directly than $E(X)$ and $\text{var}(X)$.

1. Noticing that $Y \sim \text{unif}(0, b - a)$ we have using (1)

   
   
   $E(Y) = \sum_{i=0}^{b-a} \frac{i}{b-a+1}$
   
   
   $= \frac{1}{b-a+1}(0 + 1 + 2 + \cdots + (b - a))$
   
   
   $= \frac{(b - a)(b - a + 1)}{2(b-a+1)}$
   
   
   $= \frac{b - a}{2}$

   Therefore $E(X) = E(Y) + a = \frac{b-a}{2} + a = \frac{a+b}{2}$.

2. Now $\text{var}(Y) = E(Y^2) - E(Y)^2$ in which

   
   
   $E(Y^2) = \sum_{i=0}^{b-a} \frac{i^2}{b-a+1}$
   
   
   $= \frac{1}{b-a+1}(0^2 + 1^2 + 2^2 + \cdots + (b - a)^2)$
   
   
   $= \frac{1}{b-a+1} \frac{(b-a)(b-a+1)(2(b-a)+1)}{6}$
   
   
   $= \frac{(b-a)(2(b-a)+1)}{6}$

   where (1) has been used. Therefore

   
   
   $\text{var}(X) = \text{var}(Y) = \frac{(b-a)(2(b-a)+1)}{6} - \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)(b-a+2)}{12}$

♠
Module 4. Standard discrete distributions

Example 4.2 Oz Lotto, like many lottery games, challenges players to match randomly chosen numbers. Oz Lotto randomly picks numbers between 1 and 45 (inclusive). Each number should have an equal chance of selection, so the discrete uniform distribution $U(1, 45)$ is appropriate. Figure 4.2 shows the proportion of times each number has been chosen from Game 1 to Game 473. The data appear to be approximately uniform as expected.

If the data follow a discrete uniform distribution exactly, the mean should be $(1 + 45)/2 = 23$ and the variance $(44 \times 46)/12 = 168.66$. Using the sample data, the sample mean is computed to be 23.29 and the sample variance as 166.36; these are both very close to what one might expect for a $U(1, 45)$ random variable.
A binomial distribution is used in a situation where the same ‘experiment’ is repeated a number of times, and one of two outcomes is observed. A simple example is tossing a coin ten times and observing if a head falls. The same experiment (tossing the coin) is repeated, there are only two outcomes on each trial (a head or a tail), and the probability of a head remains constant on each trial.

Consider tossing a die five times and observing the number of times a 1 is rolled. The probability of observing three 1’s can be found as follows. In the five tosses, a 1 must appear three times; there are \( \binom{5}{3} \) ways of allocating on which of the five rolls they will appear. In the five rolls, 1 must appear three time with probability \( \frac{1}{6} \); the other two rolls must produce another number with probability \( \frac{5}{6} \). So the probability will be

\[
\binom{5}{3} \left( \frac{1}{6} \right)^3 \left( \frac{5}{6} \right)^2 = 0.032,
\]

assuming independence of the events. In this way, the binomial distribution can be developed.

Formally situations giving rise to a binomial distribution are defined in terms of Bernoulli trials.

**Definition 4.3** A Bernoulli trial is an experiment with only two possible outcomes, usually labelled ‘success’ and ‘failure’. The sample space can be denoted by \( S = \{s, f\} \).

A binomial situation arises if a sequence of Bernoulli trials is observed in each of which \( P(\{s\}) = p \) and \( P(\{f\}) = q \), where \( p + q = 1 \). If \( n \) such trials are conducted consider the random variable \( X \), where \( X \) is the number of successes in \( n \) trials. Now \( X \) will have value set \( R_X = \{0, 1, 2, \ldots, n\} \). It is assumed that \( p \) is constant from trial to trial, and that the \( n \) trials are independent.
Consider the event $X = r$ (where $0 \leq r \leq n$). This could correspond to the sample point
\[ S \ S \ S \ldots S \ S \ S \ F \ F \ldots F \ F \]
which is the intersection of $n$ independent events consisting of $r$ successes and $n - r$ failures and hence the probability is $p^r q^{n-r}$.

Every other sample point in the event $X = r$ will appear as a rearrangement of the $S$’s and $F$’s in the sample point described above and will therefore have the same probability. Now the number of distinct arrangements of the $r$ $S$’s and $(n-r)$ $F$’s is \( \binom{n}{r} \) so
\[
P(X = r) = \binom{n}{r} p^r q^{n-r}, \quad r = 0, 1, \ldots, n.
\]

Note that the sum of the probabilities is 1 as the binomial expansion of $(p + q)^n$ (see (1)) is just
\[
\sum_{r=0}^{n} \binom{n}{r} p^r q^{n-r} = (p + q)^n = 1
\]
since $p + q = 1$. So we have the following:

**Definition 4.4** Let $X$ be the number of successes in $n$ independent Bernoulli trials with $P($Success$) = p$ $(0 \leq p \leq 1)$ constant in each trial. Then $X$ is said to have a binomial probability distribution with parameters $n, p$ and $pf$ given by (4.1). We write $X \sim \text{bin}(n, p)$.

Figure 4.3 shows the pf for the binomial distribution for various parameter values.

**Example 4.3** A die is thrown 4 times. What is the probability of exactly 2 sixes?

There are 4 Bernoulli trials with $p = 1/6$. Let the random variable $X$ be the number of 6’s in 4 tosses. Then
\[
P(X = 2) = \binom{4}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2 = 150/1296
\]
4.2. Binomial distribution

The following are the basic properties of the binomial distribution.

**Theorem 4.5** If $X \sim \text{bin}(n, p)$ then

1. $E(X) = np$
2. $\text{var}(X) = np(1 - p) = npq$
3. $M_X(t) = (pe^t + q)^n$

**Proof** We make use of equation (4.2) in the following.
1.

\[ E(X) = \sum_{x=0}^{n} x \left( \frac{n}{x} \right) p^x q^{n-x} \]
\[ = \sum_{x=1}^{n} x \left( \frac{n-1}{x-1} \right) p^x q^{n-x} \]
\[ = np \sum_{x=1}^{n} \left( \frac{n-1}{x-1} \right) p^{x-1} q^{n-x} \]
\[ = np \sum_{y=0}^{n-1} \left( \frac{n-1}{y} \right) p^y q^{n-1-y} \quad \text{putting } y = x - 1. \]

Now the sum is 1 since a term in the sum is the probability of \( y \) successes in \((n-1)\) Bernoulli trials and the sum is over all values in \( R_X \). Note also that in the second line the sum is over \( x \) from 1 to \( n \) because for \( x = 0 \) we have a binomial coefficient of the form \( \binom{b}{a} \) with \( a < 0 \) which in this context we define to be zero. Thus,

\[ E(X) = np. \]

2. To find the variance, we use the computational formula \( \text{var}(X) = E(X^2) - [E(X)]^2 \). Firstly, to find \( E(X^2) \), write \( E(X^2) \) as \( E[X(X-1) + X] \) then write it as \( E[X(X-1)] + E(X). \)

\[ E(X^2) = \sum_{x=0}^{n} x(x-1)P(X = x) + np \]
\[ = \sum_{x=2}^{n} x(x-1) \frac{n(n-1)}{x(x-1)} \left( \frac{n-2}{x-2} \right) p^x q^{n-x} + np \]
\[ = \sum_{x=2}^{n} n(n-1) \left( \frac{n-2}{x-2} \right) p^x q^{n-x} + np \]
\[ = n(n-1)p^2 \sum_{y=0}^{n-2} \left( \frac{n-2}{y} \right) p^y q^{n-2-y} + np, \]

putting \( y = x - 2 \). For the same reason as before, the sum is 1, so

\[ E(X^2) = n^2 p^2 - np^2 + np \]

and hence

\[ \text{var}(X) = E(X^2) - [E(X)]^2 = n^2 p^2 - np^2 + np - n^2 p^2 = np(1 - p). \]
4.2. Binomial distribution

3. The mgf of $X$ is given by

$$M_X(t) = E(\exp(tX)) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} = (pe^t + q)^n$$

(Note that the smart operator would have proven this result first and then used it to prove (i) and (ii), making use of the methods in 3.5.1—try this as an exercise.)

Example 4.4 Wilks [34, p 68] notes that in the 200 years from 1796 to 1995, Cayuga Lake has frozen only in ten of those years. Since the lake is deep, it will only freeze during exceptionally cold weather. The probability the lake freezes during any one year can be estimated as $p = 10/200 = 0.05$.

Using this information, the number of times the lake will not freeze in ten randomly chosen years in given by the random variable $X$ where $X \sim \text{bin}(10, 0.95)$. The probability that the lake will not freeze in these ten years is $P(X = 10) = \binom{10}{10} 0.95^{10} 0.05^0 \approx 0.599$, or about 60%.

Note we could define the random variable $Y$ as the number of times the lake will freeze in the ten randomly chosen years. Then, $Y \sim \text{bin}(10, 0.05)$ and we would compute $P(Y = 0)$ and get the same answer.

If the number of ‘successes’ has a binomial distribution, so does the number of ‘failures’. Specifically if $X \sim \text{bin}(n, p)$, then $Y = n - X \sim \text{bin}(n, 1 - p)$

The terms ‘success’ and ‘failure’ are not literal. ‘Success’ simply means that the event is of interest. If the event of interest is the number of damaging tornadoes, this is still called a ‘success’.

Remember that a binomial situation requires the trials to be independent, and the probability of success to be a constant $p$ throughout the trials. For
example, drawing cards from a pack without replacing them is not a binomial situation; after drawing one card, the probabilities will then change for the drawing of the next card. In this case, the hypergeometric distribution could be used (see Section 4.6).

Binomial probabilities can sometimes be approximated using the normal distribution (see Section 5.2.7) or the Poisson distribution (see Section 4.3.1).

Binomial probabilities are calculated using

- \( \text{dbinom}(x = , \text{size} = , p = ) \) for \( P(X = x) \),
- \( \text{pbinom}(q = , \text{size} = , p = ) \) for \( P(X \leq q) \)
- \( \text{pbinom}(q = , \text{size} = , p = , \text{lower.tail} = F) \) for \( P(X > q) \)

In Rcmdr, \( P(X = x) \) is calculated from the menu for Binomial probabilities and \( P(X \leq q) \) is calculated using the menu for Binomial tail probabilities.

A long time ago, these were determined using tables such as described in the textbook.

### 4.3 Poisson distribution

**Reading 4.2** DGS, Section 5.4; WMS, Section 3.8.

The Poisson distribution is a commonly used distribution to model the number of occurrences of an event which occurs randomly in time or space.

The Poisson distribution arises as a result of assumptions that are made about a random experiment, and roughly these are:

- (a) events that occur in one time-interval (or region) are independent of those occurring in any other non-overlapping time-interval (or region)

- (b) for a small time-interval the probability that an event occurs in it is proportional to the length of the interval
4.3. Poisson distribution

(c) the probability that 2 or more events occur in a very small time-interval is so small that it can be neglected.

Whenever these assumptions are valid, or approximately so, the Poisson distribution is appropriate, and quite a number of natural phenomena fall into this category.

**Reading 4.3** Read WMS, Section 3.8; DGS, Section 5.4.

**Definition 4.6** A random variable $X$ is said to have a Poisson distribution if its pf is

$$p_X(x) = \frac{e^{-\mu} \mu^x}{x!} \text{ for } x = 0, 1, 2, \ldots$$

where the parameter is $\mu > 0$. We write $X \sim \text{Pois}(\mu)$.

The pf for a Poisson distribution for different values of $\mu$ is shown in Figure 4.4.

The following are the basic properties of the Poisson distribution.

**Theorem 4.7** If $X \sim \text{Pois}(\mu)$ then

1. $E(X) = \mu$
2. $\text{var}(X) = \mu$
3. $M_X(t) = \exp[-\mu\{1 - \exp(t)e^t\}]$.

**Proof** Result 3 is proven as follows:

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} e^{-\mu} \frac{\mu^x}{x!}$$

$$= e^{-\mu} \sum_{x=0}^{\infty} \frac{(\mu e^t)^x}{x!}$$

$$= e^{-\mu} \left[1 + \mu e^t + \frac{(\mu e^t)^2}{2!} + \ldots\right]$$

$$= e^{-\mu e^t} e^{\mu t}$$

$$= e^{-\mu(1-e^t)}.$$ 

Results 1 and 2 follow from differentiating the mgf.
Figure 4.4: The pf for the Poisson distribution for $\mu = 0.5$, 1, 2 and 5.

**Example 4.5** Customers enter a waiting line ‘at random’ at a rate of 4 per minute. Assuming that the number entering the line in any given time interval has a Poisson distribution, determine the probability that at least one customer enters the line in a given $\frac{1}{2}$-minute interval, we have (since $\mu = 2$)

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-2} = .865.$$  

**Example 4.6** Clarke [7] (quoted in Hand [16, Dataset 289]) gives the number of flying bomb hits on London during World War II in a 36 square kilometre area of South London. The area was gridded into 0.25 km
4.3. Poisson distribution

<table>
<thead>
<tr>
<th>Hits</th>
<th>Number</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>229</td>
<td>0.3976</td>
</tr>
<tr>
<td>1</td>
<td>211</td>
<td>0.3663</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
<td>0.1615</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>0.0608</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>0.0122</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.0017</td>
</tr>
</tbody>
</table>

Table 4.1: The number of flying bomb hits on London during World War II in a 36 square kilometre area of South London. The area was gridded into 0.25 km squares and the number of bombs falling in each grid was counted. The proportion of the 576 grid squares receiving 0, 1, … hits was also computed.

<table>
<thead>
<tr>
<th>Hits</th>
<th>Empirical prob</th>
<th>Poisson prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.398</td>
<td>0.394</td>
</tr>
<tr>
<td>1</td>
<td>0.366</td>
<td>0.367</td>
</tr>
<tr>
<td>2</td>
<td>0.161</td>
<td>0.171</td>
</tr>
<tr>
<td>3</td>
<td>0.061</td>
<td>0.053</td>
</tr>
<tr>
<td>4</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>7</td>
<td>0.002</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 4.2: The observed and empirical probabilities for the data in Table 4.1. The Poisson distribution fits the data very well.

If the hits are random, a Poisson distribution should fit the data. The data are given in Table 4.1.

Using the proportions as estimates of the true probabilities, the sample mean can be computed using the definition of the mean.

\[
E(X) \approx \sum x p_X(x) \\
= (0 \times 0.3976) + (1 \times 0.3663) + \cdots + (7 \times 0.0017) \\
= 0.9323
\]

Using this value as an estimate of the Poisson mean, the pf for the Poisson distribution can be compared to the empirical probabilities computed above; see Table 4.2. (For example, the probability of zero hits is \( \exp(-0.9323)(0.9323)^0/0! \approx 0.3936. \))

The two probabilities are very close; the Poisson distributions fits the data very well.
Poisson probabilities are calculated using
dpois(x=,lambda=) for \( P(X = x) \),
ppois(q=,lambda=) for \( P(X \leq q) \) and
ppois(q=,lambda=,lower.tail=F) for \( P(X > q) \)

In Rcmdr, \( P(X \leq q) \) is calculated using the menu for Poisson tail probabilities.

### 4.3.1 Relationship to the binomial distribution

If the number of trials is very large and the probability of success in a single trial is very small, then the computation of binomial probability is not easy. For example, consider \( X \sim \text{bin}(n = 2000, p = .005) \). Then, the pf is

\[
p_X(x) = \binom{2000}{x}(0.005)^x0.995^{2000-x},
\]

and it is then tedious to compute \( P(X > 101) \) for example. (Try computing \( \binom{2000}{102} \) on your calculator, for example.) The Poisson distribution can be used to approximate this probability however. We could set the Poisson mean to equal the binomial mean (that is, set \( \mu = np \)). The same can be done for the variance, setting \( \mu = np(1 - p) \). Since \( \text{E}(Y) = \text{var}(Y) = \mu \) for the Poisson distribution, this can only be (approximately) true here if \( p \) is close to zero.

A general guideline is that the Poisson distribution can be used to approximate the binomial when \( n \) is large, \( p \) is small and \( np \) is less than about 7.
Figure 4.5: The Poisson distribution is an excellent approximation to the binomial distribution when $p$ is small and $n$ is large. In this example, the binomial parameters are $n = 2000$ and $p = 0.005$ so that the Poisson mean is $\mu = 2000 \times 0.05 = 10$. The binomial pf is shown using the bars; the Poisson pf using the solid line.

### 4.4 Geometric distribution

Consider now an experiment where independent Bernoulli trials are repeated until the first success occurs. What is the distribution of the number of trials required?

Let the random variable $Y$ be the number of trials necessary to obtain the first success. Since the first success may occur on the first trial, or second trial or third trial, and so on, $Y$ is a random variable with range space $\{1, 2, 3, \ldots \}$.

The distribution is easy to derive. To observe the first success on the $y$th trial, there must be $y - 1$ failures followed by one success. Since the prob-
ability of failure is $q$ and the probability of success is $p$, the probability of the first success on trial $y$ is

$$y - 1 \text{ failures } q^{y-1} \times p \text{ One success.}$$

This derivation assumes the events are independent.

**Reading 4.5** Read WMS, Section 3.5.

**Definition 4.8** A rv $X$ has a geometric distribution if the pf of $X$ is given by

$$p_X(x) = q^{x-1}p \quad \text{for } x = 1, 2, \ldots$$

where $q = 1 - p$ and $0 < p < 1$ is the parameter of the distribution. We write $X \sim \text{geom}(p)$.

The pf for a geometric distribution for various values of $p$ is shown in Figure 4.6.

The following are the basic properties of the geometric distribution.

**Theorem 4.9** If $X \sim \text{geom}(p)$ then

1. $E(X) = 1/p$
2. $\text{var}(X) = (1 - p)/p^2$
3. $M_X(t) = pe^t/\{1 - (1 - p)e^t\}$

**Proof** Results 1 and 3 are proven directly in WMS, Section 3.5 but it’s easier to prove 3 first and then use the mgf to prove 1 and 2. This is left as an exercise.

**Example 4.7** A computer system fails at random with probability 0.02 on any given day. Determine

(a) the probability that the system will first fail before day 20.
(b) the number of days the system can be expected to work until failure.
4.4. Geometric distribution

Solution  Let $Y$ be the day on which the system fails, so that $Y = \{1, 2, 3, \ldots \}$. Then $Y \sim \text{geom}(0.02)$.

(a) We seek $P(Y < 20) = P(Y \leq 19)$. This can be done by hand, or using a computer. In R the command \texttt{pgeom()} gives values from the distribution function of the geometric distribution.

\begin{verbatim}
> pgeom(18,0.02) # ie up to 18 failures at p=0.02
[1] 0.3187674
\end{verbatim}

So the probability is 0.32 that the system will fail before day 20.

(b) The expected value is $E(Y) = 1/p = 1/0.02 = 50$. The system could be expected to last fifty days before the first failure.

Example 4.8  An article appearing in *The Sunday Mail* on 21 December 1997 give the birth weight, gender, and time of birth of 44 babies born
Table 4.3: Fitting the geometric distribution to baby birth data.

<table>
<thead>
<tr>
<th>Births Till Boy Born</th>
<th>Tally</th>
<th>Empirical Probability</th>
<th>Theoretical Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>0.692</td>
<td>0.500</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.115</td>
<td>0.250</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.154</td>
<td>0.125</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.000</td>
<td>0.063</td>
</tr>
<tr>
<td>5+</td>
<td>1</td>
<td>0.040</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>1.001</td>
<td>1.001</td>
</tr>
</tbody>
</table>

in the 24-hour period of 18 December 1997 at the Mater Mother’s Hospital in Brisbane, Australia.

The gender of the births is a random process, and so the number of births observed until a boy is born should have approximately a geometric distribution. By counting the number of births until a boy is born (and restarting the count when a girl is born), Table 4.3 has been constructed. The theoretical probabilities have been based on the probability of a boy being born as being 0.5. The fit is reasonable (but not brilliant) for the small sample size; see Table 4.3.

Geometric probabilities are calculated using
\[ dgeom(x= ,prob= ) \] for \( P(X = x) \)
\[ pgeom(q= ,prob= ) \] for \( P(X \leq q) \) and
\[ pgeom(q= ,prob= ,lower.tail=F) \] for \( P(X > q) \)

In Rcmdr, \( P(X \leq q) \) is calculated using the menu for Geometric tail probabilities.

### 4.5 Negative binomial distribution

Consider now an experiment where independent Bernoulli trials are repeated until the \( r \)th success occurs. What is the distribution of the number of trials required?

Let \( rv \ Y \) be the number of trials necessary to obtain \( r \) successes. Clearly \( Y \) will be at least \( r \) and the range space for \( Y \) is \( \{r, r + 1, r + 2, \ldots\} \). To
observe the $r$th success in the $y$th trial, there must have been $y - r$ failures and $r - 1$ successes in the first $y - 1$ trials; the $r$th success then happens on the $y$th trial. There are $\binom{y - 1}{r - 1}$ ways to allocate these successes to the first $y - 1$ trials. Each of the $r - 1$ successes occur with probability $p$, and the $y - r$ failures with probability $1 - p$ (assuming events are independent). Hence the probability of observing the $r$th success in trial $y$ is

$$\binom{y - 1}{r - 1} \times (1 - p)^{y - r} \times p^r.$$ 

**Definition 4.10**  A random variable $X$ with pf

$$p_X(x) = \binom{x - 1}{r - 1} (1 - p)^{x-r} p^r$$  \hspace{1cm} (4.3)

has a negative binomial distribution with parameters $r$ (an integer $\geq 1$) and $p$ ($0 \leq p \leq 1$). We write $X \sim \text{negbin}(r, p)$.

The probability function for the negative binomial distribution for various values of $p$ and $r$ is shown in Figure 4.7.

When $r = 1$, the negative binomial distribution is the same as the geometric distribution, so the geometric distribution is a special case of the negative binomial.

The following are the basic properties of the negative binomial distribution.

**Theorem 4.11**  If $X \sim \text{negbin}(r, p)$ with pf (4.3) then

1. $E(X) = r/p$
2. $\text{var}(X) = r(1 - p)/p^2$
3. $M_X(t) = \left[pe^t/(1 - (1 - p)e^t)^r \right]$
Figure 4.7: The pf for the negative binomial distribution for \( p = 0.2 \) and \( p = 0.7 \) and \( r = 1 \) and \( r = 3 \).

**Proof** Once again it makes sense to prove (iii) first and this is left as an exercise.

Negative binomial probabilities are calculated using
- `dnbinom(x= ,size= ,prob= )` for \( P(X = x) \)
- `pnbinom(q= ,size= ,prob= )` for \( P(X \leq q) \)
- `pnbinom(q= ,size= ,prob= ,lower.tail=F)` for \( P(X > q) \)

In Rcmdr, \( P(X \leq q) \) is calculated using the menu for Negative binomial tail probabilities.

**Example 4.9** A company employs a telephone marketer to invite customers, over the telephone, to a product demonstration. The marketer must obtain ten people for the demonstration. The probability that a randomly chosen person accepts the invitation is only 0.15.

(a) How many calls must the marketer make before obtaining ten acceptances?
(b) How likely is it that the marketer will need to make more than 100 calls to secure ten acceptances?

(c) Each call take an average of 5 minutes. How long is the marketer expected to be calling to find ten acceptances?

(d) Each call costs 25 cents. If the company pays the marketer $30 per hour, what is the average cost of securing ten acceptances?

Solution

(a) In this problem, a ‘success’ is an acceptance to attend the demonstration. Let $Y$ be the number of calls necessary to secure ten acceptances. Then $Y$ has a negative binomial distribution such that $Y \sim \text{NBin}(p = 0.15, r = 10)$. The mean number of calls to be made will be $E(Y) = r/p = 10/0.15 \approx 66.7$.

(b) To determine $P(Y > 100) = 1 - P(Y \leq 100)$, using a computer is the wisest approach (MATLAB, R, Excel, . . . ). In R, the command `dnbinom()` returns probabilities from the density function of the negative binomial distribution. One answer to the problem is

```r
> x.values <- seq(1, 100, by=1)
> 1 - sum(dnbinom( x.values, size=10, prob=0.15))
[1] 0.0244252
```

The probability is about 0.0244. Alternatively use `1 - pnbinom(100,10,0.15)` to access the probability directly from the distribution function.

(c) Let $T$ be the time to make the calls in minutes. Then $T = 5Y$. Hence, $E(T) = 5E(Y) = 5 \times 66.7 = 333.5$, or about 5.56 hours.

(d) Let $C$ be the total cost in dollars. The cost of employing the marketer is, on average, $30 \times 5.56 = $166.75. Then $C = 0.25Y + 166.75$, so that $E(C) = 0.25E(Y) + 166.75 = C = 0.25 \times 66.7 + 166.75 = $183.43.
4.5.1 Alternative parameterization

Often the negative binomial distribution is used in a different parameterization. This form can be seen as the number of failures which occur in a sequence of independent trials until \( r \) successes are reached. The pf is this reparameterized form in

\[
p_X(x) = \binom{x + r - 1}{x} p^r (1 - p)^x
\]

(4.4)

where \( X \) is the number of failures until the \( r \)th success is observed. Note that the variables \( X \) and \( Y \) are related by \( Y = X + r \) for constant \( r \). Hence, this pf is defined for \( x = 0, 1, 2, \ldots \).

It is also possible to allow \( r \) to be any positive number, not just an integer. This is true for the form above or for the parameterization in Equation (4.3). It is, however, most common to relax this restriction with the alternative parameterization of Equation (4.4).

When \( r \) is permitted to be non-integer, the interpretations given above are lost, but the distribution is then more flexible. Relaxing the restriction on \( r \) in the form of Equation (4.3) gives the pf as

\[
p_X(x) = \frac{\Gamma(x + r)}{\Gamma(r) x!} p^r (1 - p)^x,
\]

(4.5)

for \( x = 0, 1, 2, \ldots \) and \( r > 0 \). In this expression, \( \Gamma(r) \) is the gamma function (see below), and is like a factorial; indeed, \( \Gamma(r) = (r - 1)! \) if \( r \) is a positive integer. See Figure 4.8 and Section 4.5.2.

This form of the negative binomial distribution is sometimes used in place of the Poisson distribution used to model count data (they are both defined on \( \{0, 1, 2, \ldots\} \)). Since the negative binomial distribution has two parameters and the Poisson only one, the negative binomial distribution often produces a better fit.

For the reparameterised version of the negative binomial distribution Theorem 4.11 becomes:

If \( X \sim \text{negbin}(r, p) \) with pf (4.5) then

- \( \mathbb{E}(X) = r(1 - p)/p \)
- \( \text{var}(X) = r(1 - p)/p^2 \)
- \( M_X(t) = \left[p/(1 - (1 - p)e^t)\right]^r \)
4.5. Negative binomial distribution

4.5.2 The gamma function

**Definition 4.12** The function $\Gamma(\cdot)$ is called the gamma function and is defined as

$$
\Gamma(r) = \int_0^{\infty} x^{r-1} \exp(-x) \, dx
$$

for $r > 0$.

The gamma function has the special property that

$$
\Gamma(r) = (r - 1)!
$$

if $r$ is a positive integer (see Figure 4.8).

Here are important properties of the gamma function

**Theorem 4.13** For the gamma function $\Gamma(\cdot)$,

1. $\Gamma(r) = (r - 1)\Gamma(r - 1)$ if $r > 0$
2. $\Gamma(1/2) = \sqrt{\pi}$

**Proof**

1. Integration by parts gives

$$
\Gamma(r) = \int_0^{\infty} -x^{r-1} \, d(e^{-x})
$$

$$
= 0 + (r - 1) \int_0^{\infty} e^{-x} x^{r-2} \, dx
$$

$$
= (r - 1)\Gamma(r - 1)
$$

Property (4.6) follows directly from this result (Exercise!).

2. Putting $r = 1/2$ in the first part of this theorem gives

$$
\Gamma(1/2) = \int_0^{\infty} e^{-x} x^{-1/2} \, dx
$$

$$
= \sqrt{2} \int_0^{\infty} e^{-z^2/2} \, dz,
$$

putting $x = z^2/2$, $z \geq 0$

$$
= \sqrt{2\sqrt{\pi}} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz
$$

Note that the integral is $\frac{1}{2}$, being half the area under a normal curve.

So

$$
\Gamma(1/2) = \sqrt{\pi}.
$$
Figure 4.8: The gamma function is like the factorial function but has a continuous argument. The line corresponds to the gamma function $\Gamma(z)$; the solid points correspond to the factorial $(z-1)! = \Gamma(z)$ for integer $z$.

Example 4.10 Consider the computer system in Example 4.7. Suppose after five failures, the system is upgraded.

(a) What is the probability that an upgrade will happen within one year?

(b) What is the expected number of days between upgrades?

Solution Let $D$ be the days on which the system fails for the fifth time. Then $D \sim \text{NBin}(p = 0.02, r = 5)$. Note that $D = \{5, 6, 7, \ldots \}$.

(a) We seek $P(D < 365)$ or if using the alternate parameterization where $X = D - r$ we seek $P(X < 360)$. This is best done using a computer (using Excel, MATLAB, R or some other program). Using R:

```
> pnbinom(360,5,0.02)
[1] 0.8553145
```

So the probability of upgrading within one year is about 85%.
4.5. Negative binomial distribution

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>70</td>
<td>0.467</td>
<td>0.318</td>
<td>0.449</td>
</tr>
<tr>
<td>1</td>
<td>38</td>
<td>0.253</td>
<td>0.364</td>
<td>0.261</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>0.113</td>
<td>0.209</td>
<td>0.140</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>0.067</td>
<td>0.080</td>
<td>0.073</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>0.060</td>
<td>0.023</td>
<td>0.038</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.020</td>
<td>0.005</td>
<td>0.019</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.013</td>
<td>0.001</td>
<td>0.010</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.007</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td>8+</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.4: Counts of mites on leaves selected at random from six similar apple trees.

(b) $E(D) = 5/0.02 = 250$. About 250 days should elapse between upgrades, on average.

Example 4.11 Bliss [6] gives data from the counts of adult European red mites on leaves selected at random from six similar apple trees; see Table 4.4. From the data, the mean mites per leaf is

$E(X) = (0.467 \times 0) + (0.253 \times 1) + \cdots \approx 1.14667.$

To compute the variance,

$E(X^2) = (0.467 \times 0^2) + (0.253 \times 1^2) + \cdots \approx 3.57333,$

so that $var(X) = 3.57333 - (1.14667)^2 = 2.258489$. The Poisson distribution has an equal mean and variance; the Poisson distribution may not model the data well.

For the second parameterization of the negative binomial distribution, the expression for the mean and variance can be set to the computed values above and solved for $p$ and $r$; then $p \approx 0.5077$ and $r \approx 1.1826$. Using these values, the estimated probability function for both the Poisson and negative binomial distributions are given in Table 4.4; the negative binomial distribution fits better as expected.
4.6 Hypergeometric distribution

When the selection of items a fixed number of times is done with replacement, the probability of an item being selected stays the same and the binomial distribution can be used. However, when the selection of items is done without replacement, the trials are not independent, making the binomial model unsuitable. In such a situation is the hypergeometric model.

Consider first a simple example. A bag containing six red balls and four blue balls. The variable of interest, say $X$, is the number of red balls drawn in three random selections from the bag, without replacing the balls. Since the balls are not replaced, $P$ (draw a red ball) is not constant and so the binomial distribution cannot be used.

The probabilities can be computed, however, using probabilistic ideas from Module 1. There are a total of $\binom{10}{3}$ ways of selecting a sample of size 3 from the bag. Consider the case $X = 0$. The number of ways of drawing no red balls is $\binom{6}{0}$ and the number of ways of drawing the three blue balls in $\binom{4}{3}$, so the probability is

$$P(X = 0) = \frac{\binom{6}{0} \binom{4}{3}}{\binom{10}{3}} \approx 0.00833.$$ 

Likewise, the number of ways to draw one red ball (and hence two blue balls) is $\binom{6}{1} \times \binom{4}{2}$, so

$$P(X = 1) = \frac{\binom{6}{1} \binom{4}{2}}{\binom{10}{3}}.$$ 

Similarly,

$$P(X = 2) = \frac{\binom{6}{2} \binom{4}{1}}{\binom{10}{3}}$$

$$P(X = 3) = \frac{\binom{6}{3} \binom{4}{0}}{\binom{10}{3}}$$

In general, if there are $N$ balls in total in the bag, and $r$ of them are red, and we select a sample of size $n$ from the bag without replacement, then the probability of finding $x$ red balls in the sample of size $n$ is

$$P(X = x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}}.$$
where \( X \) is the number of red balls in sample of size \( n \). (In the example, \( n = 3, r = 6 \) and \( N = 10 \).) In the formula, note that \( \binom{r}{x} \) is the number of ways of selecting \( x \) red balls from the \( r \) red balls in the bag; \( \binom{N-r}{n-x} \) is the number of ways of selecting all the remaining \( n-x \) to be the other colour (and there are \( N-r \) of those in the bag); and \( \binom{N}{n} \) is the number of ways of selecting a sample of size \( n \) if there are \( N \) balls in the bag in total.

**Definition 4.14** Consider a set of \( N \) items of which \( r \) are of one kind (call them successes) and other \( N-r \) are of another kind (call them failures). We are interested in the probability of \( x \) successes in \( n \) trials, when the selection (or drawing) is made without replacement. Then the rv \( X \) is said to have a hypergeometric distribution with pf

\[
p_X(x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}} \tag{4.7}
\]

where \( x = 0, 1, \ldots, n; x \leq r, \) and \( n-x \leq N-r \).

The following are the basic properties of the hypergeometric distribution.

**Theorem 4.15** If \( X \) has a hypergeometric distribution with pf (4.7) then

1. \( E(X) = nr/N \)
2. \( \text{var}(X) = n \left( \frac{r}{N} \right) \left( \frac{N-r}{N} \right) \left( \frac{N-n}{N-1} \right) \)

The moment generating function is far too difficult\(^1\) and will not be considered here.

If the population is much larger than the sample size (that is, \( N \) is much larger than \( n \)), then the probability of a success will be approximately constant.

Consider the example at the start of this Section. The probability of drawing a red ball initially is \( 6/10 = 0.6 \). The probability that the next ball is red becomes \( 5/9 = 0.556 \). But suppose there are 10000 balls in the (very big!) bag, of which 6000 are red. The probability of drawing a red ball initially is \( 6000/10000 = 0.6 \) and the probability that the next ball is red becomes \( 5999/9999 = 0.59996 \); the probability is almost the same. In this case, we might consider using the binomial distribution with \( p \approx 0.6 \).

\(^1\)This probably comes as no surprise given the pf!
In general, if \( N \) is much larger than \( n \), the population proportion then will be approximately \( p \approx r/N \), and so \( 1 - p \approx (N - r)/N \). Using this information,

\[
E(X) = n \times (r/N) \approx np
\]

and

\[
\text{var}(X) = n \left( \frac{r}{N} \right) \left( \frac{N - r}{N} \right) \left( \frac{N - n}{N - 1} \right)
\approx n (p) (1 - p) (1)
= np(1 - p),
\]

which correspond to the mean and variance of the binomial distribution.

**Example 4.12** Twenty mice are available to be used in an experiment; seven of the mice are female and 13 are male. Five mice are required and will be sacrificed. What is the probability that more than three of the mice are males?

**Solution** Let \( Y \) be the number of male mice chosen in a sample of size 5. Then \( X \) has a hypergeometric distribution (since mice are chosen without replacement) where \( N = 20 \), \( n = 5 \), \( r = 13 \) and we seek

\[
P(X > 3) = P(X = 4) + P(X = 5)
\]

\[
= \binom{13}{4} \binom{7}{1} \binom{20}{5} + \binom{13}{5} \binom{7}{0} \binom{20}{5}
\]

\[
\approx 0.3228 + 0.0830 = 0.4058
\]

The probability is about 41%.

Hypergeometric probabilities are calculated using

- `dhyper(x, m, n, k)` for \( P(X = x) \)
- `phyper(q, m, n, k)` for \( P(X \leq q) \) and
- `phyper(q, m, n, k, lower.tail=F)` for \( P(X > q) \)

In Rcmdr, \( P(X \leq q) \) is calculated using the menu for **Hypergeometric tail probabilities**.
4.7 Self-assessment exercises

The following exercises are designed to provide practice at problem-solving based on the material in this module. Solutions are provided at the end of the module. Additional exercises are available in the next section and in the textbook.

Ex. 4.1 Suppose the number of typographical\(^2\) errors on a single page has a Poisson distribution with \(\mu = 2\). Assume errors are independent from page to page. What is the probability that two pages out of ten will contain no such errors?

Ex. 4.2 A university department has a photocopier security system to track the number of photocopies made by individual staff members. The staff members must punch in their access code before being allowed to make photocopies. The access code consists of any combination of the ten digits 0 to 9, and each staff member can choose any four digit number.

(a) Determine how many access codes are possible.
(b) Determine the probability that a randomly chosen access code contains more than two digits greater than 6.
(c) Comment on how likely it is that the access codes were all chosen at random.
(d) Determine the number of access codes possible if no digit can be repeated.

Ex. 4.3 In Australia, most number plates on cars, trucks and buses consist of any 3 letters and any 3 numbers. How many vehicles (cars, trucks and buses) can I expect to pass by me while waiting to cross the road before I see the letter A on a number plate?

Ex. 4.4 The number of phone calls received at a real estate firm per hour has approximately a Poisson distribution with mean 6. Determine

(a) the probability that more than ten phone calls will be received in one hour.
(b) Suppose the secretary takes a 30 minute lunch break. What is the probability that no calls will be be received in this time?

\(^2\)This one was intentional to keep you alert.
Ex. 4.5 Example 1.2 gives data on the number of times each face appears when a standard die was rolled. Do the data appear to come from a discrete uniform distribution? Justify your answer.

Ex. 4.6 (Computer exercise) In Section 1.9 an experiment concerning 21 CAO patients was described. If it is assumed there is no difference in effect of the two experimental conditions then each patient represents an independent Bernoulli trial with

\[ P(\text{success}) = P(\text{Patient does better under condition 1}) = 0.5. \]

Using R find \( P(X = x) \) (\( \text{dbinom} \)) and \( P(X \leq x) \) (\( \text{1-pbinom} \)) for \( x = 0, 1, 2, \ldots, 21 \) when \( X \sim \text{bin}(21, 0.5) \) and compare these probabilities to those estimated in Section 1.9.

(\text{Note: Simulation won’t generally give you the exact theoretical answer but, provided the sample size is not too small, should give a reasonable estimate of the true value.})

Ex. 4.7 (Computer exercise) Use the \( \text{ppois} \) function to calculate the probability that 5 or more customers will arrive in a half minute interval. (That is \( P(X \geq 5) \)).

Simulate the arrival of customers in 100 such half minute intervals and estimate \( P(X \geq 5) \). Compare the true and estimated results.

If five customers arrived in a half minute period is there sufficient evidence to claim that \( \mu \neq 2? \)

4.8 Exercises

Ex. 4.8 Assume that the probability that an observer will see a particular bird is \( p \). If there are 6 observers in a search party

(a) write down the probability distribution for \( X \), the number of observers that see the bird.

(b) What is the probability a bird will not be seen by the search party?

(c) How large should a search party be to ensure that there is at least a 95\% chance of the bird being seen?

Ex. 4.9 The probability of brill fish being infected by parasites is known to be 0.6. Suppose brill are sampled until an infected fish is found. Let \( Y \) be the number of fish sampled. Find the probability that

(a) the first infected fish is the fifth sampled.

(b) Find the probability that more than 3 fish must be sampled to find the first infected fish.
(c) If sampling continues after the first infected fish is found find the probability that the third infected fish is the eighth sampled.

**Ex. 4.10** Find the variance of the negative binomial distribution

\[ P(Y = y) = \binom{y - 1}{n - 1} p^n q^{y-n} \]

where \( y = n, n + 1, n + 2, \ldots \), and \( p + q = 1 \).

**Ex. 4.11** The change in depth, \( Y \) of a river at a specific location has pdf given by \( f(y) = \frac{1}{4}, -2 \leq y \leq 2 \).

(a) Find the probability that \( Y \) is greater than 1.

(b) If readings on four different days are taken what is the probability that exactly two are greater than 1?

**Ex. 4.12** (a) If more than three cars per minute enter a tunnel it is known that a hazardous situation will develop. It is known that on average 1 car enters the tunnel per 2 minute interval. Write down the probability function, \( P(X = x) \) for \( x = 0, 1, 2, \ldots \) for the number of cars that enter the tunnel in a one minute period and use the Poisson functions in R to evaluate it for \( x = 0, 1, 2, \ldots, 8 \) and graph the probability function.

(b) Find the probability that more than three cars will enter in a 1 minute period.

(c) The tunnel is observed for a total of ten (independent) 1 minute periods. Write down an expression for the probability that more than three cars enter the tunnel during at least one of the ten 1 minute periods. Evaluate this expression using R (pbinom).

**Ex. 4.13** When modelling counts \( Y \) defined on \( \{0, 1, 2, \ldots\} \), it is common to use the Poisson distribution, when \( E(Y) = \text{var}(Y) \) is implied. Show that the negative binomial distribution may be a better choice if \( \text{var}(Y) > E(Y) \).

**Ex. 4.14** List the assumptions made when using the following distributions:

(a) the Poisson distribution

(b) the binomial distribution

**Ex. 4.15** In Section 4.5, different forms of the negative binomial distribution were given; one for \( Y \), the number of trials until \( r \) successes were observed, and for \( X \), the number of failures until \( r \) successes were observed.
Module 4. Standard discrete distributions

<table>
<thead>
<tr>
<th>Year</th>
<th>No.</th>
<th>Year</th>
<th>No.</th>
<th>Year</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1959</td>
<td>3</td>
<td>1969</td>
<td>7</td>
<td>1979</td>
<td>3</td>
</tr>
<tr>
<td>1960</td>
<td>4</td>
<td>1970</td>
<td>4</td>
<td>1980</td>
<td>4</td>
</tr>
<tr>
<td>1961</td>
<td>5</td>
<td>1971</td>
<td>5</td>
<td>1981</td>
<td>3</td>
</tr>
<tr>
<td>1962</td>
<td>1</td>
<td>1972</td>
<td>6</td>
<td>1982</td>
<td>3</td>
</tr>
<tr>
<td>1963</td>
<td>3</td>
<td>1973</td>
<td>6</td>
<td>1983</td>
<td>8</td>
</tr>
<tr>
<td>1964</td>
<td>1</td>
<td>1974</td>
<td>6</td>
<td>1984</td>
<td>6</td>
</tr>
<tr>
<td>1965</td>
<td>5</td>
<td>1975</td>
<td>3</td>
<td>1985</td>
<td>7</td>
</tr>
<tr>
<td>1966</td>
<td>1</td>
<td>1976</td>
<td>7</td>
<td>1986</td>
<td>9</td>
</tr>
<tr>
<td>1967</td>
<td>2</td>
<td>1977</td>
<td>5</td>
<td>1987</td>
<td>6</td>
</tr>
<tr>
<td>1968</td>
<td>2</td>
<td>1978</td>
<td>8</td>
<td>1988</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4.5: The number of tornado’s recorded in New York state from 1959 to 1988.

(a) Use the fact that \( Y = X + r \) to derive the mean and variance of \( X \) given that \( \text{E}(Y) = r/p \) and \( \text{var}(Y) = r(1 - p)/p^2 \).

(b) Use the mgf for \( Y \) to prove the mgf for \( X \) is \( M_X(t) = \left( \frac{p}{1-(1-p)\exp(t)} \right)^r \).

Ex. 4.16 The number of tornadoes recorded in New York state from 1959 to 1988 are shown in Table 4.5 (from Wilks [34]).

(a) Tabulate the number of years in which there are zero, one, two, . . . tornado’s recorded.

(b) Use the data to compute the sample mean number of tornado’s recorded per year.

(c) Use the mean found above to compute the theoretical Poisson distribution of tornado’s per year recorded in New York state.

(d) Is the Poisson distribution a good model in this example? Explain.

Ex. 4.17 In a classic experiment, Rutherford and Geiger examined the radioactive decay of polonium (see Rutherford and Geiger [27]). They counted the number of scintillations in 72 second intervals of the element. There were a total of 10097 scintillations during 2608 intervals. The data is presented in Table 4.6.

(a) Convert the counts into probabilities.

(b) Determine the mean of the distribution.

(c) Using the mean calculated above, determine the probabilities under the assumption of a Poisson distribution.
### 4.8. Exercises

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>57</td>
<td>5</td>
<td>408</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>203</td>
<td>6</td>
<td>273</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>383</td>
<td>7</td>
<td>139</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>525</td>
<td>8</td>
<td>45</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>532</td>
<td>9</td>
<td>27</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.6: The number of scintillations in 72 second intervals of radioactive polonium.

<table>
<thead>
<tr>
<th>Month</th>
<th>Births</th>
<th>Month</th>
<th>Births</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>66</td>
<td>Jul</td>
<td>70</td>
</tr>
<tr>
<td>Feb</td>
<td>63</td>
<td>Aug</td>
<td>59</td>
</tr>
<tr>
<td>Mar</td>
<td>64</td>
<td>Sep</td>
<td>54</td>
</tr>
<tr>
<td>Apr</td>
<td>48</td>
<td>Oct</td>
<td>51</td>
</tr>
<tr>
<td>May</td>
<td>64</td>
<td>Nov</td>
<td>45</td>
</tr>
<tr>
<td>Jun</td>
<td>74</td>
<td>Dec</td>
<td>42</td>
</tr>
</tbody>
</table>

Table 4.7: The number of women in Basel, Switzerland having first births by month.

(d) By comparing theoretical Poisson probabilities and the probabilities calculated above, comment on the suitability of using the Poisson distribution to model radioactive decay.

**Ex. 4.18** Table 4.7 gives the number of women having their first births in the University Hospital of Basel, Switzerland, for each month of the year.

(a) Is it reasonable to expect that the distribution will be approximately uniform over the months? Explain.

(b) Do the data appear to follow a discrete uniform distribution? Explain.

(c) Assuming the discrete uniform distribution applies, determine $P(\text{woman has a first birth after August})$

(d) Assuming the discrete uniform distribution applies, determine $P(\text{woman has a first birth after August | woman has first birth after May})$

**Ex. 4.19** The gambling game of Keno has become popular recently. In this game, players pay $1 and then select numbers between 1 and 80 inclusive. (How many numbers they choose is discussed later.) Then,
a computer selects twenty numbers between 1 and 80 inclusive. Prizes are then awarded depending on how many selected numbers are in the computer’s sample of twenty. Players can select any quantity of numbers, from just one up to forty.

(a) Consider a player who picks \( n \) numbers. Show that the probability of the player successfully picking \( x \) numbers follows a hypergeometric distribution.

(b) Consider a player who decided to select ten numbers. Determine the probability of having these ten numbers selected in the computer’s sample of twenty.

(c) Determine the expected payoff (see Exercise 1.32) for the player in the last question. Successfully selecting ten numbers from ten wins a maximum of $2 000 000

(d) Determine the expected payoff for successfully selecting one number from one, five from five, and fifteen from fifteen also. The prizes are $3, $600 and $250 000 respectively. Which strategy has the greatest expected payoff?

(e) For players that select ten numbers, having seven, eight or nine numbers matching the computers selection wins $90, $900 and $9000 respectively. Which of these options has the greatest expected payoff?

(f) Write a short discussion on the statistical aspects of Keno, focusing on the chance of players winning money on the long term. Your answer should be no more than 6 sentences.

Ex. 4.20 Basinski [4, Exhibit 7.17, page 190] gives the mean number of daily admissions and daily discharges for heart attacks victims in medium-sized hospitals in Ontario, Canada; see Table 4.8. Do the discharges appear to be discrete uniform? What about the admissions? Can you offer any explanations?

4.9 Some answers and hints

4.1 Let the number of typographical errors be \( E \). Then \( E \sim \text{Pois}(2) \). Then,
\[
P (E = 0) = \exp(-2)2^0/0! = 0.1353.
\]
Now let \( P \) be the number of pages out of 10 with no errors. Then \( P \sim \text{bin}(n = 10, p = 0.1353) \) so
\[
P (P = 2) = \binom{10}{2} (0.1353)^2 (0.8647)^8 = 0.2575.
\]

4.2 (a) \( 10^4 \)

(b) binomial with the probability of choosing a number greater than 6 being 0.3 and \( n = 4 \), so answer is \( C_4^3 (0.7)^1 + C_4^4 (0.3)^4 (0.7)^0 = 0.0837 \)
4.9. Some answers and hints

<table>
<thead>
<tr>
<th>Day</th>
<th>Admission</th>
<th>Discharges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>1193</td>
<td>1142</td>
</tr>
<tr>
<td>Tuesday</td>
<td>1091</td>
<td>1278</td>
</tr>
<tr>
<td>Wednesday</td>
<td>1057</td>
<td>1321</td>
</tr>
<tr>
<td>Thursday</td>
<td>1193</td>
<td>1244</td>
</tr>
<tr>
<td>Friday</td>
<td>1142</td>
<td>1705</td>
</tr>
<tr>
<td>Saturday</td>
<td>1091</td>
<td>843</td>
</tr>
<tr>
<td>Sunday</td>
<td>1125</td>
<td>375</td>
</tr>
</tbody>
</table>

Table 4.8: Mean daily admission and discharges for heart attacks victims in medium-sized hospitals in Ontario, Canada.

(c) unlikely (people use number easy to remember, such as years of birth); order is important: \( P_{10}^{10} = 5040. 

4.3 Let \( A \) be the event ‘seeing a number plate with an A’. \( P(A) = 1 - P(A') = 1 - (25/26)^3 = 0.1110036; \) so the probability is geometric with \( p = 0.111 \) and the expected number is \( 1/p = 9. \)

4.4 Let \( X \) be the number of calls received in one hour; \( X \sim \text{Pois}(6). \) Then \( P(X > 10) = 1 - P(X \leq 10) = 0.0426. \) In half-an-hour, the mean is three phone calls. Let \( Y \) be the number of calls received in half-an-hour; \( Y \sim \text{Pois}(3). \) Then \( P(Y = 0) = 0.0498. \)

4.5 Yes—the variations are small enough to be explained by sampling error.

4.6 improve<-0:21
   Px<-dbinom(x=improve,size=21,prob=0.5)
   PGTx<-pbinom(q=improve,size=21,prob=0.5)
   PLTx<-1-PGTx
   results<-cbind(improve,Px,PLTx,PGTx)
   dimnames(results)<-list(NULL,c("x","P\{X=x\}","P(X<=x)","P(X>x)"))
   print(round(results,4))

4.7 ppois(q=5,mu=2)
   sim.results<-table(rpois(n=100,mu=2))
   print(sim.results)
   My results gave

   sim.results
   0 1 2 3 4 5 6
   14 24 27 21 8 4 2
There are 2 intervals where at least five customers enter giving an estimated probability of 0.06. We need to combine categories 5 and 6 so that class frequency is $> 5$ and large sample theory applies.

This would indicate that it is unlikely that five (or more) customers would arrive in a half minute interval if in fact $\mu = 2$. We would have reason to suspect that $\mu$ may be greater than 2 if five (or more) customers arrive in a half minute interval.

(Try it for 1000 intervals, your estimate should improve.)

4.8 See Section 4.2 (c) $n \geq \frac{\log(0.05)}{\log q}$

4.9 See Sections 4.4 and 4.5.
(a) 0.015 (b) 0.064 (c) 0.0464

4.10 First find $E[Y(Y + 1)] = E(Y^2) + E(Y)$ using a method similar to that for finding $E(Y)$, (Theorem 4.11. Then use $\text{var}(Y)E(Y^2) - [E(Y)^2]$. 

4.11 (a) 0.25 (b) 0.2109

4.12 See 4.3 with $\mu = 0.5$ since interested in intervals of 1 minute.
(a) Use $P(X > 3) = 1 - P(X \leq 3)$. Ans = 0.001752. (b) Use `ppois` subcommand. (c) Let $X =$ number of times more than three cars enter tunnel (ie number of ‘successes’ in 10 trials, $p = 0.001752$). Ans = 0.01736.

4.13 Make sure you use the correct parameterisation of the negative binomial!
Module 5

Standard continuous distributions

Contents

5.1 Continuous uniform distribution ................. 154
5.2 Normal distribution ............................ 157
   5.2.1 The standard normal distribution ............ 158
   5.2.2 Determining normal probabilities .......... 161
   5.2.3 Plotting the Normal distribution .......... 164
   5.2.4 Calculating normal probabilities (standard normal) ... 165
   5.2.5 Quantiles from probabilities (standard normal) ... 168
   5.2.6 Probabilities and quantiles from a normal distribution with \( \mu \neq 0 \) and \( \sigma \neq 1 \) ......... 171
   5.2.7 Normal approximation to the binomial .......... 174
5.3 Exponential distribution .......................... 178
   5.3.1 Relationship between the Poisson and exponential distributions .................. 181
5.4 Gamma distribution .............................. 183
5.5 Beta distribution ................................ 189
5.6 Self-assessment exercises .......................... 192
5.7 Exercises ......................................... 195
5.8 Some answers and hints ............................ 202
Module objectives

On completion of this module, students should be able to:

- be familiar with probability functions of uniform, exponential, gamma, beta, normal and bivariate normal random variables
- know the basic properties of the continuous distributions covered in this module
- be able to apply the various continuous distributions as appropriate to problem solving
- be able to approximate the binomial distribution by a normal distribution
- know the computation of conditional mean and variance for a bivariate normal distribution

In this module, some popular continuous distributions are discussed. Properties such as definitions and applications are considered.

5.1 Continuous uniform distribution

The continuous uniform distribution has a constant pdf over a given range.

**Reading 5.1** WMS, Section 4.4.

**Definition 5.1** If a rv $X$ with range space $[a, b]$ has the pdf

$$f_X(x) = \frac{1}{b-a} \quad \text{for } a \leq x \leq b$$

then $X$ has a continuous uniform distribution and we write $Y \sim U(a, b)$ or $X \sim \text{unif}(a, b)$.

The same notation is used to denote the discrete continuous uniform distribution; the context should make it clear which is meant). A plot of the pdf for a continuous uniform distribution is shown in Figure 5.1.

The following are the basic properties of the continuous uniform distribution.
5.1. Continuous uniform distribution

Theorem 5.2 If $X \sim \text{unif}(a, b)$ then

1. $E(X) = (a + b)/2$
2. $\text{var}(X) = (b - a)^2/12$
3. $M_X(t) = \{\exp(bt) - \exp(at)\}/(b - a)$

Proof These proofs are left as exercises.

Example 5.1 If $X$ is uniformly distributed on $[-2, 2]$, find $P\left(|X| > \frac{1}{2}\right)$.

$$P\left(|X| > \frac{1}{2}\right) = P\left(X > \frac{1}{2}\right) + P\left(X < -\frac{1}{2}\right)$$

$$= \int_{\frac{1}{2}}^{2} f(x) \, dx + \int_{-2}^{-\frac{1}{2}} f(x) \, dx \text{ where } f(x) = \frac{1}{4}$$

$$= \frac{3}{4}.$$ 

Example 5.2 Most calculators, mathematical and statistical packages, and many spreadsheets, generate uniform random numbers. Four different packages were used to generate 1000 random numbers between zero and one: MATLAB, R (an open source statistical package), OpenOffice.org (an open source office package for Windows and Linux) and
Excel (a Windows spreadsheet). Histograms of the numbers were plotted (see Figure 5.2). The means and variance of the 1000 numbers are given on the plots; the mean is expected to be \((0 + 1)/2 = 0.5\) and the variance \(1/12 = 0.08333\). In all cases, this simplistic test indicates that the random numbers appear sound.

Uniform probabilities are calculated using

- \(\text{punif}(q=, \text{size}=, p=)\) for \(P(X \leq q)\)

Although there is also a function to calculate the density (\(\text{dunif}()\)), it does not provide a probability because for continuous random variables, the probability mass at a single point on the real line is zero.
5.2 Normal distribution

The most readily known continuous distribution is probably the normal distribution, sometimes called the “bell-shaped” distribution. In many ways it is the cornerstone of modern statistical theory as you will see in this and follow-up courses. It was first investigated in the 18th century when scientists observed an astonishing degree of regularity in errors of measurement. They found that the patterns which they observed could be closely approximated by continuous curves which they referred to as ‘normal curves of errors’ and attributed to laws of chance. It has many applications and many natural quantities (such as heights and weights of humans) follow normal distributions.

**Definition 5.3** If a rv $X$ has the pdf

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

for $-\infty < x < \infty$, then $X$ has a normal distribution where the two parameters are

1. the mean $\mu$ such that $-\infty < \mu < \infty$
2. the standard deviation $\sigma$ such that $\sigma > 0$.

We write $X \sim N(\mu, \sigma^2)$.

Some texts use the notation $X \sim N(\mu, \sigma)$ so it is wise to check.

Some examples of normal distribution pdfs are shown in Figure 5.3.

In drawing the graph of the normal pdf note that

(i) $f_X(x)$ is symmetrical about $\mu$

(ii) $f_X(x) \to 0$ asymptotically as $x \to \pm\infty$

(iii) $f_X'(x) = 0$ when $x = \mu$ and a maximum occurs there

(iv) $f_X''(x) = 0$ when $x = \mu \pm \sigma$ (points of inflexion)
Figure 5.3: Some examples of normal distributions. The solid lines correspond to \( \sigma = 0.5 \) and the dashed lines to \( \sigma = 1.0 \). For the left diagram, \( \mu = -3 \); for the right diagram, \( \mu = 2 \).

The proof that \( \int_{-\infty}^{\infty} f_X(x) \, dx = 1 \) is not obvious and relies on first squaring the integral and then changing to polar coordinates.

The following are the basic properties of the normal distribution.

**Theorem 5.4** If \( X \sim N(\mu, \sigma^2) \) then

1. \( E(X) = \mu \)
2. \( \text{var}(X) = \sigma^2 \)
3. \( M_X(t) = \exp \left( \mu t + \frac{t^2 \sigma^2}{2} \right) \)

**Proof** The proof of these results is delayed until after the Theorem 5.6.

### 5.2.1 The standard normal distribution

A special case of the normal distribution is the **standard normal** distribution. This is simply the normal distribution with mean zero and variance one. The standard normal distribution has prominence in statistics because of its utility in many situations and it is often denoted as
5.2. Normal distribution

- \( \phi_Z(z) \) rather than the generic \( f_Z(z) \) for the density function, and
- \( \Phi_Z(z) \) rather than the generic \( F_Z(z) \) for the distribution function.

However, this convention is not universal so you must be alert to the occasions when the generic notation is being used in the case of the standard normal distribution. Alternatively, where you see \( \phi_Z(z) \) or \( \Phi_Z(z) \), it is safe to assume that they are describing the standard normal distribution.

**Definition 5.5** The pdf for a random variable \( Z \) with a standard normal distribution is

\[
\phi_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}
\]

where \(-\infty < z < \infty\). We write \( Z \sim N(0, 1) \).

It’s useful to note that

\[
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} \, dz = 1
\]

(5.1)

by virtue of the fact that \( f_Z(z) \) is a pdf. The proof of 5.1 is given in DGS, Section 5.6.

The following are the basic properties of the standard normal distribution.

**Theorem 5.6** If \( Z \sim N(0, 1) \) then

1. \( E(Z) = 0 \)
2. \( var(Z) = 1 \)
3. \( M_Z(t) = \exp\left(\frac{t^2}{2}\right) \)

**Proof** As usual we could prove part 3 and use that result to prove parts 1 and 2. However, it’s instructive to prove parts 1 and 2 directly.

1.

\[
E(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-\frac{1}{2}z^2} \, dz
\]

\[
= \int_{-\infty}^{\infty} -d(e^{-\frac{1}{2}z^2})
\]

\[
= \left[-e^{-\frac{1}{2}z^2}\right]_{-\infty}^{\infty}
\]

\[
= 0
\]
2. \[ \text{var}(Z) = \text{E}(Z^2) - \text{E}(Z)^2 = \text{E}(Z^2) \]
\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}z^2} \, dz \]
\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -zd(e^{-\frac{1}{2}z^2}) \]
\[ = \frac{1}{\sqrt{2\pi}} \left[ -z e^{-\frac{1}{2}z^2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} \, dz \]
\[ = 1 \]
since the first term is zero and the second term makes use of (5.1).

3. \[ M_Z(t) = \text{E}(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \, dz. \]
Collecting together the terms in the exponent and completing the square, we have
\[ -\frac{1}{2}[z^2 - 2tz] = -\frac{1}{2}((z-t))^2 + \frac{1}{2}t^2. \]
Taking the constants outside the integral, we have
\[ M_Z(t) = e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}|z-t|^2} \, dz. \]
The integral here is 1 being the area under an \( N(t, 1) \) pdf curve. Hence
\[ M_Z(t) = e^{\frac{1}{2}t^2} \]

This distribution is important in practice since any normal distribution can be rescaled into a standard normal distribution using
\[ Z = \frac{X - \mu}{\sigma}. \quad (5.2) \]

**Proof of Theorem 5.4** Theorem 5.6 makes this proof straightforward.

1. We have \( Z = (X - \mu)/\sigma \). Therefore
\[ X = \mu + \sigma Z. \]
it follows
\[ \text{E}(X) = \text{E}(\mu + \sigma Z) = \mu + \sigma \text{E}(Z) = \mu \]
because \( \text{E}(Z) = 0. \)
5.2. Normal distribution

2. Also
\[ \text{var}(X) = \text{var}(\mu + \sigma Z) = \sigma^2 \text{var}(Z) = \sigma^2 \]
because \( \text{var}(Z) = 1 \).

3. Finally
\[ M_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}(\exp(t(\mu + \sigma Z))) = \exp(\mu t)\mathbb{E}(\exp(t\sigma Z)) \]
but \( \mathbb{E}(\exp(t\sigma Z)) = M_Z(t\sigma) = e^{\frac{1}{2}(\sigma t)^2} \) so
\[ M_Z(t) = \exp \left( \frac{t^2}{2} \right) \]

5.2.2 Determining normal probabilities

The probability \( P(a < X \leq b) \) where \( X \sim N(\mu, \sigma^2) \) can be written
\[ P(a < X \leq b) = F_X(b) - F_X(a) \]
where the distribution function
\[ F_X(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du \]
This integral cannot be written in terms of standard functions and in general must be evaluated for a particular \( x \) numerically (e.g., using Simpson’s rule or the like).

(Note that the distribution function is described as \( F_X(x) \) because it is the distribution function for a normal distribution with mean \( \mu \) and standard deviation \( \sigma \). In general terms it also includes the special case where \( \mu = 0 \) and \( \sigma = 1 \), i.e., \( \Phi_Z(z) \).)

We don’t have to do this because all statistical packages have a built-in procedure that evaluates \( F_X(x) \) for any \( x \). In the pre-computing era, tables were used to find

- normal densities, \( \phi_Z(z) \), for given quantiles \( z \),
- normal tail probabilities \( P(Z < z) = \Phi_Z(z) \) for a quantile \( z \),

The tables were used in an inverse manner to find quantiles that correspond to a tail area probability.

The information in the tables were values for the distribution function of the standard normal distribution
\[ F_Z(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{u^2}{2} \right\} du \]
which had been evaluated numerically. Because any normal distribution can be transformed into a standard normal distribution using $Z = \frac{X - \mu}{\sigma}$, only one set of tables was required to cover the infinite possibilities associated with analysing normally distributed data. We have that $X = \mu + \sigma Z$ from which

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

The process of converting a value $x$ into $z$ using $z = \frac{x - \mu}{\sigma}$ is called standardising. Standardising remains a current technique in probability and statistics but tables may be consigned to the status of relic. It is more convenient to calculate probabilities using the computing and in our case the probability calculator is R.

The explanation of the process of calculating probabilities for the normal distribution is depicted in Figure 5.4. In this case the required distribution is the standard normal but it can be extended to the general case immediately.

Initially, the discussion is restricted to the ideas of using the standard normal and is then followed by explanation of how to draw the curves and how to calculate probabilities using Rcmdr.

The density curve in the top frame was produced with the R function `dnorm(q=y, mean=0, sd=1)` where the grid of $y$ values was about 100 points in the interval $(-4, 4)$. The location of the mean is shown by the thin grey line at zero and quantiles $-1.28$ and $1.28$ are chosen (arbitrarily) to discuss the calculation of probabilities. In future cases, you will supply different quantiles.

Consider the following cases.

(a) $P(Z < -1.28)$. This probability is the area under the curve to the left of the quantile $-1.28$. It is termed the lower tail probability associated with $q = -1.28$. That area or probability is determined from the distribution function where the probability associated with $q = -1.28$ is $p = 0.10$.

(b) $P(Z > 1.28)$. By symmetry, this is the same as $P(Z < -1.28)$. From the distribution function we observe that the probability associated with the quantile $1.28$ is $0.9$. Since the distribution function is $\Phi_z(z) = P(Z < z)$, then $P(Z > z) = 1 - \Phi_z(z)$ so $P(Z > 1.28) = 1 - P(Z < 1.28) = 1 - 0.9 = 0.1$. Because the inequality is $>$, we term this as an upper tail probability.
Figure 5.4: Density and distribution function for the standard normal distribution.

(c) $P(-1.28 < Z < 1.28)$. This interval on the X-axis (quantiles) has a probability given by the associated interval on the Y-axis (probabilities). So $P(-1.28 < Z < 1.28) = \Phi_Z(1.28) - \Phi_Z(-1.28) = 0.9 - 0.1 = 0.8$

When you come to calculate probabilities using R, it may be of use to sketch a rough diagram of the distribution function and mark the event such as $-1.28 < Z < 1.28$ on the X-axis. The probability of this event will be the associated interval on the Y-axis.
3 special cases

For the normal distribution,

1. The probability determined from the interval \((\mu - \sigma, \mu + \sigma)\) is 0.68. 
   Hence
   
   \[
   P(X > \mu + \sigma) = 0.16 \\
   P(X < \mu - \sigma) = 0.16 \\
   P(|X| > \mu + \sigma) = 0.32
   \]

2. The probability determined from the interval \((\mu - 2\sigma, \mu + 2\sigma)\) is 0.95. 
   Hence,
   
   \[
   P(X > \mu + 2\sigma) = 0.025 \\
   P(X < \mu - 2\sigma) = 0.025 \\
   P(|X| > \mu + 2\sigma) = 0.05
   \]

3. The probability determined from the interval \((\mu - 2.6\sigma, \mu + 2.6\sigma)\) is 0.99. 
   Hence,
   
   \[
   P(X > \mu + 2.6\sigma) = 0.005 \\
   P(X < \mu - 2.6\sigma) = 0.005 \\
   P(|X| > \mu + 2.6\sigma) = 0.01
   \]

The relationships to remember are

\[
P(\mu - \sigma < X < \mu + \sigma) = 0.68 \\
P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95 \\
P(\mu - 2.6\sigma < X < \mu + 2.6\sigma) = 0.99
\]
5.2. Normal distribution

5.2.3 Plotting the Normal distribution

It is usually not necessary to do this once you have understood the relationship between quantiles and probabilities.

In Rcmdr, select the menu for plotting the normal distribution as depicted in Figure 5.7. Future directions for this will be given as Distributions > Continuous distributions > Normal dist... > Plot ....

Figure 5.5: Rcmdr menu for plotting Normal distribution

This will produce a GUI (graphical user interface) where you supply the values to make the plot as in Figure 5.6.

Figure 5.6: Entering the values to produce the plot of the standard normal distribution function
In this case, the button for plotting the distribution function has been checked. The values for the mean and standard deviation have been left at the defaults of 0 and 1. Later when working with the general form of the normal distribution, you would insert the values for mean and standard deviation that apply in the specific case.

This is the graph produced

Figure 5.7: Plot of the distribution function for the standard normal.

5.2.4 Calculating normal probabilities (standard normal)

The Rcmdr menu for calculating normal probabilities is given by Distributions > Continuous dist ... > Normal dist... > Normal probabilities.

Figure 5.8: The Rcmdr menu for calculating normal probabilities

This brings up a GUI in which to enter the values.
5.2. Normal distribution

Figure 5.9: The GUI for entering values to calculate normal probabilities

![GUI for normal probabilities](image)

The values of $-1.28$ and $1.28$ are entered in the GUI with the option for lower tail checked. The values of the mean and standard deviation remain as the defaults.

The calculated probabilities are in the Output Window of Rcmdr,

![Calculated probabilities in Rcmdr Output Window](image)

Whilst the heavy calculations have been rendered straightforward by Rcmdr, there remains the job of understanding what the outputs represent. You would report these results as

$$ P(Z < -1.28) = 0.10 $$

$$ P(Z < 1.28) = 0.90 $$

Note rounding to 2 decimals.

For $P(Z > 1.28)$ (say) use the option for upper tail in the GUI.
Observe how Rcmdr generates the script for calculating these probabilities,

\[
\text{pnorm(c(-1.28,1.28), mean=0, sd=1, lower.tail=TRUE)}
\]

The function \texttt{pnorm} calculates probabilities for the normal distribution for a given mean and standard deviation and known quantiles. It can calculate both lower and upper tail probabilities.

An alternative to using Rcmdr is to simply type the script into the R console which is fine if you can remember the function \texttt{pnorm(q= ,mean= ,sd= ,lower.tail= )} and put in the arguments, e.g.

\[
> \text{pnorm(c(-1.28,1.28), mean=0, sd=1, lower.tail=TRUE)}
\]

\[
[1] 0.1002726 0.8997274
\]

If you required upper tail, change the argument to \texttt{lower.tail=F}.

**Default values.** The default values for \texttt{pnorm} are \texttt{mean=0}, \texttt{sd=1} and \texttt{lower.tail=T}. Unless you change these, the calculations will be done with these arguments.

There is no preference at this stage and it is recommended that you start off using Rcmdr and maybe switch to the scripting way as you become more proficient.

In the following examples, the script is given to show the calculation but the results may equally be obtained using Rcmdr. For the standard normal, the default values for the mean (0) and standard deviation (1) are used so it is not necessary to specify them.

**Example 5.3** If the rv \( Z \) is distributed \( N(0,1) \) then,

<table>
<thead>
<tr>
<th>Algebra</th>
<th>Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) ( P(Z \leq 1.3) = .90 )</td>
<td>( &gt; \text{pnorm(q=1.3)} )</td>
</tr>
<tr>
<td></td>
<td>( [1] 0.90 )</td>
</tr>
<tr>
<td>(b) ( P(Z &gt; 1.6) = 0.055 )</td>
<td>( &gt; \text{pnorm(q=1.6,lower.tail=F)} )</td>
</tr>
<tr>
<td></td>
<td>( [1] 0.055 )</td>
</tr>
<tr>
<td>(c) ( P(Z &lt; -1) = 0.16 )</td>
<td>( &gt; \text{pnorm(q=-1,lower.tail=T)} )</td>
</tr>
<tr>
<td></td>
<td>( [1] 0.16 )</td>
</tr>
<tr>
<td>(d) ( P(1 &lt; Z &lt; 1.3) = \Phi(1.3) - \Phi(1) = 0.9032 - 0.8413 = .062 )</td>
<td>( &gt; \text{diff(pnorm(q=c(1,1.3)))} )</td>
</tr>
<tr>
<td></td>
<td>( [1] 0.062 )</td>
</tr>
<tr>
<td>(e) ( P(-2 &lt; Z &lt; -1) = \Phi(-1) - \Phi(-2) = 0.9772 - 0.8413 = 0.14 )</td>
<td>( &gt; \text{diff(pnorm(q=c(-2,-1)))} )</td>
</tr>
<tr>
<td></td>
<td>( [1] 0.14 )</td>
</tr>
<tr>
<td>(f) ( P(-1 &lt; Z &lt; 1) = \Phi(1) - \Phi(-1) = 0.68 )</td>
<td>( &gt; \text{diff(pnorm(q=c(-1,1)))} )</td>
</tr>
<tr>
<td></td>
<td>( [1] 0.68 )</td>
</tr>
<tr>
<td>(g) ( P(-2 &lt; Z &lt; 2) = \Phi(2) - \Phi(-2) = 0.95 )</td>
<td>( &gt; \text{diff(pnorm(q=c(-2,2)))} )</td>
</tr>
<tr>
<td></td>
<td>( [1] 0.95 )</td>
</tr>
</tbody>
</table>
Examples (f) and (g) are 2 special cases worth remembering. For a normal distribution, the probability that an observation is within 1 standard deviation of the mean ($\mu \pm \sigma$) is 0.68 and $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$. Actually for (g), the quantile values should be $(-1.96, 1.96)$ rather than $(-2, 2)$ but it is easier to remember $\pm 2$ and the discrepancy due to this approximation is negligible.

![Diagram of standard normal distribution with area = 0.95 between -1.96 and 1.96](image)

Figure 5.11: Area under standard normal curve: $z = -1.96$ to $z = 1.96$

### 5.2.5 Quantiles from probabilities (standard normal)

To calculate quantiles from probabilities, the distribution function is used in reverse to the above explanation of finding probabilities from quantiles. This is depicted in Figure 5.12.

Define $z_\nu$ as the value of $Z$ such that $P(Z < z_\nu) = \nu$ \(^1\) so that $z_{0.1} = -1.25$, $z_{0.5} = 0$ and $z_{0.9} = 1.28$.

The Rcmdr menu to get quantiles is

**Distributions > Continuous dist .. > Normal dist .. > normal quantiles**

Enter the probabilities and change the default values for mean and standard deviation if necessary.

The Output Window contains the results.

The script generated by Rcmdr uses the function

```r
qnorm(p = , mean = , sd = , lower.tail = T).
```

\(^1\nu\) is the Greek letter pronounced “nu”
Figure 5.12: The process of determining quantiles corresponding to normal probabilities

\[ qnorm(p = \gamma, \text{mean}=0, \text{sd}=1) \]

Figure 5.13: The Rcmdr GUI to get normal quantiles for a given probability.

As in the previous section with \texttt{pnorm}, \texttt{qnorm} can be used directly in the R console.

\[ > \text{qnorm}(p = c(0.1, 0.9)) \]
\[ [1] -1.3 1.3 \]

\textbf{Example 5.4} If \( Z \) is distributed \( N(0, 1) \) find \( z \) so that
5.2. Normal distribution

Figure 5.14: Output Window for quantiles corresponding to given probabilities.

<table>
<thead>
<tr>
<th>Algebra</th>
<th>Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) $P(Z \leq z) = .8849 \Rightarrow z = 1.2$</td>
<td><code>qnorm(p=0.88)</code></td>
</tr>
<tr>
<td>(b) $P(Z &gt; z) = .3446 \Rightarrow z = 0.4$</td>
<td><code>qnorm(p=0.3446,lower.tail=F)</code></td>
</tr>
<tr>
<td>(c) $P(</td>
<td>Z</td>
</tr>
<tr>
<td>$P(-z &lt; Z &lt; z) = 0.950$</td>
<td>[1] 1.96</td>
</tr>
<tr>
<td>$P(Z &lt; z) = 0.975 \Rightarrow$</td>
<td>[1] 2.0</td>
</tr>
</tbody>
</table>

As this example refers to the standard normal, there was no need to enter mean and standard deviation because the default values are for the standard normal.

Figure 5.15 depicts the event (c). It is conventional to take the symmetric interval of probability such that $0.95 = 0.975 - 0.025$ rather than other alternatives like $0.99 - 0.04$ for example. The planning of this calculation could be aided by a quick diagram.
5.2.6 Probabilities and quantiles from a normal distribution with \( \mu \neq 0 \) and \( \sigma \neq 1 \)

For probabilities and quantiles for normal distributions other than the standard normal, it is a simple matter to enter the mean and standard deviation in lieu of the default values in the Rcmdr GUI or in the script.

Example 5.5  In (a)–(e) assume \( X \sim N(4, 9) \). That is \( \mu = 4, \sigma^2 = 9 \) ⇒ \( \sigma = 3 \).

<table>
<thead>
<tr>
<th>Algebra</th>
<th>Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) ( P(X &lt; 7) = 0.84 )</td>
<td>( \text{pnorm}(q=7, \text{mean}=4, \text{sd}=3) ) [1] 0.84</td>
</tr>
<tr>
<td>(b) ( P(X &gt; 2.5) = 0.69 )</td>
<td>( \text{pnorm}(q=2.5, \text{mean}=4, \text{sd}=3, \text{lower.tail}=F) ) [1] 0.69</td>
</tr>
<tr>
<td>(c) ( P(</td>
<td>X - 4</td>
</tr>
<tr>
<td>= ( P(-0.3 &lt; (X - 4) &lt; 0.3) )</td>
<td></td>
</tr>
<tr>
<td>= ( P(3.7 &lt; X &lt; 4.3) = 0.0796 )</td>
<td></td>
</tr>
<tr>
<td>(d) ( P(-0.5 &lt; X &lt; 4.3) )</td>
<td>( \text{diff}(\text{pnorm}(c(-0.5, 4.3), \text{mean}=4, \text{sd}=3)) ) [1] 0.47</td>
</tr>
<tr>
<td>= ( P(X &lt; 4.3) - P(X &lt; -0.5) )</td>
<td></td>
</tr>
<tr>
<td>= ( 0.54 - 0.067 = 0.473 )</td>
<td></td>
</tr>
<tr>
<td>(e) ( k ) such that ( P(</td>
<td>X - 4</td>
</tr>
<tr>
<td>⇒ ( P(</td>
<td>X - 4</td>
</tr>
<tr>
<td>⇒ ( P(-k + 4 &lt; X &lt; k + 4) = 0.95 )</td>
<td></td>
</tr>
<tr>
<td>⇒ ( -k + 4 = -1.88, k + 4 = 9.88 )</td>
<td></td>
</tr>
<tr>
<td>⇒ ( k = 5.88 )</td>
<td></td>
</tr>
</tbody>
</table>
5.2. Normal distribution

Example 5.6  Pearson and Lee [26] give the forearms lengths of 140 adult males (in inches). The mean of the lengths is 18.80 inches and the variance is 1.120 (inches$^2$). Fitting a normal distribution with these parameters produces an excellent fitting model; see Figure 5.16.

![Histogram of Forearm length](image)

Figure 5.16: The forearm lengths (in inches) of 140 adult males. A normal distribution models the data well.

(a) Using these computed values, what is the probability that a forearm in a male is longer than 20.5 inches?

(b) Of the men in the study, 15% have forearms shorter than what length?

Solution  Let $X$ be a length of a male forearm; then $X \sim N(18.80, 1.120)$.

(a) To compute the probability $P(X > 20.5)$,

```r
> pnorm(q=20.5, mean=18.8, sd=sqrt(1.120), lower.tail=F)
[1] 0.0541
```

so approximately 5.4% of males in the study have a forearm longer than 20.5 inches. (Note that the sample is not a random sample of all males; the results can only be generalized to the population—which is unknown in this case—from which the sample was taken.)

Many students find it useful to (hand) draw a picture to help them understand what to do. Sketch the distribution function and mark the mean and its probability of 0.5 using dotted lines. Do a quick mental calculation to determine how many standard deviations the quantile differs from the mean. Here, $\sigma = (1.12)^{\frac{1}{2}} \approx 1$ so that the quantile 20.5 is about 2 standard deviations larger than the mean. Draw in that value for $Y$ and its probability. (These
Module 5. Standard continuous distributions

are sketched by hand and are only approximate but it pays to draw neatly.) For this example, Figure 5.17 suggests what to draw.) Use the fact that $P(\mu - \sigma < X < \mu + \sigma) = 0.68$ and $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$. So the probabilities associated with each quantile in these relationships are:

<table>
<thead>
<tr>
<th>quantile</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu - 2\sigma$</td>
<td>0.025</td>
</tr>
<tr>
<td>$\mu - \sigma$</td>
<td>0.16</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.5</td>
</tr>
<tr>
<td>$\mu + \sigma$</td>
<td>0.84</td>
</tr>
<tr>
<td>$\mu + 2\sigma$</td>
<td>0.975</td>
</tr>
</tbody>
</table>

With relation to these standard quantiles, mark the quantile in question and its probability. Although this will be rough in a hand drawn sketch, it will be a useful basis for calculating the probability with R or Rcmdr. After some practice, the drawing of the function will become superfluous.

Figure 5.17: Sketch of the distribution function with standard quantiles and the quantile of interest. standard quantiles are drawn as dotted lines and the statistics of interest are represented as the grey lines.

(b) Require $\nu$ such that $P(X < \nu) = 0.15$

Figure 5.17 is again the basis for determining (in the early stages of understanding probability) how to calculate the quantile from the probability. The given probability is 0.15 which is quite near to that for $P(X < \mu - \sigma) = 0.16$. Hence we would expect that $X_{0.15}$ is close to (but not exactly equal to) $18.8 - (1.12)^{\frac{1}{2}} = 17.7$. It is readily calculate in R (or Rcmdr)
5.2. Normal distribution

```r
> qnorm(p=0.15,mean=18.8,sd=sqrt(1.12))
[1] 17.7
```

Hence \( X_{0.15} = 17.703 \approx 17.7 \); approximately 15\% of men in the study have forearms shorter than about 17.7 inches. By comparison \( X_{0.16} = 17.748 \).

---

**Example 5.7** An examination has mean score of 500 and standard deviation 100. The top 75\% of candidates taking this examination are to be passed. Assuming the score has a normal distribution what is the lowest passing score?

```r
> qnorm(p=0.75,mean=500,sd=100,lower.tail=F)
[1] 432.55
```

Remember it is not sufficient to present your answer as the computer output. The answer is given in a sentence or 2 and in the context of the problem. An appropriate answer in this cases would be

The quantile for the upper 0.75 of candidates is 432.6 so to enable a pass rate of 0.75, the pass mark is set at 432.6.

---

**Quick approximations using the standard normal**

Since computing replaced tables, the role of the standard normal is (a) for theoretical purposes, (b) as a pivotal statistic (this follows in STAT261), comparing different normal distributions using the Z-score and (d) quick approximations when scanning results.

Given a value of a normal random variable, \( x \) and the mean \( \mu \) and standard deviation \( \sigma \), it is quite easy to calculate \( z = (x - \mu)/\sigma \).

In ??, the probabilities of a normal random variable exceeding 1, 2 or 2.6 standard deviations from the mean was written.

Recall,

\[
\begin{align*}
P(X > \mu + \sigma) &= 0.16 \\
P(X > \mu + 2\sigma) &= 0.05 \\
P(X > \mu + 2.6\sigma) &= 0.01
\end{align*}
\]
In the case of the standard normal $(\mu = 0, \sigma = 1)$ these rules become

\[
\begin{align*}
P(Z > 1) &= P(Z < -1) = 0.16 \\
P(Z > 2) &= P(Z < -2) = 0.05 \\
P(Z > 2.6) &= P(Z < -2.6) = 0.01
\end{align*}
\]

Thus we can quickly gauge the approximate probability by inspecting $Z$. If for instance $Z = 2.2$, we readily conclude that $P(Z > 2.2) < 0.05$.

**Example**

If $X$ has a normal distribution with mean 2 and variance 16, find (approximately) $P(|X - 2| \geq 8)$.

\[
Z = \frac{X - 2}{\sqrt{4}}
\]

\[
P(|X - 2| \geq 8) = P\left(\frac{|X - 2|}{4} \geq \frac{8}{4}\right)
\]

\[
= P(|Z| \geq 2)
\]

\[
= 0.05
\]

### 5.2.7 Normal approximation to the binomial

In Section 4.2, the binomial distribution was considered. There are times when it might be difficult to use the binomial distribution; consider the case of a binomial random variable $X$ where $n = 1000$, $p = 0.45$ and we seek $P(X > 997)$.

However, sometimes the normal distribution can be used to approximate binomial probabilities. This is possible since, for certain parameter values, the binomial pf starts to take on a normal distribution shape; see Figures 4.3 and 5.18.
5.2. Normal distribution

When is the binomial probability function close enough to use the normal approximation? There is no definitive answer; a common guideline suggests that if both

- \( np \geq 5 \) and
- \( n(1-p) \geq 5 \),

the approximation is satisfactory. (These are only guidelines, and other texts may suggest different guidelines.)

Figure 5.18 shows some picture of various binomial pdfs overlaid with the corresponding normal distribution; the approximation is visibly better as the guidelines given above are satisfied.

![Figure 5.18: The normal distribution approximating a binomial distribution.](image)

The guidelines suggest the approximation should be good when \( np \geq 5 \) and \( n(1-p) \geq 5 \); this is evident from the pictures. In the top row, a significant amount of the approximating normal distribution even appears when \( X < 0 \).

The normal distribution can be used to approximate probabilities in situations that are actually binomial. There is a fundamental difficulty with this
approach: modelling a discrete distribution with a continuous distribution. This is best explained through an example. The example explains the principle; the idea extends to all situations where the normal distribution is used to approximate a binomial distribution.

Consider \( mdx \) mice (which have a strain of muscular dystrophy) from a particular source for which 30% of the mice survive for at least 40 weeks. One particular experiment requires at least 35 mice to live beyond 40 weeks, and 100 mice have been sourced from this supplier. What is the probability that 35 or more of the group will survive beyond 40 weeks?

First note that the situation is binomial; if \( X \) is the number of mice from the group of 100 that survive, then \( X \sim \text{bin}(100, 0.3) \), \( E(X) = 100 \times 0.3 = 30 \) and \( \text{var}(X) = 100 \times 0.3 \times 0.7 = 21 \)

This could be approximated by the normal distribution \( Y \sim N(30, 21) \).

Since \( np = 30 \) and \( n(1 - p) = 70 \) are both much larger than 5, this approximation is expected to be quite good. Figure 5.19 shows the upper tail of the distribution near \( X = 35 \). Note that if we use the normal approximation from only \( Y = 35 \), only half of the original bar in the binomial pf is included; but since the number of mice is discrete, we want the entire bar corresponding to \( X = 35 \). So to compute the correct answer, we need to use the normal distribution to find \( P(Y > 34.5) \). This change from the \( X \geq 34.5 \) to \( Y > 34.5 \) is called using the continuity correction.

The exact answer (using the binomial distribution) is 0.1629 (rounded to four decimal places). Note when calculating the binomial probability we require \( P(X \geq 35) \equiv P(X > 34) \).

\begin{verbatim}
> pbinom(q=34,p=0.3,size=100,lower.tail=F)
[1] 0.16286
\end{verbatim}

Using the normal distribution with the continuity correction gives the answer as 0.1631;

\begin{verbatim}
> pnorm(q=34.5,mean=30,sd=sqrt(21),lower.tail=F)
[1] 0.16305
\end{verbatim}

Using the normal distribution without the continuity correction, the answer is 0.1376.

\begin{verbatim}
> pnorm(q=35,mean=30,sd=sqrt(21),lower.tail=F)
[1] 0.13762
\end{verbatim}

The solution is more accurate, as expected, using the continuity correction.
5.2. Normal distribution

Without using correction

With correction

Figure 5.19: The normal distribution approximating a binomial distribution when \( n = 100 \) and \( p = 0.3 \) to compute \( P(X \geq 35) \). The distribution is only shown for \( 33 < X < 37 \). Without the continuity correction (left), only half the \( X = 35 \) bar is counted; using the continuity correction, the entire \( X = 35 \) bar is included in the computation.

Example 5.8 Consider rolling a standard die 100 times, and counting the number of 1’s that appear uppermost. The random variable \( X \) is the number of 1’s; then, \( X \sim \text{bin}(n = 100, p = 1/6) \). Since \( np = 16.667 \) and \( n(1-p) = 83.333 \) are both greater than 5, a normal approximation should be accurate, so define \( Y \sim N(16.6667, 13.889) \) (the variance is \( np(1 - p) = 13.889 \)). Various probabilities have been computed as shown in Table 5.1 that indicate the accuracy of the approximation, and the way in which the continuity correction has been used. You should ensure you understand how the concept of the continuity correction has been applied in each situation, and be able to compute the probabilities for the normal approximation.
### Binomial Normal Approximation

<table>
<thead>
<tr>
<th>Event</th>
<th>Prob</th>
<th>Event</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X &lt; 10)$</td>
<td>0.02129</td>
<td>$P(Y &lt; 9.5)$</td>
<td>0.02724</td>
</tr>
<tr>
<td>$P(X \leq 15)$</td>
<td>0.3877</td>
<td>$P(Y &lt; 15.5)$</td>
<td>0.3771</td>
</tr>
<tr>
<td>$P(X &gt; 17)$</td>
<td>0.4006</td>
<td>$P(X &gt; 17.5)$</td>
<td>0.4115</td>
</tr>
<tr>
<td>$P(X \geq 21)$</td>
<td>0.1519</td>
<td>$P(X &gt; 20.5)$</td>
<td>0.1518</td>
</tr>
<tr>
<td>$P(X = 16)$</td>
<td>0.1065</td>
<td>$P(15.5 &lt; Y &lt; 16.5)$</td>
<td>0.1050</td>
</tr>
<tr>
<td>$P(15 &lt; X \leq 17)$</td>
<td>0.2117</td>
<td>$P(15.5 &lt; Y &lt; 17.5)$</td>
<td>0.2113</td>
</tr>
</tbody>
</table>

Table 5.1: Some events and their probabilities, computed using the binomial distribution (exact) and the normal approximation using the continuity correction. The accuracy of the approximation is very good.

## 5.3 Exponential distribution

An important distribution intimately connected with the Poisson distribution is the exponential distribution. If the number of events in a process follows a Poisson distribution the space or time between two consecutive events has an exponential distribution. (See Theorem 5.9.) Hence the exponential distribution is used to describe the interval between consecutive randomly occurring events.

**Definition 5.7** If a rv $X$ has the pdf, $f_X(x) = \frac{1}{\beta} \exp(-x/\beta)$, then $X$ has an exponential distribution with parameter $\beta$ ($>0$). We write $X \sim \text{exp}(\beta)$.

The parameter $\lambda$ or $\theta = 1/\beta$ is often used in place of $\beta$.

Plots of the pdf for various exponential distributions are given in Figure 5.20.

The following are the basic properties of the exponential distribution.

**Theorem 5.8** If $X \sim \text{exp}(\beta)$ then

1. $E(X) = \beta$
2. $\text{var}(X) = \beta^2$
3. $M_X(t) = (1 - \beta t)^{-1}$ for $t < 1/\beta$. 


5.3. Exponential distribution

![Graph](image)

Figure 5.20: The pdf of an exponential distribution for various values of $\beta$.

**Proof**  These proofs are left as an exercise.

Notice that the parameter $\beta$ represents the mean of the exponential distribution or, in the context of a Poisson process, the mean interval length between consecutive events. Then the alternative parameter $\lambda = 1/\beta$ represents the mean rate at which events occur.

**Example 5.9**  This example reveals the relationships amongst the parameters typically used to describe the exponential and Poisson distributions.

A Poisson process occurs at the mean rate of 5 events per hour. Describe the distribution of the time between consecutive events and the distribution of the number of events in one day (24 hours).

Let $N$ represent the number of events in one day and $T$ the time between consecutive events. We are given that events occur at the mean rate of $\lambda = 5$ events per hour. It follows that the mean time between consecutive events, $\beta = 1/\lambda = 0.2$ hours. Also, the mean number of events in one day is $\mu = 24 \times 5 = 120$. Consequently we have that $N \sim \text{Pois}(\mu = 120)$ and $X \sim \text{exp}(\beta = 0.2)$ (or, equivalently, $X \sim \text{exp}(\lambda = 5)$). (These parameter definitions are strongly recommended although not all texts follow them.)
Example 5.10 Cox and Lewis [8] give data collected by Fatt and Katz concerning the time intervals between successive nerve pulses along a nerve fibre. There are 799 observations which we do not give here. The mean time between pulses is $\beta = 0.2186$ seconds. An exponential distribution might be expected to model the data well. This is indeed the case; see Figure 5.21. What proportion of time intervals can be expected to be longer than 1 second?

![Figure 5.21: The time between successive nerve pulses. An exponential distribution with mean $\beta = 0.2186$ models the data well.](image)

**Solution** If $X$ is the time between successive nerve pulses (in seconds), then $X \sim \text{exp}(\beta = 0.2186)$.

The solution will then be

$$P(X > 1) = \int_1^{\infty} \frac{1}{0.2186} \exp(-x/0.2186) \, dx$$

$$= - \exp(-x/0.2186) \bigg|_1^{\infty}$$

$$= (-0) + (\exp\{-1/0.2186\})$$

$$= 0.01031.$$

There is about a 1% chance of a nerve pulse exceeding one second.

If the mean is 0.2186, the rate is $\frac{1}{0.2186}$ and that is the argument which is used in R.

```r
> pexp(q=1,rate=1/0.2186,lower.tail=F)
[1] 0.010311
```
5.3. Exponential distribution

5.3.1 Relationship between the Poisson and exponential distributions

Events occurring according to a Poisson process are characterised by an exponential distribution describing the interval between consecutive events. This relationship is captured in the following theorem.

**Theorem 5.9** Consider a Poisson process at rate \( \lambda \) and suppose observation starts at an arbitrary time point. Then the time \( T \) to the first event has an exponential distribution with mean \( \mu = 1/\lambda \); i.e.

\[
f_T(t) = \lambda e^{-\lambda t}, \quad t > 0
\]

**Proof** Let \( t = 0 \) be the arbitrary time at which observation of the Poisson process starts. Consider an interval \( (0, t] \) for some fixed value \( t > 0 \). Now \( T \) is the elapsed time after observations starts until the first event occurs.

Clearly, if the first event takes longer than \( t \) to occur, then \( T > t \) and the number of events in \( (0, t] \) is zero. Conversely, if the first event occurs at time \( t \) or earlier then \( T \leq t \) and the number of events in \( (0, t] \) is greater than zero. It follows that the events \( \{ T > t \} \) and \( \{ N(t) = 0 \} \) are equivalent where \( N(t) \) is the number of events occurring in time \( (0, t] \). Therefore

\[
P(T > t) = P(N(t) = 0)
\]

But \( N(t) \sim \text{Pois}(\lambda t) \) with mean \( \mu = \lambda t \). Hence

\[
P(T > t) = \frac{(\lambda t)^{0}e^{-\lambda t}}{0!} = e^{-\lambda t}
\]

This shows that the df of \( T \) is given by

\[
F_T(t) = P(T \leq t) = 1 - P(T > t) = 1 - e^{-\lambda t}
\]

Differentiating with respect to \( t \) yields the pdf

\[
f_T(t) = \frac{d}{dt}F_T(t) = \lambda e^{-\lambda t}
\]

which we recognise as the exponential distribution with mean \( \mu = 1/\lambda \).

Although the theorem refers to ‘time’, the variable of interest may be distance or any other continuous variable which may be applicable in measuring the interval between events.

An important feature of a Poisson process and hence of the exponential distribution is the memoryless or Markov property; that is, the future of the process at any time point does not depend on the history of the process. This property is captured in the following theorem.
**Theorem 5.10** If $T \sim \text{exp}(\lambda)$, then for $s > 0$ and $t > 0,$

$$P(T > s + t \mid T > s) = P(T > t)$$

**Proof** Using Definition 1.14,

$$P(T > s + t \mid T > s) = \frac{P(\{T > s + t\} \cap \{T > s\})}{P(T > s)}$$

But if $T > s + t$ then $T > s$. Consequently $P(\{T > s + t\} \cap \{T > s\}) = P(T > s + t)$ and so

$$P(T > s + t \mid T > s) = \frac{P(T > s + t)}{P(T > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T > t)$$

where we have used Equation (5.3).

This theorem states that the probability that the time to the next event is greater than $t$ does not depend on the time $s$ back to the previous event.

**Example 5.11** Suppose the lifespan of component $A$ is modelled by an exponential distribution with mean 12 months (rate $= \frac{1}{12}$ or 1 component for every 12 months).

(a) What is the probability that component $A$ fails in less than 6 months?

(b) Component $A$ has been in place for 12 months. What is the probability that it will fail in less than a further 6 months?

**Solution** By the memoryless property, the answers to (a) and (b) are the same and are both given by $P(T < 6)$ where $T \sim \text{exp}(\beta = 6)$; i.e

$$P(T < 6) = 1 - \exp(-6/12) = 0.3935$$

```r
> P1 <- pexp(q=6, rate=1/12, lower.tail=T)
> P1
[1] 0.39347
```
\[ P(T < 18 | T > 12) = \frac{P(12 < T < 18)}{P(T > 12)} = \frac{[1 - \exp\left(\frac{-18}{12}\right)] - [1 - \exp\left(\frac{-12}{12}\right)]}{\exp\left(\frac{-12}{12}\right)} = 0.3935 \]

\begin{verbatim}
P2 <- diff( pexp(c(12,18),rate=1/12) ) # P(12 < T < 18) P3 <- pexp(12,rate=1/12,lower.tail=F) # P(T > 12)
P2/P3
[1] 0.3934693
\end{verbatim}

Example 5.11 highlights the notion that an exponential process is ageless in the sense that the risk of ‘mortality’ remains constant with age. In other words, the probability of such an event occurring in the next small interval, whether it be the failure of a component or the occurrence of an accident, remains constant regardless of the age of the component or the length of time since the last accident. In this sense an exponential lifetime is different from a human lifetime or the lifetime of many man-made objects where the risk of ‘death’ in the next small interval increases with age.

## 5.4 Gamma distribution

**Reading 5.4** DGS, Section 5.9; WMS, Section 4.6.

The normal distribution is often used in modelling, but there is one significant shortcoming. A lot of data is only defined for non-negative values or for positive values\(^2\) yet the normal distribution allows negative values of the random variable. This is not always a problem, especially when the observations are far from zero; see Example 5.6. But some data sets have observations close to zero, and the data are often skewed to the right (positively skewed).

There are many distributions for modelling right skewed data; here the gamma distribution is considered.

**Reading 5.5** Read WMS, Section 4.6; DGS, Section 5.9.

\(^2\)The phrase non-negative permits zero; positive explicitly excludes zero.
Definition 5.11 If a rv $X$ has the pdf
\[ f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha - 1} \exp(-x/\beta) \]
then $X$ has a gamma distribution, where $\Gamma(\cdot)$ is the gamma function (see Section 4.5.2) and $\alpha, \beta > 0$. We write $X \sim \text{Gamma}(\alpha, \beta)$.

The parameter $\alpha$ is called the shape parameter and $\beta$ is called the scale parameter. Some texts use different notation for the shape and scale parameters. In broad terms, the shape parameter dictates the general shape of the distribution; the scale parameter dictates how ‘stretched out’ the distribution is.

Plots of the gamma pdf for various values of the parameters are given in Figure 5.22.

The exponential distribution is a special case of the gamma distribution with $\alpha = 1$. This means that properties of the exponential distribution can be obtained by substituting $\alpha = 1$ into the formulae for the gamma distribution.

Notice that

\[
\int_0^\infty f_X(x) \, dx = \int_0^\infty \frac{e^{-x/\beta} x^{\alpha - 1}}{\beta^\alpha \Gamma(\alpha)} \, dx
\]

\[
= \frac{1}{\Gamma(\alpha)} \int_0^\infty e^{-y} y^{\alpha - 1} \, dy \quad \text{(on putting } y = x/\beta) \\
= 1 \quad \text{(because } \int_0^\infty e^{-y} y^{\alpha - 1} \, dy = \Gamma(\alpha))
\]

as it must.
The following are the basic properties of the gamma distribution.

**Theorem 5.12** If $X \sim \text{Gamma}(\alpha, \beta)$ then

1. $E(X) = \alpha \beta$
2. $\text{var}(X) = \alpha \beta^2$
3. $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < 1/\beta$.

**Proof**

1. 

$$E(X) = \int_0^{\infty} xf_X(x) \, dx = \beta \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)} \int_0^{\infty} \frac{e^{-x/\beta} x^{(\alpha+1)-1} \, dx}{\beta^{\alpha+1} \Gamma(\alpha + 1)} = \alpha \beta.$$ 

This result follows from using (5.4) and Theorem 4.13.
Module 5. Standard continuous distributions

2. 
\[ E(X^2) = \int_0^\infty x^2 f_X(x) \, dx = \beta^2 \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-x/\beta} x^{\alpha+1}}{\beta^{\alpha+2} \Gamma(\alpha + 2)} \, dx = \alpha(\alpha+1)\beta^2. \]

where the result follows by writing \( \Gamma(\alpha + 2) = (\alpha + 1)\alpha\Gamma(\alpha) \). Hence
\[ \text{var}(X) = E(X^2) - [E(X)]^2 = \alpha(\alpha + 1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2 \]

3. 
\[ M_X(t) = E(e^{Xt}) = \int_0^\infty e^{tx} e^{-x/\beta} x^{\alpha-1} \frac{\beta^\alpha \Gamma(\alpha)}{\Gamma(\alpha)(1 - \beta t)^{\alpha-1} 1/\beta^\alpha} \, dx \]
\[ = \int_0^\infty e^{-z} z^{\alpha-1} \frac{dz}{\Gamma(\alpha)(1 - \beta t)^{\alpha-1} 1/\beta^\alpha}, \text{ putting } z = x(1 - \beta t)/\beta \]
\[ = (1 - \beta t)^{-\alpha}, \text{ since the integral remaining is 1.} \]

\[ \star \]

As usual the moments can be found by expanding \( M_X(t) \) in series. That is,
\[ M_X(t) = 1 + \alpha\beta t + \frac{\alpha(\alpha + 1)\beta^2}{2!} t^2 + \cdots \]

from which
\[ E(X) = \text{coefficient of } t = \alpha\beta, \]
\[ E(Y^2) = \text{coefficient of } t^2/2! = \alpha(\alpha + 1)\beta^2, \text{ as found earlier.} \]

Like for the normal distribution, the distribution function of the gamma cannot, in general, be computed without using numerical integration although see Example 5.13).

Example 5.12 Larsen and Marx [19, Case Study 4.6.1] use the gamma distribution to model daily rainfall in Sydney, Australia using the parameter estimates \( \alpha = 0.105 \) and \( \beta = 76.9 \). (Their example is based on Das [9].) The comparison between the data and the model is given in Table 5.2 indicating a good correspondence between the data and the theoretical distribution.
### 5.4. Gamma distribution

#### Table 5.2: The gamma distribution used to model Sydney daily rainfall.

<table>
<thead>
<tr>
<th>Rainfall (in mm)</th>
<th>Observed Frequency</th>
<th>Expected Frequency</th>
<th>Rainfall (in mm)</th>
<th>Observed Frequency</th>
<th>Expected Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–5</td>
<td>1631</td>
<td>1639</td>
<td>46–50</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>6–10</td>
<td>115</td>
<td>106</td>
<td>51–60</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>11–15</td>
<td>67</td>
<td>62</td>
<td>61–70</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>16–20</td>
<td>42</td>
<td>44</td>
<td>71–80</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>21–25</td>
<td>27</td>
<td>32</td>
<td>81–90</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>26–30</td>
<td>26</td>
<td>26</td>
<td>91–100</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>31–35</td>
<td>19</td>
<td>21</td>
<td>101–125</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>36–40</td>
<td>14</td>
<td>17</td>
<td>126–150</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>41–45</td>
<td>12</td>
<td>14</td>
<td>151–425</td>
<td>14</td>
<td>13</td>
</tr>
</tbody>
</table>

**Example 5.13** The lifetime of an electrical component in hours, say $T$, can be well modelled by the distribution $\text{Gam}(\alpha = 2, \beta = 1)$. What is the probability that a component will last for more than three hours?

**Solution** From the information, $T \sim \text{Gamma}(\alpha = 2, \beta = 1)$. The required probability is therefore

$$P(T > 3) = \int_{3}^{\infty} \frac{1}{1^2 \Gamma(2)} t^{2-1} \exp(-t/1) \, dt$$

$$= \int_{3}^{\infty} t \, \exp(-t) \, dt$$

since $\Gamma(2) = 1! = 1$. This expression can be integrated using integration by parts as follows:

$$P(T > 3) = \left[ \frac{t \exp(-t)}{3} \right]_{3}^{\infty} - \int_{3}^{\infty} \exp(-t) \, dt$$

$$= \left[ (0) - \left\{ -3 \exp(-3) \right\} \right] - \left\{ \exp(-t) \right\}_{3}^{\infty}$$

$$= 3 \exp(-3) + \exp(-3)$$

$$= 0.1991$$
Note that this is only possible since $\alpha$ is integer. If $\alpha$ was, for example, 2.5, this could not have been done. It is readily calculated using R:

```
> gamma(2.5)
[1] 1.3293
```

Maxima:

```
(%i11) gamma(2.5);
(%o11) 1.329340388179137
```

Scilab:

```
-->gamma(2.5)
ans = 1.3293404
```

It will not be necessary to evaluate a Gamma function in calculating probabilities or quantiles because that is in-built to the `pgamma` and `qgamma` functions in R.

In R, the parameters for the gamma distribution are called `shape` ($\alpha$) and `scale` ($\beta$). The above probability $P(T > 3)$ is determined by:

```
> pgamma(q=3,shape=2,scale=1,lower.tail=F)
[1] 0.19915
```

**Example 5.14** If $X \sim \text{Gamma}(\alpha, \beta)$, find the distribution of $Y = kX$ for some constant $k$.

**Solution** One way to approach this question is to use moment generating functions. Since $X$ has a gamma distribution, $M_X(t) = (1 - \beta t)^{-\alpha}$. Now,

\[
M_X(t) = \mathbb{E}(\exp(tX)) \quad \text{by definition of the mgf}
\]

\[
= \mathbb{E}(\exp(tkX)) \quad \text{since } X = kX
\]

\[
= \mathbb{E}(\exp(sX)) \quad \text{by letting } s = kt
\]

\[
= M_X(s) \quad \text{by definition of the mgf}
\]

\[
= M_X(kt)
\]

\[
= (1 - \beta kt)^{-\alpha}.
\]

This is just the mgf for random variable $X$ with $k\beta$ in place of $\beta$, so the distribution of $X$ is $\text{Gamma}(\alpha, k\beta)$. 

Apart from the uniform distribution, all the continuous distributions considered so far have densities which are positive over an infinite interval. It is useful to have another class of distributions that can be used to model phenomena constrained to a finite interval. The beta distribution fits this category.

Definition 5.13 A random variable $X$ with probability density function

$$f_X(x) = \frac{x^{m-1}(1-x)^{n-1}}{B(m,n)}, \quad 0 \leq x \leq 1, \quad m > 0, \quad n > 0$$

where

$$B(m,n) = \int_{0}^{1} x^{m-1}(1-x)^{n-1} \, dx, \quad m > 0, \quad n > 0 \quad (5.5)$$

is said to have a beta distribution with parameters $m$ and $n$. We write $X \sim \text{beta}(m,n)$.

$B(m,n)$ defined by $(5.5)$ is known as the beta function with parameters $m$ and $n$.

We see from $\int_{0}^{1} f_X(x) \, dx = 1$ that

$$\int_{0}^{1} \frac{x^{m-1}(1-x)^{n-1}}{B(m,n)} \, dx = 1. \quad (5.6)$$

Some properties of the beta function follow.

Theorem 5.14 The beta function $(5.5)$ satisfies the following:

1. The beta function is symmetric in $m$ and $n$. That is, if $m$ and $n$ are interchanged, the function remains unaltered; ie

$$B(m,n) = B(n,m).$$
Module 5. Standard continuous distributions

2. $B(1, 1) = 1$

3. $B\left(\frac{1}{2}, \frac{1}{2}\right) = \pi$

4. For positive constants $m, n$ (not necessarily integers)

$$B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m + n)}.$$

Proof

1. To prove this, put $z = 1 - x$ and hence $dz = -dx$ in (5.5). We obtain

$$B(m, n) = -\int_{1}^{0} (1 - z)^{m-1} z^{n-1} dz$$
$$= \int_{0}^{1} z^{n-1}(1 - z)^{m-1} dz$$
$$= B(n, m)$$

2. Put $x = \sin^2 \theta$, $dx = 2 \sin \theta \cos \theta d\theta$ in (5.5). We have

$$B(m, n) = 2 \int_{0}^{\pi/2} \sin^{2m-1} \theta \cos^{2n-1} \theta d\theta.$$ 

So, for $m = n = \frac{1}{2}$,

$$B\left(\frac{1}{2}, \frac{1}{2}\right) = 2 \int_{0}^{\pi/2} d\theta = \pi$$

3. The proof (not required) involves expressing $\Gamma(m)$ as

$$2 \int_{0}^{\infty} e^{-u^2} u^{2m-1} du,$$

then writing $\Gamma(m)\Gamma(n)$ as a double integral. Changing to polar coordinates, the double integral is written as a repeated integral which is $

\Gamma(m + n) \ B(m, n)$.

Typical graphs for the beta pdf are given below. Note that, if $m = n$, the distribution is symmetric about $x = \frac{1}{2}$, and in the special case where $m = n = 1$, the beta distribution becomes the uniform distribution on (0,1). Some basic properties of the beta distribution follow.
5.5. Beta distribution

![Figure 5.23: Various beta pdf's.](image)

Theorem 5.15 If $X \sim \text{beta}(m, n)$ then

1. $E(X) = \frac{m}{m+n}$
2. $\text{var}(X) = \frac{mn}{(m+n)^2(m+n+1)}$
3. A mode occurs at $x = \frac{m-1}{m+n-2}$ for $m, n > 1$.

Proof Assume $X \sim \text{beta}(m, n)$, then

$$
\mu'_r = \mathbb{E}(X^r) = \int_0^1 \frac{x^r x^{m-1}(1-x)^{n-1}}{B(m,n)} \, dx
$$

$$
= \frac{B(m+r, n)}{B(m,n)} \int_0^1 \frac{x^{m+r-1}(1-x)^{n-1}}{B(m+r,n)} \, dx
$$

$$
= \frac{\Gamma(m+r)\Gamma(n)}{\Gamma(m+r+n)\Gamma(m)\Gamma(n)}
$$

$$
\Gamma(m+r)\Gamma(n)
$$

$$
\Gamma(m+r+n)\Gamma(m)
$$

Putting $r = 1$ and $r = 2$ into the above expression, and using the fact that $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$, it is easy to verify that the mean and variance are $E(X) = \frac{m}{m+n}$ and $\text{var}(X) = \frac{mn}{(m+n)^2(m+n+1)}$.

Any mode $\theta$ of the distribution for which $0 < \theta < 1$ will satisfy $f'_X(\theta) = 0$. From the definition we see that for $m, n > 1$ and $0 < x < 1$,

$$
f'_X(x) = (m-1)x^{m-2}(1-x)^{n-1} - (n-1)x^{m-1}(1-x)^{n-2}/B(m,n) = 0
$$
implies

\[(m - 1)(1 - x) = (n - 1)x\]

which is satisfied by \(x = (m - 1)/(m + n - 2)\).

The mgf of the beta distribution cannot be written in terms of standard functions.

The distribution function of the beta must be evaluated numerically in general, except when \(m\) and \(n\) are integers as shown in the example below.

**Example 5.15** The bulk storage tanks of a fuel retailed are filled each Monday. The retailer has observed that over many weeks the proportion of the available fuel supply sold is well modelled by a beta distribution with \(m = 4\) and \(n = 2\). According to this model, on average what proportion of fuel is sold each week? Is it likely that at least 90% of the supply will sell in a given week?

If \(X\) denotes the proportion of the total supply sold in a given week.

\[E(X) = m/(m + n) = 4/6 = 2/3\]

In the second part our interest is in

\[
P(X > 0.9) = \int_{0.9}^{1} \frac{\Gamma(4 + 2)}{\Gamma(4)\Gamma(2)} x^3(1 - x) dx
\]

\[= 20 \int_{0.9}^{1} (x^3 - x^4) dx
\]

\[= 20(0.004)
\]

\[= 0.08\]

It is unlikely that 90% of the supply will be sold in a given week.

### 5.6 Self-assessment exercises

The following exercises are designed to provide practice at problem-solving based on the material in this module. Solutions are provided at the end of the module. Additional exercises are available in the next section and in the textbook.

In Questions 5.1 and 5.2 use tables of the standard normal to obtain the answers then check these answers using the \texttt{pnorm} or \texttt{qnorm} functions.
5.6. Self-assessment exercises

Ex. 5.1 If $Z \sim N(0, 1)$ find

(a) $P(Z \leq 1.75)$
(b) $P(Z > 2.25)$
(c) $P(Z < -0.85)$
(d) $P(Z \geq -1.64)$
(e) $P(|Z| < 1.84)$

Ex. 5.2 If $Z \sim N(0, 1)$ find $k$ such that

(a) $P(Z < -k) = 0.25$
(b) $P(|Z| \leq k) = .25$

Ex. 5.3 The life time in years of new restaurants initiated in a given city is random variable with an approximate gamma distribution with $\alpha = 1$ and $\beta = 5$. Find

(a) the mean and the variance of the life time of new restaurants in that city;
(b) the probability that at most 5 of the 20 new restaurants will fail within the next year.

Ex. 5.4 The exponential distribution is sometime called the lifetime distribution, since it is commonly used to model the lifetimes of manufactured components. Assume that the lifetime of diodes follows an exponential distribution with mean $\lambda = 40$ hours.

(a) Find the probability that a diode will last longer than 60 hours.
(b) Find the probability that three diodes out of five will last longer than 60 hours.

Ex. 5.5 In the following situations, determine which distribution studied in the course would be most appropriate for modelling the variable of interest, justifying your answer.

(a) The number of people using an Automatic Teller Machine per 15 minutes.
(b) The number of independent shots taken at a target until a missile hits the target.
(c) The heights of two year old children at a Day Care Centre.
(d) The time interval between successive births at a busy hospital.
Ex. 5.6 If $H$ is a binomial random variable such that $H \sim \text{bin}(n, p)$, then the normal approximation can be used under certain conditions.

(a) State the conditions under which the normal approximation is satisfactory.

(b) In this context, explain what is meant by a continuity correction, and why it is necessary.

(c) If $H$ represents the number of heads in 100 tosses of a fair coin, use the normal approximation to determine $P(H < 45)$.

Ex. 5.7 (Computer exercise) Generate a random sample of size 1000 from a uniform distribution on the interval $[-1, 1]$. ($\text{runif}(n=1000, \text{min}=-1, \text{max}=1)$)

(a) Find the mean and standard deviation of the sample. Is it as predicted by Theorem 5.2?

(b) Draw a histogram letting R choose the number of bins and midpoints for the bins. A common problem encountered here is that the two extreme bins are centred at the endpoints causing them to only be half the height you would expect. (Why?) If this occurs redraw the histogram choosing the centres of the bins to be at $-0.9, -0.7, -0.5, \ldots, 0.7, 0.9$.

Ex. 5.8 (Computer exercise)

(a) Generate a sample of size 1000 from an exponential distribution with rate parameter $\lambda = 0.5$.

(b) Find the sample mean and variance and compare with the theoretical result from Theorem 5.8.

(c) Estimate $P(X \geq 4)$ and compare it with the true result found by integrating the density function.

(d) Draw a histogram, (be careful, make sure the first bin is not centred at zero).

(e) Draw a boxplot of the random sample. What does it show?

Ex. 5.9 (Computer exercise) Plot the pdf and distribution function for the standard normal curve on the one set of axes.
5.7 Exercises

Ex. 5.10 If $X \sim N(12, 16)$ find

(a) $P(X < 9)$
(b) $P(X \geq 17)$
(c) $P(0 < X \leq 18)$

Ex. 5.11 If $X \sim N(5, 2.5^2)$ find $k$ such that

(a) $P(X > k) = 0.75$
(b) $P(|X - 5| > k) = 0.01$

Ex. 5.12 At one locality near a busy airport the noise level of a jet aircraft taking off is approximately normally distributed with a mean of 101 decibels and a standard deviation of 5 decibels.

(a) What is the probability that a randomly selected jet will generate a noise level greater than 110 decibels.
(b) Estimate the probability in (a) by simulating the noise level of 1000 such jets.
(c) What is the probability that if four jets are selected at random none will generate a noise level greater than 110 decibels?
(d) If a noise regulation is introduced which specifies that 95% of all jets taking off must have noise levels less than 105 decibels by how much must the mean noise level of aircraft be reduced assuming the standard deviation remains the same?
(e) (Computer exercise) Use R to simulate the problem in part (d).

Ex. 5.13 Given that the rv $X$ has mgf given by $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$ show that

(a) $Y = aX + b$ has a normal distribution and write down its mean and variance.
(b) Hence show that $Z = (X - \mu)/\sigma$ has a standard normal distribution and write down its mgf.

Ex. 5.14 A large service company receives calls from would be subscribers. If a company representative is not free to talk to them they are put on hold. From past experience the company knows that most potential customers will be lost if they are required to hold longer than 4 minutes. It is assumed that the holding time is normally distributed with standard deviation 0.7 minutes.
(a) If the average holding time is 2.8 minutes what proportion of potential customers will be lost?

(b) The holding time is varied by providing more (or fewer) representatives. What holding time must be attained if the company aims to lose no more than 1% of potential customers.

**Ex. 5.15** If $Y$ is a random variable with pdf given by

$$ f_Y(y) = \frac{y^{m-1}(1-y)^{n-1}}{B(m, n)}, \quad 0 \leq y \leq 1, $$

find

(a) the mode of the distribution

(b) $E(Y)$

(c) $\text{var}(Y)$

**Ex. 5.16** For a particular variety of corn on average $2/3$ of all seeds will germinate.

(a) If 120 seeds are planted find the number expected to germinate and its variance.

(b) Find a lower bound on the probability that the number germinating will be between 70 and 90.

**Ex. 5.17** A task is set which is claimed will take an average worker 2 minutes. If it is assumed that the time taken is exponentially distributed what could be said if a worker takes 4 minutes to complete the task?

**Ex. 5.18** Consider $n$ random variables $Z_i$ such that $Z_i \sim \exp(\beta)$ for every $i = 1, \ldots, n$. (That is, $Z_1, Z_2, \ldots, Z_n$ all have the distribution $\exp(\beta)$.) Show that the distribution of $Z_1 + Z_2 + \cdots + Z_n$ has the gamma distribution $\text{Gamma}(n, \beta)$.

**Ex. 5.19** The exponential distribution is said to be *memoryless* in that

$$ P(X > s + t \mid X > t) = P(X > s) $$

for all $s, t \geq 0$.

(a) Explain, in simple terms, why the above mathematical expression implies that the distribution is *memoryless*.

(b) Prove that the exponential distribution is indeed memoryless.
(c) Consider the exponential distribution
\[ f_P(p) = \frac{1}{\alpha} \exp(-p/\alpha) \quad \text{for } p \geq 0. \]
Determine \( P(P > 4 \mid P > 1) \).

**Ex. 5.20** The amount of time in minutes, \( T \), spent waiting for an elevator on the ground floor of a building is given by the function
\[ f_T(t) = \frac{1}{\Gamma(1/2)} t^{-1/2} e^{-t} \quad \text{for } t > 0 \]
where \( \Gamma(1/2) = \sqrt{\pi} \).

(a) By noting that the distribution is in the gamma family, write down the moment generating function of \( T \).
(b) Use the mgf to determine the mean and variance of \( T \).

**Ex. 5.21** Consider the distribution
\[ f_Z(z) = k \exp(-z^4) \quad \text{for } -\infty < z < \infty \]
where \( k = \frac{\pi}{\sqrt{2\Gamma(3/2)}} \approx 1.8128 \). Using a numerical method, determine \( P(0 < z < 1) \) accurate to at least 2 decimal places.

**Ex. 5.22** In 1951, Gupta [15] studied the lifetimes of 300 electric lamps, given in Table 5.3. The data has also been looked at by others.

(a) Convert the counts to probabilities.
(b) Determine the sample mean and sample variance of the lifetime of the electric lamps using the given data.
(c) Sketch a histogram of the data, and comment on its normality.
(d) Assume the data has a normal distribution with the mean and variance as found in part (b). Determine the theoretical probabilities of a lamp lasting between 1300 and 1500 hours, and compare to the probability computed from the given data. Comment on your answers.
(e) Using the theoretical normal distribution developed above, determine the probability that a lamp will last for longer than 1600 hours given that it has lasted 1300 hours.

**Ex. 5.23** Rainfall is often modelled using the gamma distribution or the exponential distribution for simplicity. However, sometimes other distributions are used. An example of such a distribution is described by
Module 5. Standard continuous distributions

<table>
<thead>
<tr>
<th>Lifetimes (hours)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>950-1000</td>
<td>2</td>
</tr>
<tr>
<td>1000-1050</td>
<td>2</td>
</tr>
<tr>
<td>1050-1100</td>
<td>3</td>
</tr>
<tr>
<td>1100-1150</td>
<td>6</td>
</tr>
<tr>
<td>1150-1200</td>
<td>7</td>
</tr>
<tr>
<td>1200-1250</td>
<td>12</td>
</tr>
<tr>
<td>1250-1300</td>
<td>16</td>
</tr>
<tr>
<td>1300-1350</td>
<td>20</td>
</tr>
<tr>
<td>1350-1400</td>
<td>24</td>
</tr>
<tr>
<td>1400-1450</td>
<td>27</td>
</tr>
<tr>
<td>1450-1500</td>
<td>29</td>
</tr>
<tr>
<td>1500-1550</td>
<td>29</td>
</tr>
<tr>
<td>1550-1600</td>
<td>28</td>
</tr>
<tr>
<td>1600-1650</td>
<td>25</td>
</tr>
<tr>
<td>1650-1700</td>
<td>21</td>
</tr>
<tr>
<td>1700-1750</td>
<td>16</td>
</tr>
<tr>
<td>1750-1800</td>
<td>12</td>
</tr>
<tr>
<td>1800-1850</td>
<td>8</td>
</tr>
<tr>
<td>1850-1900</td>
<td>6</td>
</tr>
<tr>
<td>1900-1950</td>
<td>3</td>
</tr>
<tr>
<td>1950-2000</td>
<td>2</td>
</tr>
<tr>
<td>2000-2050</td>
<td>1</td>
</tr>
<tr>
<td>2050-2100</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.3: Lifetimes of Electric Lamps (Source: Gupta [15])

the following functions

\[ f_R(r) = \frac{1 - p}{\Gamma(\alpha)\beta^\alpha} r^{\alpha-1} \exp(-r/\beta) \quad \text{if } r > 0 \]

\[ p_R(r) = p \quad \text{if } r = 0 \]

where \( R \) is the daily rainfall in millimetres.

(a) Determine if \( R \) is a discrete, continuous or mixed distribution, justifying your answer.

(b) Find \( f_{R|R>0}(r \mid R > 0) \), and show that it is a gamma distribution.

(c) Explain, in the context of the question, the difference between \( P(R > 10) \) and \( P(R > 10 \mid R > 0) \).

(d) Given the data in Table 5.4, determine a value for \( p \).

(e) Four days at random are independently drawn from the data in Table 5.4. Calculate the probability of getting more than two dry days in the four days.
### 5.7. Exercises

<table>
<thead>
<tr>
<th>Rainfall (in mm)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2237</td>
</tr>
<tr>
<td>0 &lt; Rain ≤ 25</td>
<td>809</td>
</tr>
<tr>
<td>25 &lt; Rain ≤ 50</td>
<td>260</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>93</td>
</tr>
<tr>
<td>100</td>
<td>81</td>
</tr>
<tr>
<td>125</td>
<td>56</td>
</tr>
<tr>
<td>150</td>
<td>43</td>
</tr>
<tr>
<td>175</td>
<td>12</td>
</tr>
<tr>
<td>200</td>
<td>17</td>
</tr>
<tr>
<td>225</td>
<td>5</td>
</tr>
<tr>
<td>250</td>
<td>12</td>
</tr>
<tr>
<td>275</td>
<td>5</td>
</tr>
<tr>
<td>300</td>
<td>3</td>
</tr>
<tr>
<td>325</td>
<td>1</td>
</tr>
<tr>
<td>350</td>
<td>5</td>
</tr>
<tr>
<td>375</td>
<td>2</td>
</tr>
<tr>
<td>400</td>
<td>3</td>
</tr>
<tr>
<td>700</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 5.4: Daily rainfall for Melbourne, from 1 January 1981 to 31 December 1990 (Original Source: Australian Bureau of Meteorology)

**Ex. 5.24** The life of a certain electrical component has probability density function

\[ f_T(t) = \alpha e^{-\alpha t}, \quad t \in (0, \infty) \]

(a) Find the probability that the life of the component will be more than \( t_1 \) hours.

(b) Find the conditional probability that the lifetime of the component will be more than \( t_1 + t_2 \) hours given that the component has already been in operation for \( t_1 \) hours.

(c) Assume that a machine using such a component will fail when the component fails. If the machine is inspected at time \( t_1 \) and will not be inspected again until time \( t_2 \) is the best strategy to replace the component at time \( t_1 \) with a new one even though it shows no sign of failing? (Assume the exponential model is correct!) Explain your answer.

**Ex. 5.25** If a company employs \( n \) sales persons, its gross sales in thousands of dollars may be regarded as a rv having a gamma distribution with \( \alpha = 80\sqrt{n} \) and \( \beta = 2 \). If the sales cost is $8000 per salesperson, how many salespersons should the company employ to maximise the expected profit?
Ex. 5.26 The Cauchy distribution has the pdf
\[ f_X(x) = \frac{k}{1 + x^2} \]
for \(-\infty < x < \infty\).

(a) Find the value of \(k\).

(b) Sketch the Cauchy distribution over the range \(-6 \leq x \leq 6\).

(c) Determine \(P(X > 2 \mid X < 4)\).

(d) Find the distribution function for \(X\).

(e) Using the distribution function, determine the median of the Cauchy distribution.

Ex. 5.27 (a) Suppose in Question 5.24 the mean lifetime is 500 hours (= 1/\(\alpha\)). Calculate the exact probabilities in parts (a) and (b) using the \texttt{pexp()} function if \(t_1 = 500\) and \(t_1 + t_2 = 1000\).

(b) Estimate the probability in part (a) by simulating the lifetimes of 1000 components.

Ex. 5.28 The von Mises distribution is often used to model directional data. (It was used, for example, by Fisher and Lee \cite{11} to model the direction and distance travelled by blue periwinkles after they had been transplanted downshore from the height at which they normally live.) The von Mises distribution has the probability density function
\[ f_Y(y; \lambda) = \begin{cases} \kappa \exp\{\lambda \cos(y - \mu)\} & \text{if } 0 \leq y < 2\pi \\ 0 & \text{otherwise} \end{cases} \]
where \(\lambda > 0\) and \(0 \leq \mu < 2\pi\).

The density function cannot be integrated using standard methods (so don’t try!), and so the constant, \(\kappa\) (kappa), must be determined by using a numerical integration method.

(a) Using such a method, determine the value of the constant \(\kappa\) in the case where \(\lambda = 1\) and \(\mu = \pi\), accurate to two significant figures. (You are encouraged to use a computer package such as MATLAB, R or Excel, though this is not essential.)

(b) Numerically determine \(P(Y > 2)\) when \(\lambda = 1\) and \(\mu = \pi\).

Ex. 5.29 (Computer exercise) Use the \texttt{dexp} and \texttt{pexp} commands to calculate the density function and corresponding distribution function for an exponential distribution with parameter \(\theta = 0.5\). Plot the resulting pdf and df on the one set of axes for \(x = 0, 1, 2, \ldots, 15\).
5.7. Exercises

Ex. 5.30 (Computer exercise) Plot the pdf’s for the three normal distributions that each have mean 0 but have standard deviations 0.5, 1 and 2 on the one set of axes. (In each case calculate the pdf for \( x \in [-3, 3] \) at \( x \) increments of 0.1.)

Ex. 5.31 (Computer exercise) Simulate Example 5.7 for a class size of 50. What is the lowest mark to pass if the top 75% pass. How does it compare with the theoretical result (which essentially assumes an infinite population). (Try it for a larger class size. As the class size increases you would expect the theoretical result and estimated result to get closer.)

Ex. 5.32 (Computer exercise) To see the effect of the second parameter \( \beta \), plot the gamma functions with a common \( \alpha = 2 \) but \( \beta = 1, 2 \) and 4. Use \( x \in [0, 6] \) with an increment of 0.1.

Ex. 5.33 (Computer exercise) On the one graph, plot the chi-square distribution for \( \nu = 1, 2, 3, 4 \). Use \( x \in [0.1, 5] \) with increments of 0.1.

Ex. 5.34 Testing for Normality If we are given a random sample we can use R to examine if it is likely to have been drawn from a normal distribution. To do this we use a normal probability plot.

(a) Generate a sample of 100 random variables from a normal distribution with mean 3 and standard deviation 2, `rnorm(n=100,mean=3,sd=2)`.

(b) Get a normal probability graph of this data by

```
qqnorm(y)
qqline(y,col=2)
```

If the data is normally distributed the graph should be approximately linear.

(c) Generate a sample of size 100 from an exponential distribution with parameter 1 and obtain a normal probability plot for this data. What do you observe?

(Try a few distributions for yourself. You could try uniform, Cauchy, gamma, binomial, etc.)

Ex. 5.35 (Computer exercise) Estimate \( \int_0^1 e^{-x} dx \) by considering the integral as the expected value of \( e^{-X} \) where \( X \) is distributed uniformly on \([0, 1]\), this is called estimation by Monte-Carlo integration. That is

\[
E(e^{-X}) = \int_0^1 e^{-x} f(x) dx, \quad \text{where } f(x) = 1, \quad 0 \leq x \leq 1
\]

(In general, if \( X \) is uniformly distributed on \([0, 1]\) then \( E(g(X)) \) can be estimated by first generating a random sample of size \( N \) from the
uniform distribution on \([0,1]\) and then using the fact that \(E(g(X))\) can be approximated by \(\frac{1}{N} \sum_{i=1}^{N} g(x_i)\).}

5.8 Some answers and hints

5.1 (a) 0.9599 (b) 0.01222 (c) 0.1977 (d) 0.9495 (e) 0.9342

5.2 (a) \(k = 0.67\) (b) \(k = 0.32\)

5.3 Let \(X\) be lifetime of new restaurants in the city in years. Then \(X \sim \text{Gamma}(\alpha = 1, \beta = 5)\).

(a) The expected number of years before a new restaurant fails is \(E(X) = \alpha \times \beta = 1 \times 5 = 5\). Also, \(\text{var}(X) = \alpha \times \beta^2 = 1 \times 5^2 = 25\).

(b) The probability that one randomly selected new restaurant failed within the first year is

\[
p = \int_{x=0}^{1} \frac{1}{5} \exp\left(-\frac{x}{5}\right) dx = 1 - \exp\left(-\frac{1}{5}\right) \approx 0.18.
\]

Let \(X\) be the number of new restaurants failing within the first year. Then \(X \sim \text{bin}(n = 20, p = 0.18)\) so \(P(X \leq 5) = 0.781\) using the binomial distribution.

5.4 Let \(D\) be the lifetime of a diode in hours so that \(D \sim \text{exp}(40)\); then \(P(D > 60) = \int_{60}^{\infty} \exp(-t/40)/40 dt \approx 1 - 0.777 = 0.223\); let \(N\) be the number of diodes out of five that last longer than 60 hours; then \(N \sim \text{bin}(n = 5, p = 0.223)\), so that \(P(N = 3) \approx 0.067\).

5.5 Poisson or negative binomial; geometric; normal; exponential.

5.6 \(P(H < 45) = P(Z < (44.5 - 50)/5) = P(Z < -1.1) = 0.136\).

5.7 \(\text{urv}<\text{runif}(n=1000,\text{min}=-1,\text{max}=1)\)

\[
\begin{align*}
\text{print(}\text{mean(}\text{urv)}\text{))} \\
\text{print(}\sqrt{\text{var(}\text{urv)}\text{))} \\
\text{hist(}\text{urv,}\text{breaks=seq(-1,1,0.2))}
\end{align*}
\]

5.8 \(v<\text{rexp}(n=1000,\text{rate}=0.5)\)

\[
\begin{align*}
\text{mnv}<\text{mean(}v\text{)} \\
\text{sdv}<\text{var(}v\text{)} \\
\text{Pgt4}<\text{1-pexp(}q=4,\text{rate}=0.5) \\
\text{print(}c(\text{mnv,}\text{sdv,}\text{Pgt4})) \\
1<\text{length(}v[\text{v}>4]) \\
\text{hist(}v\text{)}
\end{align*}
\]
5.8. Some answers and hints

I obtained 136 observations greater than 4, giving an estimate of $P(X > 4)$ of 0.136. That is about 13% of workers would be expected to take 4 minutes or more. A worker taking 4 minutes to complete the task should not be considered unusual under the assumptions made.

5.9

```r
x <- seq(-3, 3, 0.1)
phix <- dnorm(x, mean = 0, sd = 1)
PHIx <- pnorm(x, mean = 0, sd = 1)
plot(x, PHIx, type = 'l', las = 1, xlab = 'x', ylab = 'density, distribution')
lines(x, phix, lty = 2)
```

5.10  
(a) 0.2266  
(b) 1.1056  
(c) 0.93185

5.11  
(a) 3.3  
(b) 6.45

5.12  
(a) 0.0359  
(b) Use `rnorm` to simulate 1000 observations. Use the comparison operator, ($>$), and count the number of observations greater than 110.  
(c) 0.8639  
(d) Let $\mu$ be the mean noise level to be attained so $X \sim N(\mu, 5^2)$. Reduction = $101 - \mu = 4.2$ decibels  
(e) Find the observation in (b) above which there are 5% of the observations. That is sort the column (`sort`) and look for the observation in position 950 in the sorted column. The required reduction is the value required to make this observed value 105.

5.13  
(a) Use Theorem 3.17 and show $M_Y(t)$ has the form of mgf for normal distribution.

5.14  
(a) 0.0436  
(b) 2.3 mins

5.15  
(a) $\frac{m - 1}{m + n - 2}$  
(b) $\frac{m}{m + n}$  
(c) $\frac{mn}{(m + n)^2(m + n + 1)}$

5.16  
Let $X$ be the number of seeds that germinate.  
(a) $E(X) = 80$  
(b) $\text{var}(X) = 80/3$  
(b) $\text{diff}(\text{pbinom}(q = \text{c}(70, 89), \text{size} = 120, \text{prob} = 2/3))$

5.24  
(a) Integrate to find $P(T > t_1) = e^{-\alpha t_1}$.  
(b) This is a conditional probability, $P(A \mid B)$ where $A = \{T > t_1 + t_2\}$, $B = \{T > t_1\}$.  
(c) Explain which has the greater probability of failing (assuming the exponential model) in the next $t_2$ hours, a new component or a component which has already lasted $t_1$ hours?

5.25  
Let $X$ be the gross sales in thousands of dollars. Then $X \sim \text{Gamma}(\alpha = 80\sqrt{n}, \beta = 2)$. Since the cost per salesperson is $8 \ 000, the profit (in thousand dollars) is a random variable denoted by $Y = X - 8n$ (that is, the difference between the total gross sale and the total cost of employing $n$ salespersons). The expected profit is given by
$E(Y) = E(X - 8n) = E(X) - 8n = 160\sqrt{n} - 8n$. The value of $n$ maximizing expected profit is found by differentiating.

\[
\frac{d}{dn} E(Y) = 80n^{-1/2} - 8,
\]

which is zero when $n = 100$. (NOTE: The second derivative can be used to confirm this is a maximum.)

5.26 (a) $k = 1/\pi$ (c) $[\arctan(4) - \arctan(2)]/[\arctan(4) + \pi]$ (c) $F_X(x) = \arctan(2)/\pi + 1/2$ (d) median is 0.

5.27 (a) Use `pexp(rate = 1/500, q = 500)`. (b) Use `rexp()`. Count the number of observations greater than 500.

5.29 \(x<-0:0.15\)
\(PDFx<-dexp(x,rate=0.5)\)
\(CDFx<-pexp(x,rate=0.5)\)
\(plot(x,PDFx,ylim=c(0,1),type='l',las=1)\)
\(lines(x,CDFx,lty=3)\)

5.30 \(x<-seq(-3,3,0.1)\)
\(sd<-c(0.5,1,2)\)
\(fx<-array(0,dim=c(length(x),length(sd)))\)
\(\# \text{ set up the plot frame with a dummy}\)
\(plot(x,rep(0,length(x)),type='n',las=1,xlab="x",ylab="f(x)",ylim=c(0,1))\)
\(\text{for ( } k \text{ in seq(along=sd)){}\)
\(fx[,k]<-dnorm(x,mean=0,sd=sd[k])\)
\(lines(x,fx[,k],lty=k)\)
\(}\)

5.31 \(marks<-rnorm(mean=500,sd=100,n=50)\)
\(pass.mark<-quantile(marks,probs=0.25)\)

5.32 \(x<-seq(0,6,0.1)\)
\(alpha<-2\)
\(beta<-c(1,2,4)\)
\(ymax<-1.5\)
\(plot(x,rep(0,length(x)),type='n',xlab="x",\)
\(\text{ylab="gamma density",las=1,ylim=c(0,ymax))}\)
\(\text{for ( } i \text{ in seq(along=beta)){}\)
\(lines(x,dgamma(x,shape=alpha, scale=1/beta[i]),lty=i)\)
\(}\)
\(# \text{ legend with upper left corner denoted by (p1,p2) coordinate}\)
\(\text{p1<-2}\)
\(\text{p2<-0.9*ymax}\)
\(\text{legend(p1,p2,legend=c("beta=1","beta=2","beta=4"),lty=1:3)}\)
5.8. Some answers and hints

5.33 \( \text{nu}<-1:4 \)
\( \text{x<-seq}(0.1,5,0.1) \)
\( \text{ymax<-1.5} \)
\( \text{plot(x,rep(0,length(x)),type='n',las=1,ylim=c(0,ymax),} \)
\( \quad \text{xlab="x",ylab="Chi-Square density"}) \)
\( \text{for ( j in seq(along=nu))}{} \)
\( \text{lines(x,dchisq(x,df=nu[j]),lty=j}) \}
\( \text{p1<-2} \)
\( \text{p2<-0.9*ymax} \)
\( \text{legend(p1,p2,legend=c("nu=1","nu=2","nu=3","nu=4"),lty=1:4)} \)

5.35 \( \text{x<-runif(n=1000,min=0,max=1)} \)
\( \text{expX<-exp(-x[x>0])} \quad \text{# read this as \(-x\) s.t. \(x>0\)} \)
\( \text{print(mean(expX))} \)
Module 5. Standard continuous distributions
Module 6

Bivariate distributions

Contents

6.1 Introduction .............................................. 208
    6.1.1 Marginal distributions .......................... 216
    6.1.2 Conditional distributions ....................... 219
6.2 Independent random variables ......................... 222
6.3 Expectations involving bivariate distributions .... 226
    6.3.1 Moments of a bivariate distribution ............ 228
    6.3.2 Properties of covariance and correlation ...... 231
6.4 Conditional expectations ............................... 233
    6.4.1 Conditional mean ................................. 233
    6.4.2 Conditional variance ............................. 234
6.5 The multivariate extension ............................. 237
    6.5.1 Vector formulation ................................ 238
6.6 Multinomial distribution ............................... 240
6.7 The bivariate normal distribution .................... 243
6.8 Self-assessment exercises .............................. 249
6.9 Exercises ................................................ 250
6.10 Some answers and hints ............................... 253
Module objectives

Upon completion of this module students should be able to:

- understand the concept of the joint probability function and the distribution function of two random variables
- find the marginal and conditional probability functions of random variables in both discrete and continuous cases
- understand and apply the concept of independence of two random variables

Reading 6.1  DGS, Section 3.4; WMS, Sections 5.1 and 5.2.

6.1 Introduction

Not all random experiments are sufficiently simple to have the outcome denoted by a single number \( x \). In many situations we are interested in observing two or more numerical characteristics simultaneously.

This section only discusses the two-variable, or bivariate, case.

Let \( E \) be an experiment and \( S \) a sample space associated with \( E \).

Definition 6.1  Let \( X = X(s) \) and \( Y = Y(s) \) be two functions, each assigning a real number to each sample point \( s \in S \). Then \( (X, Y) \) is called a two-dimensional random variable, or a random vector.

The range space of \( (X, Y) \), \( R_{X \times Y} \), will be a subset of the Euclidean plane. Each outcome \( X(s), Y(s) \) may be represented as a point \((x, y)\) in the plane. As in the 1-dimensional case, it is necessary to distinguish between discrete and continuous random variables.

Definition 6.2  (a) \( (X, Y) \) is a 2-dimensional discrete random variable if the range space \( R_{X \times Y} \) is finite or countably infinite. That is, if values of \( (X, Y) \) may be represented as \( (x_i, y_j) \), \( i = 1, 2, \ldots \), \( j = 1, 2, \ldots \).

(b) \( (X, Y) \) is a 2-dimensional continuous random variable if the range space \( R_{X \times Y} \) is a non-denumerable set of the Euclidean plane; for example, \( R_{X \times Y} = \{(x, y) : a \leq x \leq b, c \leq y \leq d \} \).
Definition 6.3  (a) Let \((X, Y)\) be a 2-dimensional discrete random variable. With each \((x_i, y_j)\) we associate a number \(p_{X,Y}(x_i, y_j)\) representing \(P(X = x_i, Y = y_j)\) and satisfying

\[
p_{X,Y}(x_i, y_j) \geq 0, \text{ for all } (x_i, y_j)
\]

\[
\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} p_{X,Y}(x_i, y_j) = 1.
\]

Then the function \(p_{X,Y}\), defined for all \((x_i, y_j) \in \mathbb{R}\) is called the probability function of \((X, Y)\). Also,

\[
\{x_i, y_j, p_{X,Y}(x_i, y_j); i, j = 1, 2, \ldots\}
\]

is called the probability distribution of \((X, Y)\).

(b) Let \((X, Y)\) be a continuous random variable assuming values in a 2-dimensional set \(R\). The joint probability density function, \(f_{X,Y}\) is a function satisfying

\[
f_{X,Y}(x, y) \geq 0, \text{ for all } (x, y) \in \mathbb{R}, \quad (6.1)
\]

\[
\int \int_{\mathbb{R}} f_{X,Y}(x, y) \, dx \, dy = 1. \quad (6.2)
\]

Note that the second of these indicates that the volume under the surface \(f_{X,Y}(x, y)\) is one. Also, for \(\Delta x, \Delta y\) sufficiently small,

\[
f_{X,Y}(x, y) \Delta x \Delta y \approx P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y). \quad (6.3)
\]

Probabilities of events can be determined by the probability function or the probability density function as follows.

Definition 6.4 For any event \(A\), the probability of \(A\) is given by

\[
P(A) = \sum_{(x,y) \in A} p(x,y), \quad \text{for } (X, Y) \text{ discrete} \quad (6.4)
\]

\[
P(A) = \int \int_{(x,y) \in A} f(x,y) \, dx \, dy \quad \text{for } (X, Y) \text{ continuous} \quad (6.5)
\]

As in the univariate case, the (cumulative) distribution function can be used to represent a sum of probabilities or a volume under a surface. It is denoted by \(F_{X,Y}(x,y)\) and defined by
Definition 6.5 Bivariate distribution function

\[ F(x, y) = P(X \leq x, Y \leq y), \text{ for } (X, Y) \text{ discrete} \quad (6.6) \]

\[ F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v) \, du \, dv, \text{ for } (X, Y) \text{ continuous} \quad (6.7) \]

As in the univariate case, a bivariate distribution can be expressed in various ways:

(i) by enumerating the range space and corresponding probabilities

(ii) by a formula

(iii) by a table

The following examples illustrate these concepts.

Example 6.1 Consider an experiment where, simultaneously, two coins are tossed and one die is rolled. Let \( X_1 \) be the number of heads that show on the two coins, and \( X_2 \) the number on the top face of the die. Then \((X_1, X_2)\) is a discrete, bivariate random variable.

Note the possible values of \( X_1 \) are \( R_{X_1} = \{0, 1, 2\} \) and the possible values of \( X_2 \) are \( R_{X_2} = \{1, 2, 3, 4, 5, 6\} \). So the sample space for the random vector \((X_1, X_2)\) is

\[ S = \{(0, 1), (0, 2), \ldots, (0, 6); (1, 1), (1, 2), \ldots, (1, 6); (2, 1), (2, 2), \ldots, (2, 6)\} \]

Example 6.2 Consider the following discrete distribution where probabilities \( P(X = x, Y = y) \) are shown as a graph in Fig. 6.1 and as a Table in Table 6.1. Find \( P(X + Y \geq 2) \).

Note that the probabilities in the table add to one.

Treating \( X + Y \geq 2 \) as an event \( A \), we have,

\[ P(X + Y \geq 2) = P(X = 2, Y = 0 \text{ or } X = 1, Y = 1 \text{ or } X = 0, Y = 2) \]
\[ = P(X = 2, Y = 0) + P(X = 1, Y = 1) + P(X = 0, Y = 2) \]
\[ = \frac{1}{9} + \frac{1}{4} + \frac{1}{3} \]
\[ = \frac{25}{36} \]
6.1. Introduction

Figure 6.1: A bivariate probability function

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1/36</th>
<th>1/6</th>
<th>1/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1/9</td>
<td>1/3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: A bivariate probability function

Example 6.3 Consider the following continuous bivariate distribution with joint pdf

$$f_{X,Y}(x, y) = 1, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1,$$

and find $P \left( 0 \leq x \leq \frac{1}{2}, \ 0 \leq y \leq \frac{1}{2} \right)$. This is sometimes called the bivariate uniform distribution. See Fig. 6.2. Note that the volume under the surface is one.

Now the probability of the event above is the volume above the square with vertices $(0, 0), (0, \frac{1}{2}), (\frac{1}{2}, 0), (\frac{1}{2}, \frac{1}{2})$ and is thus $\frac{1}{4}$.

Example 6.4 Consider again the experiment in Examples 6.1. As an ex-
ample, the joint df at \((1, 2)\) would be computed as follows:

\[
F_{X_1,X_2}(1, 2) = \sum_{x_1 \leq 1} \sum_{x_2 \leq 2} p_{X_1,X_2}(x_1, x_2)
\]

\[
= p_{X_1,X_2}(0, 1) + p_{X_1,X_2}(0, 2) + p_{X_1,X_2}(1, 1) + p_{X_1,X_2}(1, 2)
\]

\[
= 1/24 + 1/24 + 1/12 + 1/12 = 6/24.
\]

The complete joint df is given below. It is reasonably complicated even for this simple example.

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(x_2 &lt; 1)</th>
<th>(0 \leq x_1 &lt; 1)</th>
<th>(1 \leq x_1 &lt; 2)</th>
<th>(x_1 \geq 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_2 \leq 1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1 (\leq x_2 &lt; 2)</td>
<td>0</td>
<td>1/24</td>
<td>3/24</td>
<td>4/24</td>
<td></td>
</tr>
<tr>
<td>2 (\leq x_2 &lt; 3)</td>
<td>0</td>
<td>2/24</td>
<td>6/24</td>
<td>8/24</td>
<td></td>
</tr>
<tr>
<td>3 (\leq x_2 &lt; 4)</td>
<td>0</td>
<td>3/24</td>
<td>9/24</td>
<td>12/24</td>
<td></td>
</tr>
<tr>
<td>4 (\leq x_2 &lt; 5)</td>
<td>0</td>
<td>4/24</td>
<td>12/24</td>
<td>16/24</td>
<td></td>
</tr>
<tr>
<td>5 (\leq x_2 &lt; 6)</td>
<td>0</td>
<td>5/24</td>
<td>15/24</td>
<td>20/24</td>
<td></td>
</tr>
<tr>
<td>(x_2 \geq 6)</td>
<td>0</td>
<td>6/24</td>
<td>18/24</td>
<td>24/24</td>
<td></td>
</tr>
</tbody>
</table>

**Example 6.5** Consider the bivariate discrete distribution which results when two dice are thrown. Let \(X\) be the number of 5’s and \(Y\) the number of 6’s that result. Now range spaces of \(X\) and \(Y\) are \(R_X = \{0, 1, 2\}\), \(R_Y = \{0, 1, 2\}\) and the range space for the experiment is the Cartesian product of \(R_X\) and \(R_Y\), with the interpretation that some
6.1. Introduction

of the resulting points may have probability zero. The probabilities in Table 6.2 are \( P(X = x, Y = y) \) for the \((x, y)\) pairs in the range space. The probabilities in the table are found by considering that we really have two repetitions of a simple experiment with 3 possible outcomes, \{5, 6, 5 or 6\}, with probabilities \( \frac{1}{6}, \frac{1}{6}, \frac{2}{3} \), the same on each repetition. Of course the event \( X = 2, Y = 1 \) cannot occur in two trials, so has probability zero.

Example 6.5 is a special case of the multinomial distribution, (a generalization of the binomial distribution), which will be described later.

Example 6.6 Consider the example in Example 6.1. Since the toss of the coin and the roll of the die are independent, the probabilities are computed as follows:

\[
P(X_1 = 0, X_2 = 1) = P(X_1 = 0) \times P(X_2 = 1) = \frac{1}{4} \times \frac{1}{6} = \frac{1}{24}
\]

\[
P(X_1 = 1, X_2 = 1) = P(X_1 = 1) \times P(X_2 = 2) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}
\]

and so on. The complete joint pf can be displayed in a graph (often tricky), a function, or a table. Here, the joint pf could be given (but is not obvious) as the function

\[
p_{X_1, X_2}(x_1, x_2) = \begin{cases} \left(\frac{1}{12}\right)0.5^{x_1-1} & \text{for } (x_1, x_2) \in S \text{ defined earlier} \\ 0 & \text{elsewhere} \end{cases}
\]

In tabular form (probably clearer in this example), we would have the joint pf as given in Table 6.3.
Table 6.3: The joint pdf for Example 6.6.

Example 6.7  A bank operates both a drive-up and a walk-up window. On a randomly selected day, let $X_1$ be the proportion of time the drive-up facility is in use (at least one customer is being served or waiting to be served), and $X_2$ is the proportion of time the walk-up window is in use. Then the set of possible values for $X_1$ and $X_2$ is the rectangle $R = \{(x_1, x_2) \mid 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$. From experience, the joint pdf of $(X_1, X_2)$ is given by

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} c(x_1 + x_2^2) & \text{for } 0 \leq x_1 \leq 1; 0 \leq x_2 \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Determine a value for $c$.

(b) Compute the probability neither facility is busy more than half the time.

Solution

(a) Obviously, $f_{X_1, X_2}(x_1, x_2) \geq 0$ for all $x_1$ and $x_2$ from the definition provided $c > 0$. Secondly, we need

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) \, dx_1 \, dx_2 = 1.$$
6.1. Introduction

Hence,

\[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2) \, dx_1 \, dx_2 = \int_{0}^{1} \int_{0}^{1} f_{X_1,X_2}(x_1, x_2) \, dx_1 \, dx_2 \]

\[ = \int_{0}^{1} \left\{ \int_{0}^{1} f_{X_1,X_2}(x_1, x_2) \, dx_1 \right\} \, dx_2 \]

\[ = c \int_{x_2=0}^{1} \left\{ \int_{x_1=0}^{1} (x_1 + x_2^2) \, dx_1 \right\} \, dx_2 \]

\[ = c \int_{x_2=0}^{1} (1/2 + x_2^2) \, dx_2 \]

\[ = c(1/2 + 1/3) = 5c/6, \]

and so \( c = 6/5. \)

(b) The question is asking to find \( P(0 \leq X_1 \leq 0.5, 0 \leq X_2 \leq 0.5); \)

call this event \( A. \) Then,

\[ P(A) = \int_{0}^{0.5} \int_{0}^{0.5} f_{X_1,X_2}(x_1, x_2) \, dx_1 \, dx_2 \]

\[ = \frac{6}{5} \int_{0}^{0.5} \left\{ \int_{0}^{0.5} x_1 + x_2^2 \, dx_1 \right\} \, dx_2 \]

\[ = \frac{6}{5} \int_{0}^{0.5} (1/8 + x_2^2/2) \, dx_2 \]

\[ = 1/10. \]

Example 6.8 From Example 6.7,

\[ F_{X_1,X_2}(x_1, x_2) = \frac{6}{5} \int_{0}^{x_1} \int_{0}^{x_2} (t_1 + t_2^2) \, dt_2 \, dt_1 \]

\[ = \frac{6}{5} \int_{0}^{x_1} (t_1 t_2 + t_2^3/3) \bigg|_{t_2=x_2} \, dt_1 \]

\[ = \frac{6}{5} \int_{0}^{x_1} (t_1 x_2 + x_2^3/3) \, dt_1 \]

\[ = \frac{6}{5} \left( \frac{x_1 x_2}{2} + \frac{x_1 x_2^3}{3} \right) \]

for \( 0 < x_1 < 1 \) and \( 0 < x_2 < 1. \) So

\[ F_{X_1,X_2}(x_1, x_2) = \begin{cases} 
0 & \text{if } x_1 < 0 \text{ or } x_2 < 0 \\
\frac{6}{5} \left( x_1 x_2/2 + x_1 x_2^3/3 \right) & \text{if } 0 \leq x_1 \leq 1 \text{ and } 0 \leq x_2 \leq 1 \\
1 & \text{if } x_1 > 1 \text{ and } x_2 > 1 
\end{cases} \]
6.1.1 Marginal distributions

**Reading 6.2** DGS, Section 3.5; WMS, Section 5.3.

With each two-dimensional random variable \((X, Y)\) we associate two one-dimensional random variables, namely \(X\) and \(Y\). We now find the probability distributions of each of \(X\) and \(Y\) separately.

In the case of a discrete random vector \((X, Y)\), the event \(X = x_i\) should be thought of as the union of the mutually exclusive events

\[ \{X = x_i, Y = y_1\}, \{X = x_i, Y = y_2\}, \{X = x_i, Y = y_3\}, \ldots \]

Thus,

\[ P(X = x_i) = P(X = x_i, Y = y_1) + P(X = x_i, Y = y_2) + \ldots = \sum_j p_{X,Y}(x_i, y_j) \]

Hence, when \((X, Y)\) is a discrete random vector we have:

**Definition 6.6** Given \((X, Y)\) with joint probability function \(p(x, y)\), the marginal probability functions of \(X\) and \(Y\) are, respectively

\[ P(X = x) = \sum_y p_{X,Y}(x, y) \quad \text{and} \quad P(Y = y) = \sum_x p_{X,Y}(x, y) \quad (6.8) \]

Analogously, when the random vector, \((X, Y)\), is continuous:

**Definition 6.7** If \((X, Y)\) has joint pdf \(f(x, y)\), the marginal pdfs of \(X\) and \(Y\), denoted by \(f_X(x)\), \(f_Y(y)\) respectively, are

\[ f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, dx. \quad (6.9) \]

**Example 6.9** The joint pdf of \(X\) and \(Y\) is

\[ f(x, y) = \begin{cases} k(3x^2 + xy), & 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 0 & \text{elsewhere} \end{cases} \]

find (i) \(k\) (ii) the marginal pdfs of \(X\) and \(Y\) (iii) \(P(Y < X)\).
6.1. Introduction

Solution

(i) For this to be a pdf, (5.2) must be satisfied. Now
\[ k \int_0^2 \int_0^1 (3x^2 + xy) \, dx \, dy = k \int_0^2 \left[ x^3 + \frac{x^2 y}{2} \right]_0^1 \, dy \]
\[ = k \int_0^2 (1 + \frac{y}{2}) \, dy \]
\[ = k \left[ y + \frac{1}{4} y^2 \right]_0^1 \]
\[ = 3k \]
so \( k \) must be \( \frac{1}{3} \).

(ii) \( f_X(x) = \int_0^2 \left( x^2 + \frac{xy}{3} \right) \, dy = \left[ x^2 y + \frac{xy^2}{6} \right]_{y=0}^{y=1} \)
So \( f_X(x) = 2x^2 + \frac{2x}{3}, 0 \leq x \leq 1. \)

Also \( f_Y(y) = \int_0^1 \left( x^2 + \frac{xy}{3} \right) \, dx = \left[ \frac{1}{3} x^3 + \frac{1}{6} x^2 y \right]_{x=0}^{x=1} \).
So \( f_Y(y) = \frac{1}{6} (2 + y), 0 \leq y \leq 2. \)

(iii) If \( A = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 2 \} \) then
\[ P(Y < X) = \int_{(x,y) \in A} f(x,y) \, dx \, dy \]
\[ = \int_0^1 \int_y^1 (3x^2 + xy) \, dx \, dy \]
\[ = \frac{1}{3} \int_0^1 \left[ x^3 + \frac{1}{2} x^2 y \right]_y^1 \, dy \]
\[ = \frac{1}{3} \int_0^1 \left( 1 + \frac{1}{2} y - \frac{3}{2} y^3 \right) \, dy \]
\[ = \frac{7}{24} \]

Example 6.10 Recall again Example 6.5, where we will now find the marginal
distributions of \( X \) and \( Y \). The probabilities in the first row, for in-
stance, are summed and appear as the first term in the final column
and this is the probability that \( Y = 0 \). Similarly, for the other rows.
Recalling that $X$ is the number of 5’s resulting from two dice being thrown, it is clear that the distribution of $X$ is $\text{bin}(2, \frac{1}{6})$, and the probabilities given in the last row of the table agree with this. That is, 
$$P(X = x) = \binom{2}{x} \left( \frac{1}{6} \right)^x \left( \frac{5}{6} \right)^{2-x}, \quad x = 0, 1, 2.$$ Of course, the distribution of $Y$ is the same.

### Example 6.11
Consider again the experiment in Examples 6.1 and 6.6. From Table 6.3, the marginal distribution for $X_1$ is found simply by summing over the values for $X_2$ in the table. When $x_1 = 0$,
$$p_{X_1}(0) = \sum_{x_2} p_{X_1, X_2}(0, x_2) = 1/24 + 1/24 + 1/24 + \cdots = 6/24.$$ Likewise,
$$p_{X_1}(1) = \sum_{x_2} p_{X_1, X_2}(1, x_2) = 6/12$$
$$p_{X_1}(2) = \sum_{x_2} p_{X_1, X_2}(2, x_2) = 6/24.$$ So the marginal distribution of $X_1$ is
$$p_{X_1}(x_1) = \begin{cases} 
 1/4 & \text{if } x_1 = 0 \\
 1/2 & \text{if } x_1 = 1 \\
 1/4 & \text{if } x_1 = 2 \\
 0 & \text{otherwise} 
\end{cases}$$ Note this is equivalent to adding the row probabilities in Table 6.3. In this example, the marginal distribution is easily found from the total column of Table 6.3.
6.1. Introduction

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>P(Y = y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/36</td>
<td>1/6</td>
<td>1/4</td>
<td>4/9</td>
</tr>
<tr>
<td>1</td>
<td>1/9</td>
<td>1/3</td>
<td>0</td>
<td>4/9</td>
</tr>
<tr>
<td>2</td>
<td>1/9</td>
<td>0</td>
<td>0</td>
<td>1/9</td>
</tr>
</tbody>
</table>

P(X = x) | 1/4  | 1/2  | 1/4 | 1

Figure 6.3: Joint distribution for Example 6.2

6.1.2 Conditional distributions

Consider (X, Y) with joint probability function as in Example 6.2, with marginal distributions of X and Y as shown in Table 6.3.

Suppose we want to evaluate the conditional probability P(X = 1 | Y = 1). We use the fact that P (A | B) = P (A ∩ B) / P (B). So

\[ P(X = 1 | Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{1/3}{4/9} = \frac{3}{4}. \]

So, for each \( x \in R_X \) we could find P(X = x, Y = 1) and this is then the conditional distribution of X given that Y = 1.

**Definition 6.8** For a discrete random vector (X, Y) with probability function \( p_{X,Y}(x, y) \) the conditional probability distribution of X given \( Y = y \) is defined by

\[
p_{X|Y=y}(x | y) = \frac{P(X = x | Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}
\]

for \( x \in R_X \) and provided \( p_Y(y) > 0 \).

Similarly, in the continuous case we have:
Definition 6.9 If \((X, Y)\) is a continuous 2-dimensional random variable with joint pdf \(f_{X,Y}(x, y)\) and respective marginal pdfs \(f_X(x)\), \(f_Y(y)\), then the conditional probability distribution of \(X\) given \(Y = y\) is defined by

\[
f_{X|Y=y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{6.13}
\]

for \(x \in \mathbb{R}_X\) and provided \(f_Y(y) > 0\).

Note that the above conditional pdfs satisfy the requirements for a univariate pdf; that is, \(f_{X|Y}(x \mid y) \geq 0\) for all \(x\) and \(\int_0^\infty f_{X|Y}(x \mid y) \, dx = 1\).

Example 6.12 It was shown in Example 6.9 that if the joint pdf of \(X\) and \(Y\) was

\[
f_{X,Y}(x, y) = \begin{cases} \frac{1}{3}(3x^2 + xy), & 0 < x < 1, 0 < y < 2 \\ 0, & \text{elsewhere,} \end{cases}
\]

then the marginal pdfs of \(X\) and \(Y\) were

\[
f_X(x) = 2x^2 + \frac{2}{3}x, \quad 0, x, 1,
\]

and that

\[
f_Y(y) = \frac{1}{6}(2 + y), \quad 0 < y < 2.
\]

Hence, the conditional distribution of \(X \mid Y = y\) is

\[
f_{X|Y=y}(x \mid y) = \frac{(3x^2 + xy)/3}{(2 + y)/6} = \frac{2x(3x + y)}{2 + y}, 0 < x < 1,
\]

and the conditional distribution of \(Y \mid X = x\) is

\[
f_{Y|X=x}(y \mid x) = \frac{3x + y}{2(3x + 1)}, \quad 0 < y < 2.
\]

It is easy to verify that both these conditional density functions are in fact density functions.

Interpretation of a conditional pdf

To interpret for example, \(f_{X|Y=y}(x \mid y)\), consider slicing through the surface \(f_{X,Y}(x, y)\) with the plane \(y = c\) say, for \(c\) a constant (see Fig 6.4). The intersection of the plane with the surface, will be proportional to a 1-dimensional pdf. This is \(f_{X,Y}(x, c)\), which will not in general be a density function since the area under this curve will be \(f_Y(c)\). Dividing by the constant \(f_Y(c)\) ensures the area under \(\frac{f_{X,Y}(x, c)}{f_Y(c)}\) is one. This is a 1-dimensional pdf, namely that of \(X\) given \(Y = c\), that is \(f_{X|Y=c}(x \mid c)\).
Example 6.13 Consider again the experiment in Examples 6.1 and 6.6. From Table 6.3, the conditional distribution for $X_2$ given $X_1 = 0$ can be found. Note we need to first find $p_{X_1}(x_1)$, which was done in Example 6.11. Then,

$$p_{X_2|X_1=0}(x_2 \mid 0) = \frac{p_{X_1,X_2}(0,x_2)}{p_{X_1}(0)} = \frac{p_{X_1,X_2}(0,x_2)}{1/4},$$

from which we can deduce

$$p_{X_2|X_1=0}(x_2 \mid 0) = \begin{cases}
    \frac{1}{24} = 1/6 & \text{for } x_2 = 1 \\
    \frac{1}{24} = 1/6 & \text{for } x_2 = 2 \\
    \frac{1}{24} = 1/6 & \text{for } x_2 = 3 \\
    \frac{1}{24} = 1/6 & \text{for } x_2 = 4 \\
    \frac{1}{24} = 1/6 & \text{for } x_2 = 5 \\
    \frac{1}{24} = 1/6 & \text{for } x_2 = 6
\end{cases}$$

Note the conditional distribution $p_{X_2|X_1=x_1}(x_2 \mid x_1)$ is a probability function for $X_2$. 

Figure 6.4: $f_{X,Y}(x,y)$ sliced by the plane $Y = c$. 

6.1. Introduction
6.2 Independent random variables

Recall that events $A$ and $B$ are independent if and only if
\[ P(A \cap B) = P(A)P(B) \]

An analogous definition applies for rvs.

**Definition 6.10** The random variables $X$ and $Y$ with joint df $F_{X,Y}$ and marginal df’s $F_X$ and $F_Y$ are independent if and only if
\[ F_{X,Y}(x, y) = F_X(x) \times F_Y(y) \quad (6.14) \]

for all $x$ and $y$.

If $X$ and $Y$ are not independent they are said to be dependent.

The following theorem is often used to establish independence or dependence of rvs. The proof is omitted.

**Theorem 6.11** The discrete random variables $X$ and $Y$ with joint probability function $p_{X,Y}$ and marginals $p_X$ and $p_Y$ are independent if and only if
\[ p_{X,Y}(x, y) = p_X(x) \times p_Y(y) \text{ for every } (x, y) \in R_{X \times Y} \quad (6.15) \]

The continuous random variables $(X, Y)$ with joint pdf $f_{X,Y}$ and marginal pdfs $f_X$ and $f_Y$ are independent if and only if
\[ f_{X,Y}(x, y) = f_X(x) \times f_Y(y) \quad (6.16) \]

for all $x$ and $y$.

Note that to show independence for continuous rvs (and analogously for discrete rvs) we must show $f_{X,Y}(x, y) = f_X(x) \times f_Y(y)$ for all pairs $(x, y)$. If $f_{X,Y}(x, y) \neq f_X(x) \times f_Y(y)$ for any one particular pair of $(x, y)$, then $X$ and $Y$ are dependent.

**Example 6.14** The random variables $X$ and $Y$ have the following joint probability distribution.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1/30</td>
<td>1/30</td>
<td>2/30</td>
<td>1/30</td>
</tr>
<tr>
<td>y</td>
<td>2/30</td>
<td>2/30</td>
<td>4/30</td>
<td>2/30</td>
</tr>
<tr>
<td></td>
<td>3/30</td>
<td>3/30</td>
<td>6/30</td>
<td>3/30</td>
</tr>
</tbody>
</table>
Find the marginal distributions and determine whether or not $X$ and $Y$ are independent. Summing across the rows we obtain the marginal probability function of $Y$. That is,

\[ P(Y = 1) = 1/6, \quad P(Y = 2) = 1/3, \quad P(Y = 3) = 1/2. \]

Summing each column, we obtain the marginal probability function of $X$. That is,

\[ P(X = 1) = 1/5, \quad P(X = 2) = 1/5, \quad P(X = 3) = 2/5, \quad P(X = 4) = 1/5. \]

Clearly (5.16) is satisfied for all pairs $(x, y)$, so $X$ and $Y$ are independent.

**Example 6.15** Given random variables $X$ and $Y$ with joint pdf

\[ f(x, y) = \begin{cases} 4xy & \text{for } 0 < x < 1, \ 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases} \]

show that $X$ and $Y$ are independent.

We need to find the marginal distributions of $X$ and $Y$. Now

\[ f_X(x) = \int_0^1 4xy \, dy = 2x, \quad x \in (0, 1). \]

Similarly $f_Y(y) = 2y, \ y \in (0, 1)$.

Thus we have $f_X(x) f_Y(y) = f(x, y)$ and $X$ and $Y$ are independent.

**Example 6.16** Consider again the experiment in Examples 6.1 and 6.6.

The marginal distribution of $X_1$ was found in Example 6.11. The marginal distribution of $X_2$ is (check!)

\[ p_{X_2}(x_2) = \begin{cases} \frac{1}{24} & \text{for } x_2 = 1 \\ \frac{1}{24} & \text{for } x_2 = 2 \\ \frac{1}{24} & \text{for } x_2 = 3 \\ \frac{1}{24} & \text{for } x_2 = 4 \\ \frac{1}{24} & \text{for } x_2 = 5 \\ \frac{1}{24} & \text{for } x_2 = 6 \end{cases} \]
To determine if $X_1$ and $X_2$ are independent, each $x_1$ and $x_2$ pair must be considered. As an example, we see

\[
p_{X_1}(0) \times p_{X_2}(1) = 1/4 \times 1/6 = 1/24 = p_{X_1,X_2}(0,1)
\]
\[
p_{X_1}(0) \times p_{X_2}(2) = 1/4 \times 1/6 = 1/24 = p_{X_1,X_2}(0,2)
\]
\[
p_{X_1}(1) \times p_{X_2}(1) = 1/2 \times 1/6 = 1/12 = p_{X_1,X_2}(1,1)
\]
\[
p_{X_1}(2) \times p_{X_2}(1) = 1/4 \times 1/6 = 1/24 = p_{X_1,X_2}(2,1)
\]

In fact, this is true for all pairs, and so $X_1$ and $X_2$ are independent random variables. Independence is, however, obvious from the description of the experiment (see Example 6.1). This can be most easily seen from Table 6.3.

---

**Example 6.17** Consider the continuous random variables $X_1$ and $X_2$ with joint pdf

\[
f_{X_1,X_2}(x_1, x_2) = \begin{cases} 
\frac{2}{7}(x_1 + 2x_2) & \text{for } 0 < x_1 < 1, 1 < x_2 < 2 \\
0 & \text{elsewhere}
\end{cases}
\]

The marginal distribution of $X_1$ is

\[
f_X(x_1) = \int_1^2 \frac{2}{7}(x_1 + 2x_2) \, dx_2
\]

\[
= \frac{2}{7}(x_1x_2 + x_2^2) \bigg|_{x_2=1}^{x_2=2} = \frac{2}{7}(x_1 + 3)
\]

for $0 < x_1 < 1$ (and zero elsewhere). Likewise, the marginal distribution of $X_2$ is

\[
f_{X_2}(x_2) = \int_0^1 \frac{2}{7}(x_1 + 2x_2^2) \, dx_1
\]

\[
= \frac{2}{7}(x_1^2/2 + 2x_1x_2) \bigg|_{x_1=0}^{x_1=1} = \frac{1}{7} (1 + 4x_2)
\]

for $1 < x_2 < 2$ (and zero elsewhere). (Note both the marginal distributions must be valid density functions, and so you should check $\int_0^1 f_{X_1}(x_1) \, dx_1 = 1$ and $\int_0^2 f_{X_2}(x_2) \, dx_2 = 1$.)

Since

\[
f_{X_1}(x_1) \times f_{X_2}(x_2) = \frac{2}{49}(x_1 + 3)(1 + 4x_2) \neq f_{X_1,X_2}(x_1, x_2),
\]

the random variables $X_1$ and $X_2$ are not independent.
The conditional distribution of $X_1$ given $X_2 = x_2$ is
\[ f_{X_1|X_2=x_2}(x_1 | x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{(2/7)(x_1 + 2x_2)}{(1/7)(1 + 4x_2)} \]
for $0 < x_1 < 1$ and any given value of $1 < x_2 < 2$. (Since the conditional density must be a valid pdf, you should check \( \int_0^1 \frac{(2/7)(x_1 + 2x_2)}{(1/7)(1 + 4x_2)} \, dx_1 = 1 \). So, for example,
\[ f_{X_1|X_2=1.5}(x_1 | 1.5) = \frac{(2/7)(x_1 + 2 \times 1.5)}{(1/7)(1 + 4 \times 1.5)} = \frac{2}{5}(x_1 + 3) \]
for $0 < x_1 < 1$ and is zero elsewhere. And,
\[ f_{X_1|X_2=1}(x_1 | 1) = \frac{(2/7)(x_1 + 2)}{(1/7)(1 + 4 \times 1)} = \frac{2}{5}(x_1 + 2) \]
for $0 < x_1 < 1$ and is zero elsewhere. Since the distribution of $X_1$ depends on the given value of $X_2$, $X_1$ and $X_2$ are not independent.

---

**Example 6.18** Consider the two continuous random variables $Y_1$ and $Y_2$ with joint pdf
\[ f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} k(y_1 + y_2) & \text{for } 0 < y_1 < y_2 < 1 \\ 0 & \text{elsewhere} \end{cases} \]

(a) Find a value for $k$.
(b) Determine if $Y_1$ and $Y_2$ are independent.

**Solution**

(a) A diagram of the region over which $Y_1$ and $Y_2$ are defined is shown in Figure 6.5. To find $k$, proceed as follows (being careful with the integration limits).
\[ k \int_0^1 \left\{ \int_0^{y_2} y_1 + y_2 \, dy_1 \right\} \, dy_2 = k \int_0^1 y_1^2/2 + y_1y_2 \bigg|_{y_1=0}^{y_2} \, dy_2 = 3k/2 \int_0^1 y_2^2 \, dy_2 = k/2, \]
and so $k = 2$. 

(b) Since $0 < y_1 < 1$ when $y_2 = 1$ but $0 < y_1 < 0.5$ when $y_2 = 0.5$, the values $Y_1$ can take depends on the value of $Y_2$. Hence $Y_1$ and $Y_2$ cannot be independent.

6.3 Expectations involving bivariate distributions

**Reading 6.4** WMS, Sections 5.5 and 5.6.

In a manner analogous to the univariate case, we make the following definition concerning expectation of functions of two rvs.

**Definition 6.12** Let $(X, Y)$ be a 2-dimensional random variable and let $u(X, Y)$ be a function of $X$ and $Y$. Then the expectation or expected value of $E(u(X, Y))$ is

(i) for $(X, Y)$ discrete with probability function $p_{X,Y}(x, y), (x, y) \in R$,

$$E[u(X, Y)] = \sum_{(x, y) \in R} u(x, y)p_{X,Y}(x, y), \quad (6.17)$$

(ii) for $(X, Y)$ continuous with pdf $f_{X,Y}(x, y) > 0, (x, y) \in R$

$$E[u(X, Y)] = \int\int_{R} u(x, y)f_{X,Y}(x, y) \, dx \, dy. \quad (6.18)$$
6.3. Expectations involving bivariate distributions

Although we don’t give it here, this definition can obviously be extended to the expectation of a function of any number of random variables.

Example 6.19 Consider the joint distribution of $X$ and $Y$ in Example 6.6. Determine $E(X + Y)$; ie the mean of the number of heads plus the number showing on the die.

From the Definition 6.12 we have $u(X, Y) = X + Y$ and so

$$E(X + Y) = \sum_{x=0}^{2} \sum_{y=1}^{6} (x + y)p_{X,Y}(x, y)$$

$$= 1 \times (1/24) + 2 \times (1/24) + \ldots + 6 \times (1/24)$$

$$+ 2 \times (1/12) + 3 \times (1/12) + \ldots + 7 \times (1/12)$$

$$+ 3 \times (1/24) + 4 \times (1/24) + \ldots + 8 \times (1/24)$$

$$= 21/24 + 27/12 + 33/24$$

$$= 4.5$$

Notice the answer is just $E(X) + E(Y) = 1 + 3.5 = 4.5$. This is no coincidence as we see from Theorem 6.13 below.

Example 6.20 Consider Example 6.7. Determine $E(XY)$.

Now $u(X, Y) = XY$ and we have

$$E(XY) = \frac{6}{5} \int_{0}^{1} \int_{0}^{1} xy(x + y^2)\, dx\, dy$$

$$= \frac{6}{5} \int_{0}^{1} x^2 y + \frac{x^2 y^3}{2} \bigg|_{x=0}^{x=1} \, dy$$

$$= \frac{6}{5} \int_{0}^{1} y^3 + \frac{y^3}{2} \, dy$$

$$= \frac{6}{5} \left( \frac{y^2}{6} + \frac{y^4}{8} \right) \bigg|_{0}^{1}$$

$$= \frac{6}{5} \left( \frac{1}{6} + \frac{1}{8} \right)$$

$$= \frac{7}{20}$$

For this example, unlike the previous one, we cannot find an alternative simple calculation based on $E(X)$ and $E(Y)$ because $E(XY) \neq E(X)E(Y)$. 
**Theorem 6.13** If $X$ and $Y$ are any rvs and $a$ and $b$ are any constants then

$$E(aX + bY) = aE(X) + bE(Y)$$

This theorem won’t surprise after seeing Theorem 3.3 but it is a very powerful and useful result. The proof given here is for the discrete case but the continuous case is analogous.

**Proof**

$$E(aX + bY) = \sum \sum_{(x,y) \in R} (ax + by)p_{X,Y}(x,y), \text{ by definition}$$

$$= \sum_x \sum_y axp_{X,Y}(x,y) + \sum_x \sum_y byp_{X,Y}(x,y)$$

$$= a \sum_x x \sum_y p_{X,Y}(x,y) + b \sum_y y \sum_x p_{X,Y}(x,y)$$

$$= a \sum_x xp_{X}(x) + b \sum_y yp_{Y}(y)$$

$$= aE(X) + bE(Y) \quad \blacklozenge$$

Note that this result is true whether or not $X$ and $Y$ are independent.

Theorem 6.13 naturally generalises to the expected value of a linear combination of random variables as follows.

**Theorem 6.14** If $X_1, X_2, \ldots, X_n$ are rvs and $a_1, a_2, \ldots, a_n$ are any constants then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

**Proof** The proof follows directly from Theorem 6.13 by induction. \quad \blacklozenge

### 6.3.1 Moments of a bivariate distribution

The idea of a moment in the univariate case naturally extends to the bivariate case. Hence we can define $\mu'_{rs} = E(X^r Y^s)$ or $\mu_{rs} = E((X - \mu_X)^r(Y - \mu_Y)^s)$ as raw and central moments for a bivariate distribution.

The most important of these moments is the covariance.
6.3. Expectations involving bivariate distributions

Reading 6.5 DGS, Section 4.6 (excluding the Cauchy-Schwartz inequality); WMS, Section 5.7.

**Definition 6.15** The covariance of $X$ and $Y$ is defined as

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)$$

for $X, Y$ discrete

or

$$\text{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X - \mu_X)(Y - \mu_Y) f_{XY}(x, y) \, dx \, dy$$

for $X, Y$ continuous

The covariance is a measure of how $X$ and $Y$ vary jointly in the sense that a positive covariance indicates that ‘on average’ $X$ and $Y$ increase (or decrease) together whereas a negative covariance indicates that ‘on average’ as $X$ increases and $Y$ decreases (and vice versa). We say that covariance is a measure of linear dependence.

Covariance is best evaluated from the computational formula:

**Theorem 6.16** For any rvs $X$ and $Y$,

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

**Proof** The proof is a lovely application of Theorems 6.13 and 3.3.

$$\text{cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

$$= E(XY - \mu_XY - \mu_YX + \mu_X\mu_Y)$$

$$= E(XY) - \mu_XE(Y) - \mu_YE(X) + \mu_X\mu_Y$$

$$= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y$$

$$= E(XY) - \mu_X\mu_Y.$$

Notice to compute the covariance, $E(X)$, $E(Y)$, $E(XY)$ need to be computed, and so the joint and marginal distributions of $X$ and $Y$ are needed.

Covariance has units given by the product of the units of $X$ and $Y$. For example, if $X$ is in metres and $Y$ is in seconds then $\text{cov}(XY)$ has the units metre-seconds.

In order to compare the strength of covariation amongst pairs of rvs it is desirable to eliminate the effect of the units. Correlation does this by scaling the covariance in terms of the standard deviations.
**Definition 6.17** The correlation coefficient between the random variables $X$ and $Y$ is denoted by $\text{corr}(X, Y)$ or $\rho_{X,Y}$ and is defined as

$$
\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}
$$

If there is no confusion over which random variables are involved, we shall simply write $\rho$ rather than $\rho_{XY}$.

It can be shown that $-1 \leq \rho \leq 1$.

**Example 6.21** Consider two discrete rvs $X$ and $Y$ with the joint pf given below.

<table>
<thead>
<tr>
<th>$y$</th>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>1/8</td>
<td>1/4</td>
<td>1/8</td>
<td></td>
<td>1/2</td>
</tr>
<tr>
<td>1</td>
<td>1/6</td>
<td>1/12</td>
<td>1/4</td>
<td></td>
<td>1/2</td>
</tr>
<tr>
<td>Total</td>
<td>7/24</td>
<td>1/3</td>
<td>3/8</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

To compute the correlation coefficient, the following steps are required.

(a) $\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$, so var($X$), var($Y$) and are needed;

(b) To find var($X$) and var($Y$), E($X$) and E($X^2$), E($Y$) and E($Y^2$) are needed, so the marginal pfs of $X$ and $Y$ are required.

So first, the marginal pfs are

$$p_X(x) = \sum_{y=-1,1} p_{X,Y}(x, y) = \begin{cases} 
7/24 & \text{for } x = 0 \\
8/24 & \text{for } x = 1 \\
9/24 & \text{for } x = 2 \\
0 & \text{otherwise}
\end{cases}$$

and

$$p_Y(y) = \sum_{x=0}^2 p_{X,Y}(x, y) = \begin{cases} 
1/2 & \text{for } y = -1 \\
1/2 & \text{for } y = 1 \\
0 & \text{otherwise}
\end{cases}$$

So then,

$$E(X) = (7/24 \times 0) + (8/24 \times 1) + (9/24 \times 2) = 26/24$$

$$E(X^2) = (7/24 \times 0^2) + (8/24 \times 1^2) + (9/24 \times 2^2) = 44/24$$

$$E(Y) = (1/2 \times -1) + (1/2 \times 1) = 0$$

$$E(Y^2) = (1/2 \times (-1)^2) + (1/2 \times 1^2) = 1$$
6.3. Expectations involving bivariate distributions

\[ \text{giving } \text{var}(X) = \frac{44}{24} - \left(\frac{26}{24}\right)^2 = 0.6597222 \text{ and } \text{var}(Y) = 1 - 0^2 = 1. \text{ Then,} \]

\[ \mathbb{E}(XY) = \sum_x \sum_y xy p_{X,Y}(x, y) \]
\[ = (0 \times -1 \times 1/8) + (0 \times 1 \times 1/6) + \cdots + (2 \times 1 \times 1/4) \]
\[ = 1/12. \]

Hence,

\[ \text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 1/12 - (26/24 \times 0) = 1/12, \]

and

\[ \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \]
\[ = \frac{1/12}{\sqrt{0.6597222 \times 1}} \]
\[ = 0.1025978, \]

so the correlation coefficient is about 0.10, and therefore there is a small positive linear correlation between \(X\) and \(Y\).

6.3.2 Properties of covariance and correlation

1. The correlation has no units.

2. The covariance has units; if \(X_1\) is measured in kilograms and \(X_2\) in centimetres, then the units of the covariance are kg-cm.

3. If the units of measurements change, the numerical value of the covariance will change, but the numerical value of the correlation will stay the same. (For example, if \(X_1\) is changed from kilograms to grams, the correlation will not change in value, but the covariance will.)

4. The correlation is a number between \(-1\) and 1 (inclusive). When the correlation coefficient (or covariance) is negative, a negative linear relationship is said to exist between the two variables; likewise, when the correlation coefficient (or covariance) is positive, a positive linear relationship is said to exist between the two variables.

5. When the correlation coefficient (or covariance) is zero, no linear dependence is said to exist.
Theorem 6.18 For random variables $X$, $X$ and $Z$, and constants $a$ and $b$

1. $\text{cov}(X,Y) = \text{cov}(Y,X)$
2. $\text{cov}(aX, bY) = ab \text{cov}(X,Y)$
3. $\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \text{cov}(X,Y)$
4. If $X$ and $Y$ are independent, then $E(XY) = E(X)E(Y)$ and hence $\text{cov}(X,Y) = 0$
5. $\text{cov}(X,Y) = 0$ does not imply $X$ and $Y$ are independent, except for the special case of the bivariate normal distribution.

Be aware that a zero correlation coefficient in an indication of no linear dependence only.

Example 6.22 Consider $X_1$ with the pf

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$-1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{X_1}(x_1)$</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Define $X_2$ to be explicitly related to $X_1$ such that $X_2 = X_1^2$ (that is, we know there is a relationship between $X_1$ and $X_2$, but it is not linear). The joint pf for $(X_1, X_2)$ is

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$X_1$</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
</tr>
<tr>
<td>1</td>
<td>1/3</td>
<td>0</td>
<td>1/3</td>
<td>2/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

Then

$$\text{cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = 0 - 0 \times 2/3 = 0$$

so $\text{corr}(X_1, X_2) = 0$. But $X_1$ and $X_2$ are certainly related as $X_2$ was explicitly defined as a function of $X_1$. Since the correlation is a measure of the strength of the linear relationship between two random variables, a correlation of zero simply is indication of no linear relationship between $X_1$ and $X_2$. (As is the case in this example, there may be a different relationship between the variables, but no linear relationship.)
# 6.4 Conditional expectations

**Reading 6.6** DGS, Section 4.7; WMS, Section 5.11.

Conditional expectations are simply expectations computed from a conditional distribution.

### 6.4.1 Conditional mean

The conditional mean is the expected value computed from a conditional distribution.

**Definition 6.19** The conditional expected value or conditional mean of a random variable $X$ for given $Y = y$ is denoted by $E(X \mid Y = y)$ and is defined as

$$ E(X \mid Y = y) = \begin{cases} \sum_x x p_{X \mid Y}(x \mid y) & \text{if } p_{X \mid Y}(x \mid y) \text{ is the conditional pmf} \\ \int_{-\infty}^{\infty} x f_{X \mid Y}(x \mid y) \, dx & \text{if } f_{X \mid Y}(x \mid y) \text{ is the conditional pdf} \end{cases} $$

$E(X \mid Y = y)$ is typically denoted $\mu_{X \mid Y = y}$.

**Example 6.23** Consider the two rvs $X$ and $Y$ with joint pdf

$$ f_{X,Y}(x, y) = \begin{cases} \frac{3}{5} (x + xy + y^2) & \text{for } 0 < x < 1 \text{ and } -1 < y < 1 \\ 0 & \text{otherwise} \end{cases} $$

To find $f_{Y \mid X = x}(y \mid x)$, we first need $f_X(x)$.

$$ f_X(x) = \int_{-1}^{1} f_{X,Y}(x, y) \, dy = \frac{3}{15} (6x + 2) $$

for $0 < x < 1$. Then,

$$ f_{Y \mid X = x}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{(3/5)(x + xy + y^2)}{(3/15)(6x + 2)} = \frac{3(x + xy + y^2)}{6x + 2} $$
for $-1 < y < 1$ and given $0 < x < 1$. The expected value of $Y$ given $X = x$ is then

$$E(Y | X = x) = \int_{-1}^{1} y f_{Y|X=x}(y | x) dy$$

$$= \int_{-1}^{1} y \frac{3(x + xy + y^2)}{6x + 2} dy$$

$$= \frac{3}{6x + 2} \int_{-1}^{1} y(x + xy + y^2) dy$$

$$= \frac{3}{6x + 2} \left( \frac{x}{3x + 1} \right).$$

This expression indicates that the conditional expected value of $Y$ depends on the given value of $X$; for example,

$$E(Y | X = 0) = 0/1 = 0$$

$$E(Y | X = 0.5) = \frac{0.5}{3 \times 0.5 + 1} = 0.2$$

$$E(Y | X = 1) = 1/4$$

Since $E(Y | X = x)$ depends on the value of $X$, $X$ and $Y$ are not independent.

### 6.4.2 Conditional variance

The conditional variance is the variance computed from a conditional distribution.

**Definition 6.20** The conditional variance of a random variable $X$ for given $Y = y$ is denoted by $\text{var}(X | Y = y)$ and is defined as

$$\text{var}(X | Y = y) = \begin{cases} \sum_{x=-\infty}^{x=\infty} (x - \mu_{X|y})^2 p(x | y) & \text{if } p(x | y) \text{ is the conditional } \text{pf} \\ \int_{-\infty}^{x=\infty} (x - \mu_{X|y})^2 f(x | y) dx & \text{if } f(x | y) \text{ is the conditional } \text{pdf} \end{cases}$$

where $\mu_{X|y}$ is the conditional mean of $X$ given $Y = y$.

$\text{var}(X | Y = y)$ is typically denoted $\sigma_{X|Y=y}^2$. 
Example 6.24 Refer to Example 6.23. The conditional variance of $Y$ given $X = x$ can be found by first computing $E(Y^2 \mid X = x)$.

$$E(Y^2 \mid X = x) = \int_{-1}^{1} y^2 f_{Y \mid X = x}(y \mid x) \, dy$$

$$= \frac{3}{6x + 2} \int_{-1}^{1} y^2 (x + xy + y^2) \, dy$$

$$= \frac{3}{6x + 2} \times \frac{10x + 6}{15}$$

$$= \frac{5x + 3}{5(3x + 1)}.$$  

So the conditional variance is

$$\text{var}(Y \mid X = x) = E(Y^2 \mid X = x) - (E(Y \mid X = x))^2$$

$$= \frac{5x + 3}{5(3x + 1)} - \left( \frac{x}{3x + 1} \right)^2$$

$$= \frac{10x^2 + 12x + 3}{5(3x + 1)^2}$$

for given $0 < x < 1$. Hence the variance of $Y$ depends on the value of $X$ that is given; for example,

$$\text{var}(Y \mid X = 0) = \frac{3}{5} = 0.6$$

$$\text{var}(Y \mid X = 0.5) = \frac{10 \times (0.5)^2 + (12 \times 0.5) + 3}{5(3 \times 0.5 + 1)^2} \approx 0.368$$

$$\text{var}(Y \mid X = 1) = \frac{25}{80} = 0.3125$$

In general, to compute the conditional variance of $X \mid Y = y$ given a joint probability function, the following steps are required.

1. Find the marginal distribution of $Y$.
2. Use this to compute the conditional probability function $f_{X \mid Y = y}(x \mid y) = f_{X,Y}(x, y) / f_X(x)$.
3. Find the conditional mean $E(X \mid Y = y)$.
4. Find the conditional second raw moment $E(X^2 \mid Y = y)$.
5. Finally, compute $\text{var}(X \mid Y = y) = E(X^2 \mid Y = y) - (E(X \mid Y = y))^2$. 
Example 6.25 Two discrete random variables $U$ and $V$ have the joint pf given below.

<table>
<thead>
<tr>
<th>$v$</th>
<th>0</th>
<th>1/9</th>
<th>1/18</th>
<th>1/6</th>
<th>1/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4/9</td>
<td>7/18</td>
<td>1/6</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Find the conditional variance of $V$ given $U = 11$.

Solution Using the steps outlined above:

1. First, find the marginal distribution of $U$. From the joint pf table,

   $$ p_U(u) = \begin{cases} 
   4/9 & \text{for } u = 10 \\
   7/18 & \text{for } u = 11 \\
   1/6 & \text{for } u = 12 \\
   0 & \text{otherwise} 
   \end{cases} $$

2. Secondly, compute the conditional probability function

   $$ p_{V|U=11}(v \mid u = 11) = \frac{p_{U,V}(u,v)}{p_U(u = 11)} $$

   $$ = \begin{cases} 
   1/18 & \text{if } v = 0 \\
   1/3 & \text{if } v = 1 \\
   7/18 & \text{if } v = 0 \\
   6/7 & \text{if } v = 1 
   \end{cases} $$

   using $p_U(u = 11) = 7/18$ from the step 1.

3. Thirdly, find the conditional mean

   $$ E(V \mid U = 11) = \sum_v v p_{V|U=11}(v \mid u) = \left( \frac{1}{7} \times 0 \right) + \left( \frac{6}{7} \times 1 \right) = 6/7 $$

4. Fourthly, find the conditional second raw moment

   $$ E(V^2 \mid U = 11) = \sum_v v^2 p_{V|U=11}(v \mid u) = \left( \frac{1}{7} \times 0^2 \right) + \left( \frac{6}{7} \times 1^2 \right) = 6/7 $$

5. Finally, compute

   $$ \text{var}(V \mid U = 11) = E(V \mid U = 11) - (E(V \mid U = 11))^2 $$

   $$ = (6/7) - (6/7)^2 $$

   $$ \approx 0.1224 $$
6.5 The multivariate extension

Results involving expectations naturally generalise from the bivariate to the multivariate case.

We have already seen the expectation of a linear combination of rvs in Theorem 6.14. The variance of a linear combination of rvs is given in the following theorem.

**Theorem 6.21** If $X_1, X_2, \ldots, X_n$ are rvs and $a_1, a_2, \ldots, a_n$ are any constants then

$$\text{var} \left( \sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + 2 \sum_{i<j} a_i a_j \text{cov}(X_i, X_j)$$

**Proof** For convenience, put $Y = \sum_{i=1}^{n} a_i X_i$. The by definition of variance

$$\text{var}(Y) = E(Y - E(Y))^2$$

$$= E[a_1 X_1 + \cdots + a_n X_n - a_1 \mu_1 - \cdots - a_n \mu_n]^2$$

$$= E[a_1 (X_1 - \mu_1) + \cdots + a_n (X_n - \mu_n)]^2$$

$$= E \left[ \sum_{i} a_i^2 (X_i - \mu_i)^2 + 2 \sum_{i<j} a_i a_j (X_i - \mu_i)(X_j - \mu_j) \right]$$

$$= \sum_{i} a_i^2 E(X_i - \mu_i)^2 + 2 \sum_{i<j} a_i a_j E(X_i - \mu_i)(X_j - \mu_j) \quad \text{using Theorem 6.14}$$

$$= \sum_{i} a_i^2 \sigma_i^2 + 2 \sum_{i<j} a_i a_j \text{cov}(X_i, X_j).$$

In statistical theory, an important special case of Theorem 6.21 occurs when the $X_i$ are independently and identically distributed. That is, each of $X_1, X_2, \ldots, X_n$ has got the same distribution and are independent of each other. (We see the relevance of this in Chapter 8.) Because of its importance this special case is called a corollary of Theorems 6.14 and 6.21.

**Corollary 6.22** If $X_1, X_2, \ldots, X_n$ are independently distributed rvs, each with mean $\mu$ and variance $\sigma^2$, and $a_1, a_2, \ldots, a_n$ are any constants, then

$$E \left( \sum_{i=1}^{n} a_i X_i \right) = \mu \sum_{i=1}^{n} a_i$$

$$\text{var} \left( \sum_{i=1}^{n} a_i X_i \right) = \sigma^2 \sum_{i=1}^{n} a_i^2$$
Module 6. Bivariate distributions

\section*{6.5.1 Vector formulation}

Linear combinations of rvs are most elegantly dealt with using the methods and notation of vectors and matrices.

In the bivariate case we define

\[ X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \]  
\hspace{1cm} (6.19)

\[ E(X) = E\left( \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mu \]  
\hspace{1cm} (6.20)

\[ \text{var}(X) = \text{var}\left( \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \Sigma \]  
\hspace{1cm} (6.21)

The matrix $\Sigma$ is called the \textit{variance-covariance} matrix. Notice that this matrix is square and symmetric since $\sigma_{12} = \sigma_{21}$.

The linear combination $Y = a_1 X_1 + a_2 X_2$ can be expressed

\[ Y = a_1 X_1 + a_2 X_2 = [a_1, a_2] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = a'X \]  
\hspace{1cm} (6.22)

where the (column) vector $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$.

With the standard rules of matrix multiplication (see eg, Appendix 1 of WMS), Theorems 6.14 and 6.21 applied to 6.22 then give respectively (check the details for yourself)

\[ E(Y) = E(a'X) = [a_1, a_2] \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = a'\mu \]  
\hspace{1cm} (6.23)

and

\[ \text{var}(Y) = \text{var}(a'X) \]
\[ = [a_1, a_2] \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \]
\[ = a'\Sigma a \]  
\hspace{1cm} (6.24)

The vector formulation of these results apply directly in the multivariate case as described below.
Write
\[ X = (X_1, X_2, \ldots, X_n)' \]
\[ \text{E}(X) = (\mu_1, \ldots, \mu_n)' = \mu' \]
\[ \text{var}(X) = \Sigma \]
\[ a' = [a_1, a_2, \ldots, a_n] \]

Now Theorems 6.14 and 6.21 re-expressed in vector form become:

**Theorem 6.23** If \( X \) is a random vector of length \( n \) with mean \( \mu \) and variance \( \Sigma \) and \( a \) is any constant vector of length \( n \) then
\[ \text{E}(a'X) = a'\mu \]
and
\[ \text{var}(a'X) = a'\Sigma a \]

**Proof** Exercise!  

These elegant statements concerning linear combinations are a feature of vector formulations that extend to many statistical results in the theory of statistics.

One obvious advantage of this formulation is the implementation in vector-based computer programming used by packages such as MATLAB and R.

As a further example we finish this module by providing without proof (although the proof is relatively straightforward) one further result, this time involving two linear combinations.

**Theorem 6.24** If \( X \) is a random vector of length \( n \) with mean \( \mu \) and variance \( \Sigma \) and \( a \) and \( b \) are any constant vectors, each of length \( n \), then
\[ \text{cov}(a'X, b'X) = a'\Sigma a \]

**Example 6.26** Suppose the random variables \( X_1, X_2, X_3 \) have respective means 1, 2, and 3, respective variances 4, 5, and 6, and covariances \( \text{cov}(X_1, X_2) = -1, \text{cov}(X_1, X_3) = 1 \) and \( \text{cov}(X_2, X_3) = 0 \). Consider the rvs \( Y_1 = 3X_1 + 2X_2 - X_3 \) and \( Y_2 = X_3 - X_1 \). Determine \( \text{E}(Y_1), \text{E}(Y_2), \text{var}(Y_1), \text{var}(Y_2) \) and \( \text{cov}(Y_1, Y_2) \)
A vector formulation of this problem allows us to use Theorems 6.23 and 6.24 directly. Putting \( a' = (3, 2, -1)' \) and \( b' = (-1, 0, 1)' \) we have

\[
Y_1 = a'X \quad \text{and} \quad Y_2 = b'X
\]

where \( X' = (X_1, X_2, X_3)' \).

We also define \( \mu' = (1, 2, 3)' \) and \( \Sigma = \begin{bmatrix}
4 & -1 & 1 \\
-1 & 5 & 0 \\
1 & 0 & 6
\end{bmatrix} \) as the mean and variance-covariance matrix respectively of \( X \).

Then

\[
E(Y_1) = a'\mu = (3, 2, -1)' \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = 4
\]

and

\[
\text{var}(Y_1) = a'\Sigma a = (3, 2, -1)' \begin{bmatrix}
4 & -1 & 1 \\
-1 & 5 & 0 \\
1 & 0 & 6
\end{bmatrix} \begin{bmatrix}
3 \\ -1 \\ 0
\end{bmatrix} = 44
\]

Similarly \( E(Y_2) = 2 \) and \( \text{var}(Y_2) = 8 \).

Finally we have

\[
\text{cov}(Y_1, Y_2) = a'\Sigma b = (3, 2, -1)' \begin{bmatrix}
4 & -1 & 1 \\
-1 & 5 & 0 \\
1 & 0 & 6
\end{bmatrix} \begin{bmatrix}
-1 \\ 0 \\ 1
\end{bmatrix} = -12
\]

### 6.6 Multinomial distribution

**Reading 6.7** DGS, Section 5.11; WMS, Section 5.9.

The multinomial distribution is a generalization of the binomial distribution and is an example of a discrete multivariate distribution.

**Definition 6.25** Consider an experiment with the sample space partitioned as \( S = \{B_1, B_2, \ldots, B_k\} \). Let \( p_i = P(B_i), \ i = 1, 2, \ldots, k \) where \( \sum_{i=1}^k p_i = 1 \). Suppose there are \( n \) repetitions of the experiment in which \( p_i \) is constant. Let the random variable \( X_i \) be the number of times (in the \( n \) repetitions) that the event \( B_i \) occurs. In this situation, the random vector...
(X_1, X_2, \ldots, X_k) is said to have a multinomial distribution with probability function

\[ P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}, \quad (6.25) \]

where \( R_X = \{(x_1, \ldots, x_k) : x_i = 0, 1, \ldots, n, i = 1, 2, \ldots, k, \sum_{i=1}^k x_i = n\} \).

The part of (6.25) involving factorials arises as the number of ways of arranging \( n \) objects, \( x_1 \) of which are of the first kind, \( x_2 \) of which are of the second kind, etc. The above distribution is really \((k - 1)\)-variate since \( x_k = n - \sum_{i=1}^{k-1} x_i \). In particular if \( k = 2 \), the multinomial distribution reduces to the binomial distribution which is a univariate distribution.

If we consider \( X_i \), it is the number of times (out of \( n \)) that the event \( B_i \), which has probability \( p_i \), occurs. So the random variable \( X_i \) clearly has a binomial distribution with parameters \( n, p_i \). We see then that the marginal probability distribution of one of the components of a multinomial distribution is a binomial distribution.

Notice that the distribution in Example 6.5 is an example of a trinomial distribution. The probabilities shown in Table 6.2 can be expressed algebraically as

\[ P(X = x, Y = y) = \frac{2!}{x!y!(2 - x - y)!} \left( \frac{1}{6} \right)^x \left( \frac{1}{6} \right)^y \left( \frac{2}{3} \right)^{2-x-y} \]

for \( x, y = 0, 1, 2; x + y \leq 2 \).

The following are the basic properties of the multinomial distribution.

**Theorem 6.26** Suppose \((X_1, X_2, \ldots, X_k)\) has the multinomial distribution given in Definition 6.25. Then for \( i = 1, 2, \ldots, k \)

1. \( E(X_i) = np_i \)
2. \( \text{var}(X_i) = np_i(1 - p_i) \)
3. \( \text{cov}(X_i, X_j) = -np_ip_j \) for \( i \neq j \)

**Proof** We will use \( x \) for \( x_1 \) and \( y \) for \( x_2 \) in 3. for convenience.

1. & 2. follow from the fact that \( X_i \sim \text{bin}(n, p_i) \).
2. Consider only the case \( k = 3 \), and note that 
\[
\sum_{(x,y) \in \mathbb{R}} \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y (1 - p_1 - p_2)^{n-x-y} = 1
\]
Then, putting \( p_3 = 1 - p_1 - p_2 \),
\[
E(XY) = \sum_{(x,y)} xy P(X = x, Y = y) = \sum_{(x,y)} \frac{n!}{(x-1)!(y-1)!(n-x-y)!} p_1^x p_2^y p_3^{n-x-y}
\]
\[
= n(n-1)p_1p_2 \sum_{(x,y)} \frac{(n-2)!}{(x-1)!(y-1)!(n-x-y)!} p_1^{x-1} p_2^{y-1} p_3^{n-x-y}
\]
So \( \text{cov}(X, Y) = n^2 p_1 p_2 - np_1 p_2 - (np_1)(np_2) = -np_1 p_2 \).

Example 6.27 Suppose that the four basic blood groups O, A, B and AB are known to occur in the following proportions 9 : 8 : 2 : 1. Given a random sample of 8 individuals, what is the probability that there will be 3 each of types O and A and 1 each of types B and AB? 
The probabilities are \( p_1 = .45, p_2 = .4, p_3 = .1, p_4 = .05 \), and
\[
P(X_O = 3, X_A = 3, X_B = 1, X_{AB} = 1) = \frac{8!}{3!3!1!1!}(.45)^3(.4)^3(.1)(.05)
\]
\[=.033 \]
The R code for this uses the function \texttt{dmultinom}.

\[
> \text{dmultinom(size=8,x=c(3,3,1,1),p=c(0.45,0.4,0.1,0.05))}
\]
\[0.033 \]
The above problem could be simulated in R by
\[
\text{n <- 1000}
\text{simulated.sample <- sample(c("O","A","B","AB"),prob=c(0.45,0.4,0.1,0.05),replace=T,size=n)}
\text{table(simulated.sample)}
\text{simulated.sample}
A  AB  B  O
416 45 102 437
6.7 The bivariate normal distribution

Reading 6.8 DGS, Section 5.12; WMS, Section 5.10.

Definition 6.27 If a pair of rvs $X$ and $Y$ have the joint pdf

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X \sigma_Y \sqrt{1-\rho^2}} \exp(-Q/2)$$

where

$$Q = \frac{1}{1-\rho^2} \left[ \left( \frac{x - \mu_X}{\sigma_X} \right)^2 - 2\rho \left( \frac{x - \mu_X}{\sigma_X} \right) \left( \frac{y - \mu_Y}{\sigma_Y} \right) + \left( \frac{y - \mu_Y}{\sigma_Y} \right)^2 \right],$$

then $X$ and $Y$ have a bivariate normal distribution. We write

$$(X, Y) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho).$$

For notational convenience with the bivariate normal distribution we often use $X_1, X_2$ instead of $X, Y$.

A typical graph of the bivariate normal surface above the $x$–$y$ plane is shown below.

![Figure 6.6: The bivariate normal density function.](image)

It can be shown that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dx \, dy = 1.$$

Some important facts about the bivariate normal distribution are contained in the theorem below.
Theorem 6.28 For \((X, Y)\) with pdf given in (6.26),

(a) the marginal distributions of \(X\) and of \(Y\) are \(N(\mu_X, \sigma_X^2)\) and \(N(\mu_Y, \sigma_Y^2)\) respectively

(b) the parameter \(\rho\) appearing in (6.26) is the correlation coefficient between \(X\) and \(Y\)

(c) the conditional distributions of \(X\) given \(Y = y\) and of \(Y\) given \(X = x\) are respectively

\[
N\left[ \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y), \sigma_X^2 (1 - \rho^2) \right], \quad N\left[ \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X), \sigma_Y^2 (1 - \rho^2) \right].
\]

Proof

(a) Recall that the marginal pdf of \(X\) is \(f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy\). In the integral, put \(u = (x - \mu_X)/\sigma_X, \, v = (y - \mu_Y)/\sigma_Y, \, dy = \sigma_Y \, dv\) and complete the square (in the exponent) on \(v\).

\[
g(x) = \frac{1}{2\pi \sigma_X \sqrt{1 - \rho^2} \sigma_Y} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[ u^2 - 2\rho uv + v^2 \right] \right\} \sigma_Y \, dv
\]

\[
= \frac{1}{2\pi \sigma_X \sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[ (v - \rho u)^2 + u^2 - \rho^2 u^2 \right] \right\} \, dv
\]

\[
= \frac{e^{-u^2/2}}{\sqrt{2\pi \sigma_X}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi (1 - \rho^2)}} \exp \left\{ -\frac{1}{2(1 - \rho^2)} (v - \rho u)^2 \right\} \, dv
\]

Replacing \(u\) by \((x - \mu_X)/\sigma_X\), we see from the pdf that \(X \sim N(\mu_X, \sigma_X^2)\). Similarly for the marginal pdf of \(Y\), \(f_Y(y)\).

(b) To show that \(\rho\) in (6.26) is actually the correlation coefficient of \(X\) and \(Y\), recall that

\[
\rho_{X,Y} = \text{cov}(X,Y)/\sigma_X \sigma_Y = \frac{\text{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{(x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} f(x,y) \, dx \, dy
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} uv \frac{1}{2\pi \sqrt{1 - \rho^2} \sigma_X \sigma_Y} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[ u^2 - 2\rho uv + v^2 \right] \right\} \sigma_X \sigma_Y \, du \, dv.
\]
6.7. The bivariate normal distribution

The exponent is

\[ -\frac{[(u - \rho v)^2 + v^2 - \rho^2 v^2]}{2(1 - \rho^2)} = -\frac{1}{2} \left\{ \frac{(u - \rho v)^2}{(1 - \rho^2)} + v^2 \right\} \]

\[ \rho_{X,Y} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{ue^{-u^2/2}}{\sqrt{2\pi}} \frac{v}{\sqrt{2\pi(1 - \rho^2)}} \exp\{-(u - \rho v)^2/2(1 - \rho^2)\} \, du \, dv \]

\[ = \text{E}(U) \quad \text{where} \quad u \sim N(\rho v, 1 - \rho^2) \]

\[ = \rho \int_{-\infty}^{\infty} \frac{v^2}{\sqrt{2\pi}} e^{-v^2/2} \, dv \]

\[ = \rho \quad \text{since the integral is} \quad \text{E}(V^2) \quad \text{where} \quad V \sim N(0,1). \]

(c) In finding the conditional pdf of \( X \) given \( Y = y \), we use

\[ f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \]

Then in this ratio, the constant is

\[ \frac{\sqrt{2\pi} \sigma_Y}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} = \frac{1}{\sqrt{2\pi} \sigma_X \sqrt{1 - \rho^2}} \]

The exponent is

\[ \exp \left\{ -\frac{\left( x - \mu_X \right)^2}{\sigma_X^2} - \frac{2\rho (x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} \right\} / 2(1 - \rho^2) \]

\[ = \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[ \frac{(x - \mu_X)^2}{\sigma_X^2} - \frac{2\rho (x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} \right] (1 - 1 + \rho^2) \right\} \]

\[ = \exp \left\{ -\frac{1}{2\sigma_X^2(1 - \rho^2)} \left[ (x - \mu_X)^2 - 2\rho \frac{\sigma_X}{\sigma_Y} (x - \mu_X)(y - \mu_Y) + \frac{\rho^2 \sigma_X^2}{\sigma_Y^2} (y - \mu_Y)^2 \right] \right\} \]

\[ = \exp \left\{ -\frac{1}{2(1 - \rho^2)\sigma_X^2} \left[ x - \mu_X - \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y) \right]^2 \right\}. \]

So the conditional distribution of \( X \) given \( Y = y \) is

\[ N \left( \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y), \sigma_X^2(1 - \rho^2) \right). \]

Recall the interpretation of the conditional distribution of \( X \) given \( Y = y \) (Section 6.1.2) and note the shape of this density in Figure 6.6.
A couple of comments are worth noting about Theorem 6.28.

1. From (a) and (c) we have $E(X) = \mu_X$ and $E(X \mid Y = y) = \mu_X + \rho \sigma_X (y - \mu_Y) / \sigma_Y$ (and similarly for $Y$). Notice that $E(X \mid Y = y)$ is a linear function of $y$; i.e., if $(X, Y)$ is bivariate normal, the regression line of $Y$ on $X$ (and $X$ on $Y$) is linear.

2. An important result follows from (b). If $X$ and $Y$ are uncorrelated (i.e., if $\rho = 0$) then $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ and thus $X$ and $Y$ are independent. That is, if two normally distributed random variables are uncorrelated, they are also independent.

A more succinct way of writing the expressions for the bivariate normal density function is to use the matrix representation,

Write $\sigma_{12} = \sigma_{21} = \rho \sigma_1 \sigma_2$ and

\[
\begin{pmatrix} X_1 \\
X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\
X_2 \end{pmatrix},
\]

\[
E(\mathbf{X}) = E\left( \begin{pmatrix} X_1 \\
X_2 \end{pmatrix} \right) = \begin{pmatrix} \mu_1 \\
\mu_2 \end{pmatrix} = \mu,
\]

\[
\text{var}(\mathbf{X}) = \text{var}\left( \begin{pmatrix} X_1 \\
X_2 \end{pmatrix} \right) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\
\sigma_{21} & \sigma_2^2 \end{pmatrix} = \Sigma
\]

Then,

\[
f(\mathbf{x}; \mathbf{\mu}, \Sigma) = (2\pi)^{-1} |\Sigma|^{-1} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{\mu})^T \Sigma^{-1} (\mathbf{x} - \mathbf{\mu}) \right\}
\]

\[
\begin{pmatrix} x_1 \\
x_2 \end{pmatrix} \sim N\left( \begin{pmatrix} \mu_1 \\
\mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)
\]

\[
x_1 \sim N(\mu_1, \Sigma_{11})
\]

\[
x_2 \sim N(\mu_2, \Sigma_{22})
\]

\[
(x_2 \mid x_1) \sim N(\mu_{2,1}, \Sigma_{22,1})
\]

where \( \mu_{2,1} = \mu_2 - \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1) \)

\( \Sigma_{22,1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \)
6.7. The bivariate normal distribution

Example 6.28 Marsh [20] gives data from 200 married men and their wives from the OPCS study of heights and weights of adults in Great Britain in 1980. Histograms of the husbands’ and wives’ heights are given in Figure 6.7; the marginal distributions are approximately normal. The scatterplot of the heights is shown in Figure 6.8.

From the histograms, there is reason to suspect that a bivariate normal distribution would be appropriate.

Using $H$ to refer to heights of husbands and $W$ to the heights of wives, the sample statistics are:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Husbands</th>
<th>Wives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean:</td>
<td>1732</td>
<td>1602</td>
</tr>
<tr>
<td>Sample std dev:</td>
<td>68.8</td>
<td>62.4</td>
</tr>
<tr>
<td>Correlation:</td>
<td>+0.364</td>
<td></td>
</tr>
</tbody>
</table>

Note that $\rho$ is positive; this implies taller men marry taller women on average.

Using this sample information, the bivariate normal distribution can be computed. This 3-dimensional density function can be difficult to plot on a two-dimensional page; but see Figure 6.9.

The pdf for the bivariate normal distribution for the heights of the husbands and wives could be written down in the form of Equation (6.26) for the values of $\mu_H$, $\mu_W$, $\sigma_H^2$, $\sigma_W^2$ and $\rho$ above; but this is tedious.

Given the information, what is the probability that a randomly chosen man in the UK in 1980 who is 173 centimetres tall had married a
Module 6. Bivariate distributions

Figure 6.8: The scatterplot of husbands’ heights and wives’ heights.

woman taller than himself?

Solution  The information implies that $H = 1730$ is given (remembering the data are given in millimetres). So we need the conditional distribution of $W \mid H = 1730$. Using the results above, this conditional distribution will have mean

$$ b = \mu_W + \rho \frac{\sigma_W}{\sigma_H} (y_H - \mu_H)$$

$$= 1602 + 0.364 \frac{62.4}{68.8} (1730 - 1732)$$

$$= 1601.34$$

and variance

$$\sigma^2 (1 - \rho^2) = 62.4^2 (1 - 0.364^2)$$

$$= 3377.85.$$  

In summary, $W \mid (H = 1730) \sim N(1601.34, 3377.85)$. Note that this conditional distribution has a univariate normal distribution, and so probabilities such as $W > 1730$ are easily determined. Then,

$$P (W > 1730 \mid H = 1730) = P \left( Z > \frac{1730 - 1601.34}{\sqrt{3377.85}} \right)$$

$$= P (Z > 2.2137)$$

$$= 0.013$$

ie approximately 1% of males 173cm tall had married women taller than themselves in the UK in 1980.
6.8 Self-assessment exercises

Ex. 6.1 Discrete rvs $X$ and $Y$ have the joint pf shown below.

<table>
<thead>
<tr>
<th>$y$</th>
<th>$0$</th>
<th>$1$</th>
<th>$2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>0/12</td>
<td>1/6</td>
<td>1/24</td>
</tr>
<tr>
<td>$1$</td>
<td>1/4</td>
<td>1/4</td>
<td>1/40</td>
</tr>
<tr>
<td>$2$</td>
<td>1/8</td>
<td>1/20</td>
<td>0</td>
</tr>
<tr>
<td>$3$</td>
<td>1/120</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Determine

(a) $P (X = 1, Y = 2)$
(b) $P (X + Y \leq 1)$
(c) $P (X > Y)$
(d) the marginal pf of $X$
(e) the pf of $Y \mid X = 1$.

Ex. 6.2 The rv A has mean 3 and variance 2. The rv B has mean 4 and variance 3. Assume A and B are independent. Find

(a) $E(A + B)$
(b) $\text{var}(A + B)$
(c) $E(2A - 3B)$
(d) $\text{var}(2A - 3B)$
Ex. 6.3 (Computer exercise) Estimate the probabilities in Example 6.5 using simulation.

Ex. 6.4 Suppose $X$ and $Y$ are independently uniformly distributed on $[-1, 1]$.

(a) Verify geometrically that $P (X^2 + Y^2 \leq 1) = \pi/4$.
(b) Hence use R to estimate the value of $\pi$.

6.9 Exercises

Ex. 6.5 $(X, Y)$ has joint probability function given by

$$P (X = x, Y = y) = k|x - y|, \text{ for } x = 0, 1, 2; y = 1, 2, 3$$

(a) Find $k$.
(b) Construct a table of probabilities for this distribution.
(c) Find $P (X \leq 1, Y = 3)$
(d) Find $P (X + Y \geq 3)$

Ex. 6.6 (a) For what value of $k$ is $f(x, y) = kxy$, $0 \leq x \leq 1, 0 \leq y \leq 1$, a joint pdf?
(b) Find $P (X \leq x_0, Y \leq y_0)$.
(c) Hence evaluate $P (X \leq \frac{3}{8}, Y \leq \frac{5}{8})$.

Ex. 6.7 Consider the discrete distribution defined by the table of probabilities

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{16}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{16}$</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{16}$</td>
<td>$\frac{2}{16}$</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(a) Find the marginal distributions.
(b) Find $P (X = 2 \text{ or } Y = 3)$.
(c) Find $P (X = 2 \mid Y = 3)$.
(d) Find $P (X \leq 3 \mid Y < 4)$
(e) Find $P (Y = 3 \mid (X - Y)^2 = 1)$
(f) Show that $X$ and $Y$ are not independent.
Ex. 6.8 The pdf of \((X, Y)\) is given by
\[
f_{X,Y}(x, y) = 1 - \alpha(1 - 2x)(1 - 2y), \quad 0 \leq x \leq 1; \quad 0 \leq y \leq 1
\]
where \(-1 \leq \alpha \leq 1\).

(a) Find the marginal distributions of \(X\) and \(Y\).
(b) For what value of \(\alpha\) are \(X\) and \(Y\) independent?
(c) Evaluate the correlation coefficient \(\rho_{XY}\).
(d) Find \(P(X < Y)\).

Ex. 6.9 For the random vector \((X, Y)\) the conditional pdf of \(Y\) given \(X = x\) is
\[
f_Y(y \mid x) = \frac{2(x + y)}{2x + 1}, \quad 0 < y < 1.
\]
The marginal pdf of \(X\) is given by
\[
g_X(x) = x + \frac{1}{2}, \quad 0 < x < 1.
\]
(a) Find \(F_Y(y \mid x)\) and hence evaluate \(P(Y < 3/4 \mid X = 1/3)\).
(b) Find the joint pdf, \(f_{X,Y}(x, y)\), of \(X\) and \(Y\).
(c) Find \(P(Y < X)\)

Ex. 6.10 If \((X, Y)\) has pdf given by \(f_{X,Y}(x, y) = 4xy, 0 \leq x \leq 1, 0 \leq y \leq 1\), find \(E(X \mid Y = 1/4)\).

Ex. 6.11 (Computer exercise) Use R to estimate \(P(0 \leq X \leq 0.5, 0 \leq Y \leq 0.5)\) if the random vector, \((X, Y)\), has the bivariate uniform distribution in Example 6.3. (Generate a random sample of size 500 from the bivariate uniform distribution then count the number of pairs, \((x, y)\), which satisfy \(0 < x < 0.5\) and \(0 < y < 0.5\).)

Ex. 6.12 (Computer exercise) Estimate the probabilities in Example 6.5 using simulation.

Ex. 6.13 Assume the proportion of mature males in the population of crocodiles is \(p_1\) and the proportion of mature females is \(p_2\). (NOTE: \(p_1 + p_2 < 1\))

(a) If a sample of 5 individuals is randomly selected what is the probability that 2 are mature males and 1 is a mature female? Evaluate this for \(p_1 = 0.3\) and \(p_2 = 0.15\).
(b) Consider a sample of size \(n\). Let \(X\) denote the number of mature males and \(Y\) the number of mature females in the sample. Find \(E(X - Y)\).
(c) What is the variance of \((X - Y)\)?

(d) Use R to simulate the results in (a)–(c) above.

Ex. 6.14 (Computer exercise) Suppose a fair die is tossed 120 times. Estimate \(E(X_i)\), \(\text{var}(X_i)\) and \(\text{cov}(X_i, X_j)\) for \(i, j = 1, 2, 3, 4, 5, 6, i \neq j\), using simulation. (Note: Write an R script to solve this problem.)

Ex. 6.15 Let \(X\) and \(Y\) be the body mass index (BMI) and percentage body fat for netballers attending the AIS. Assume \(X\) and \(Y\) have a bivariate normal distribution with \(\mu_X = 23, \mu_Y = 21, \sigma_X = 3, \sigma_Y = 6\) and \(\rho_{XY} = 0.8\). Find

(a) the expected BMI of a netballer who has a percent body fat of 30,

(b) the expected percentage body fat of a netballer who has a BMI of 19.

Ex. 6.16 Let \(X\) and \(Y\) have a bivariate normal distribution with parameters \(\mu_x = 1, \mu_y = 4, \sigma^2_x = 4, \sigma^2_y = 9\) and \(\rho = 0.6\). Find

(a) \(P (-1.5 < X < 2.5)\) (b) \(P (-1.5 < X < 2.5 | y = 3)\)

(c) \(P (0 < Y < 8)\) (d) \(P (0 < Y < 8 | x = 0)\)

Ex. 6.17 Jim and Julie each roll a fair, six-sided die and observe the numbers on the top faces. Event \(A\) is defined to be the maximum of the two numbers, and event \(B\) is defined to be the minimum of the two numbers.

(a) Construct the joint probability function for \(A\) and \(B\).

(b) Determine the marginal distribution for \(B\).

Ex. 6.18 For a rv \(Z\) which has a normal distribution with mean 0 and variance 1, show that \(E(|Z|) = \sqrt{2/\pi}\).

Hence (or otherwise) find \(E(|X - \mu|)\) where \(X \sim N(\mu, \sigma^2)\).

Ex. 6.19 Wilks [34, p 101] states that the maximum temperatures measured at Ithaca \((I)\) and Canandaigua \((C)\) in January 1987 are both symmetrical. He goes on to say that the two temperatures could be modelled with a bivariate normal distribution with \(\mu_I = 29.87, \mu_C = 31.77, \sigma_I = 7.71, \sigma_C = 7.86\) and \(\rho_{IC} = 0.957\). (All measurements are in degrees Fahrenheit.)

(a) Explain, in context, what a correlation coefficient of 0.957 means.

(b) Determine the marginal distributions of \(C\) and of \(I\).

(c) Find the conditional distribution of \(C \mid I\).
6.10. Some answers and hints

(d) Plot the pdf of Canandaigua maximum temperature.
(e) Plot the conditional pdf of Canandaigua maximum temperature given that the maximum temperature at Ithaca is 25°F.
(f) Comment on the differences between the two pdfs plotted above.
(g) Find \( P(C < 32 \mid I = 25) \).
(h) Find \( P(C < 32) \).
(i) Comment on the differences between the answers in (g) and (h).

Ex. 6.20 The random variable \( U \) has a normal distribution with mean 5 and variance 2. The random variable \( V \) is independent of \( U \), and has a gamma distribution with mean 4 and variance 1. Find the mean and variance of \( 2U - V \).

6.10 Some answers and hints

6.1 (a) This corresponds to the cell \( X = 1, Y = 2 \): the answer is 1/20.
(b) \( P(X + Y \leq 1) = P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 1, Y = 0) = 1/2 \).
(c) \( P(X > Y) = P(X = 1, Y = 0) + P(X = 2, Y = 0) + P(X = 2, Y = 1) = 7/30 \).
(d) \[
p_X(x) = \begin{cases} 
7/15 & \text{if } x = 0 \\
7/15 & \text{if } x = 1 \\
1/15 & \text{if } x = 2 \\
0 & \text{otherwise}
\end{cases}
\]
(e) Only consider the column corresponding to \( X = 1 \):
\[
p_{Y \mid X=1}(y \mid x = 1) = \begin{cases} 
(1/6)/(7/15) = 15/42 & \text{if } y = 0 \\
(1/4)/(7/15) = 15/28 & \text{if } y = 1 \\
(1/20)/(7/15) = 3/28 & \text{if } y = 2 \\
0 & \text{otherwise}
\end{cases}
\]

6.2 (a) 7 (b) 5 (c) -6 (d) 35

6.3 samsz <- 500
die1 <- sample(1:6,replace=T,size=samsz)
die2 <- sample(1:6,replace=T,size=samsz)
confusion.table <- table(die1,die2)
prob56 <- diag(confusion.table)[5:6]/samsz
Module 6. Bivariate distributions

\begin{verbatim}
fives <- (die1==5) + (die2==5)
sixes <- (die1==6) + (die2==6)
freq56 <- table(fives,sixes)
freq56/samsz
\end{verbatim}

\textbf{6.4} (a) Ratio of (area of circle)/(Area of square).
(b) (i) Generate a random sample of size 500 from the bivariate uniform distribution. (ii) Count the number of observations, \((x, y)\), that satisfy \(x^2 + y^2 \leq 1:\)

\begin{verbatim}
nsim<-500
x<-runif(n=nsim,min=-1,max=1)
y<-runif(n=nsim,min=-1,max=1)
inside<-sum(x**2+y**2 <= 1)
inside/nsim
\end{verbatim}

and so estimate \(\pi/4\). (iii) Repeat to increase the sample size.

\textbf{6.5} (a) \(k = 1/11\) (c) 5/11 (d) 5/11

\textbf{6.6} (a) \(k = 4\) (b) \(x_1^3y_0^2\) (c) 0.0549

\textbf{6.7} (b) 1/2 (c) 2/5 (d) 11/12 (e) 1/2

\textbf{6.8} (a) Use Definition 6.7. \(f_X(x) = 1, 0 \leq x \leq 1, f_Y(y) = 1, 0 \leq y \leq 1.\)
(b) Use Definition 6.10. (c) \(\rho_{XY} = -\frac{\alpha}{3}\). (d) \(\frac{1}{2}\)

\textbf{6.9} (a) \(F_Y(y \mid x) = \frac{y(2x + y)}{2x + 1}\) (b) \(f_{X,Y}(x, y) = x + y, x \in [0, 1], y \in [0, 1]\)
(c) 1/2

\textbf{6.10} Use Definition 6.19. Ans = 2/3

\textbf{6.11} \textbf{METHOD:} First generate 500 random vectors \((X, Y)\) from the bivariate uniform distribution in Example 6.3. Do this by generating two columns of 500 random numbers from a uniform distribution on \([0, 1]\). Each row then forms a random vector from the required distribution. Count the number of \((x, y)\) that satisfy \(0 < x < 0.5\) and \(0 < y < 0.5.\)

\begin{verbatim}
samsz<-500
u1<-runif(n=samsz,min=0,max=1)
u2<-runif(n=samsz,min=0,max=1)
# a logical vector T/F depending on whether conditions are met
datain<-(u1 < 0.5) & (u2 < 0.5)
# coercing the logical vector into a 1/0 vector and summing
N.cond.met<-sum(datain)
Pulu2lt.half<-N.cond.met/samsz
print(Pulu2lt.half)
\end{verbatim}
6.10. Some answers and hints

(For one simulation I obtained a sum of 120 giving and estimate of 0.24.) Try repeating the simulation to give a total sample size of 1000.

6.12 Method: Simulate throwing 2 dice 500 times. Each row now represents one outcome of throwing two dice. Count how many pairs of fives and sixes are showing when the two dice are thrown and get estimates of the probabilities.

samsz<-500
die1<-sample(1:6,replace=T,size=samsz)
die2<-sample(1:6,replace=T,size=samsz)
confusion.table<-table(die1,die2)
prob56<-diag(confusion.table)[5:6]/samsz

fives<-(die1==5) + (die2==5)
sixes<-(die1==6) + (die2==6)
freq56<-table(fives,sixes)
freq56/samsz

6.13 (a) See Definition 6.25, with $k = 3, p_3 = 0.55$. Ans=0.1225
(b) Use Theorem 6.26.
(c) Use Theorem 6.26. Note $X$ and $Y$ are not independent.
(d) Represent mature males as 1, mature females as 2 and others as 3.
(i) Make a random vector of 1, 2, 3’s with probabilities $p_1, p_2, p_3$.
(ii) Count the frequency of 1, 2 or 3.
To simulate (b) and (c) first subtract counts[2] from counts[1] to get the mean and variance of the difference.
(iii) Repeat 1000 times.

nsim <- 10000
X2Y1 <- 0
diff <- array(0,nsim)
counts <- array(0,3)

for (i in 1:nsim){
crocs <- sample(1:3,prob=c(0.3,0.15,0.55),size=5,replace=T)
for (j in 1:3){ counts[j] <- length(crocs[crocs==j])}

if (counts[1]==2 & counts[2]==1) X2Y1 <- X2Y1+1
diff[i] <- counts[1]-counts[2]
}
print(X2Y1/nsim)
print(mean(diff))
print(var(diff))
6.14 Method: First simulate tossing a die 120 times and record the results in c1. Count the number of 1’s, 2’s, . . . , 6’s, (out of 120) that occur in each of 100 simulations and save the results in a matrix. Apply the mean and var functions to the columns (MARGIN=2) of the matrix.

```r
nsim<-100
frequencies<-matrix(0,nrow=100,ncol=6)
for (i in 1:nsim){
  face<-sample(1:6,replace=T,size=120)
  for ( j in 1:6){frequencies[i,j]<-sum(face==j)}
}
mnXi<-apply(frequencies,MARGIN=2,FUN=mean)
varXi<-apply(frequencies,MARGIN=2,FUN=var)
```

6.15 (a) Use Theorem 6.28. Ans = 26.6 (b) 14.6

6.16 (a) 0.6678 (b) E(X | Y = 3) = 0.6, var(X | Y = 3) = 2.56, Ans = 0.7878
(c) 0.8164 (d) E(Y | X = 0) = 3.1, var(Y | X = 0) = 5.76, Ans = 0.8808

6.18 $|Z| = \begin{cases} 
Z & Z \geq 0 \\
-Z & Z < 0 
\end{cases}$

Thus integral should be written as sum of two integrals over $(-\infty, 0)$ and $[0, \infty)$. Note $\frac{X - \mu}{\sigma} \sim N(0,1)$
Module 7

Transformations of random variables

Contents

7.1 Introduction ................................................. 258
7.2 The change of variable method ............................. 258
  7.2.1 Discrete random variables .............................. 259
  7.2.2 Continuous random variables ......................... 263
7.3 The distribution function method .......................... 267
7.4 The moment generating function method .................. 271
7.5 The chi-square distribution ............................... 273
7.6 Self-assessment exercises ................................. 275
7.7 Exercises ................................................... 276
7.8 Some answers and hints .................................... 278

Module objectives

On completion of this module, you should be able to:

• derive the distribution of a transformed variable, given the distribution of the original variable using either
– the distribution function method
– the change of variable method
– the moment generating function method

as appropriate

• find the joint distribution of two transformed variables in a bivariate situation

7.1 Introduction

Reading 7.1  DGS, Section 3.8; WMS, Sections 6.1 to 6.4.

The type of problem considered in this chapter is as follows. Given a rv $X$ with known distribution, and a function $u$, how do we find the probability distribution of a rv $Y = u(X)$?

Among several available techniques, three are considered here:

• the distribution function method
• the change of variable method
• the moment generating function method

A concept that is important in this context is that of 1:1 transformations as contained in the following definition.

Definition 7.1  Given rv’s $X$ and $Y$ with range spaces $R_X$, $R_Y$ respectively, the function $u$ is said to represent a 1:1 transformation (or mapping) if to each $x \in R_X$ there corresponds exactly one $y \in R_Y$.

When $Y = u(X)$ is a 1:1 transformation, the inverse function is uniquely defined; that is, $X$ can be written uniquely in terms of $Y$. This feature is important when considering the distribution of $Y$ when the distribution of $X$ is known.

7.2 The change of variable method

The discrete and continuous cases are considered separately. The method is relatively straightforward for one-to-one transformations; that is, functions that are only decreasing functions (such as $Y = 1 - X$), or only increasing functions (such as $Y = \exp(X)$.) Considerable care needs to be exercised if the transformation is not 1–1. Examples are given below.
7.2. The change of variable method

7.2.1 Discrete random variables

Univariate case

Let \( X \) be a discrete random variable with \( p_X(x) \). Let \( R_X \) denote the set of discrete points at each of which \( p_X(x) > 0 \). Let \( y = u(x) \) define a \textit{one-to-one transformation} that maps \( R_X \) onto \( R_Y \), the set of discrete points at each of which the transformed variable \( Y \) has a non-zero probability. If we solve \( y = u(x) \) for \( x \) in terms of \( y \), say \( x = w(y) \), then for each \( y \in R_Y \), we have \( x = w(y) \in R_X \).

Example 7.1  Given \( P(X = x) = x/15 = p_X(x), x \in R_X = \{1, 2, 3, 4, 5\} \), find the probability function of \( Y \) where \( Y = 2X + 1 \).

Note that \( R_Y = \{3, 5, 7, 9, 11\} \) and the mapping \( y = 2x + 1 = u(x) \) is 1:1. We wish to find a ‘formula’ for \( P(Y = y), y \in R_Y \).

So the probability function of \( Y \) is

\[
P(Y = y) = \frac{y - 1}{30}, \quad y = 3, 5, 7, 9, 11.
\]

(Note that the probabilities given by this ‘formula’ add to 1.)

The above procedure for \( Y = u(X) \) a 1:1 mapping can be stated \textit{more generally} as follows.

\[
P(Y = y) = P(u(X) = y) = P(X = u^{-1}(y)) = p_X(u^{-1}(y)), \quad y \in R_Y
\]

Example 7.2  Let \( X \) have a binomial distribution with \( p_X(x) \)

\[
p_X(x) = \begin{cases} \binom{3}{x}(0.2)^x(0.8)^{3-x} & \text{for } x = 0, 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}
\]

The pf of \( Y = X^2 \) can be found using the change of variable methods as follows.

First note that \( Y = X^2 \) is not a one-to-one transformation in general, but it is here since \( X \) has non-zero probability only for \( x = 0, 1, 2, 3 \).

The transformation \( y = u(x) = x^2, R_X = \{x \mid x = 0, 1, 2, 3\} \) maps onto \( R_Y = \{y \mid y = 0, 1, 4, 9\} \). The inverse function becomes \( x = w(y) = \sqrt{y} \), and hence the pf of \( Y \) becomes

\[
p_Y(y) = p_X(\sqrt{y}) = \begin{cases} \binom{3}{\sqrt{y}}(0.2)^{\sqrt{y}}(0.8)^{3-\sqrt{y}} & \text{for } y = 0, 1, 4, 9 \\ 0 & \text{otherwise} \end{cases}
\]
We now consider the case where the function \( u \) is not 1:1.

**Example 7.3** Suppose \( P(X = x) \) is as in Example 7.1 and define \( Y = |X - 3| \). Then since \( R_Y = \{0, 1, 2\} \) the mapping is clearly not 1:1. In fact, the event \( Y = 0 \) occurs if \( X = 3 \), the event \( Y = 1 \) occurs if \( X = 2 \) or 4, and the event \( Y = 2 \) occurs if \( X = 1 \) or 5.

To find the probability distribution of \( Y \) means finding the probabilities associated with the values 0, 1, 2 in \( R_Y \), and we have

\[
\begin{align*}
P(Y = 0) &= P(X = 3) = \frac{3}{15} \\
P(Y = 1) &= P(X = 2 \text{ or } 4) = \frac{2}{15} + \frac{1}{15} = \frac{1}{2} \\
P(Y = 2) &= P(X = 1 \text{ or } 5) = \frac{1}{15} + \frac{5}{15} = \frac{2}{5}
\end{align*}
\]

So the probability function of \( Y \) can be expressed as table,

<table>
<thead>
<tr>
<th>( y )</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(Y = y) )</td>
<td>( \frac{1}{5} )</td>
<td>( \frac{2}{5} )</td>
<td>( \frac{2}{5} )</td>
</tr>
</tbody>
</table>

**Bivariate case**

The bivariate case is similar to the univariate case. Here we have a joint pf \( p_{X_1,X_2}(x_1, x_2) \) of two discrete random variables \( X_1 \) and \( X_2 \) defined on the two-dimensional set of points \( R^2_X \) for which \( p(x_1, x_2) > 0 \). There are now two one-to-one transformations:

\[
\begin{align*}
y_1 &= u_1(x_1, x_2) \\
y_2 &= u_2(x_1, x_2)
\end{align*}
\]

that map \( R^2_X \) onto \( R^2_Y \) (the two-dimensional set of points for which \( p(y_1, y_2) > 0 \)).

The two inverse functions are

\[
\begin{align*}
x_1 &= w_1(y_1, y_2) \\
x_2 &= w_2(y_1, y_2)
\end{align*}
\]

Then the joint pf of the new (transformed) random variables is

\[
p_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 
p_{X_1,X_2}(w_1(y_1, y_2), w_2(y_1, y_2)) & \text{where } (y_1, y_2) \in R^2_Y \\
0 & \text{elsewhere}
\end{cases}
\]
Example 7.4 Let the two discrete random variables $X_1$ and $X_2$ have the joint pf

<table>
<thead>
<tr>
<th>$x_2$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Consider the two one-to-one transformations

$$Y_1 = X_1 + X_2$$
$$Y_2 = 2X_1.$$

The joint pf of $Y_1$ and $Y_2$ can be found by noting where the $(x_1, x_2)$ pairs are mapped to in the $y_1, y_2$ space:

$$
\begin{align*}
(x_1, x_2) & \mapsto (y_1, y_2) \\
(-1, 0) & \mapsto (-1, -2) \\
(-1, 1) & \mapsto (0, -2) \\
(-1, 2) & \mapsto (1, -2) \\
(1, 0) & \mapsto (1, 2) \\
(1, 1) & \mapsto (2, 2) \\
(1, 2) & \mapsto (3, 2)
\end{align*}
$$

The joint pf can then be constructed as follows:

<table>
<thead>
<tr>
<th>$y_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_2$</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>-2</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

Sometimes, a joint pf of two random variables is given, but only one new random variable is required. In this case, a second (dummy) transformation is developed, usually one that is simple. See the following example.

Example 7.5 Let $X_1$ and $X_2$ be two independent random variables with the joint pf

$$p_{X_1, X_2}(x_1, x_2) = \frac{\mu_1^{x_1} \mu_2^{x_2} \exp(-\mu_1 - \mu_2)}{x_1! x_2!}$$

for $x_1$ and $x_2 = 0, 1, 2, \ldots$
Module 7. Transformations of random variables

(This is the joint pdf of two independent Poisson random variables.) Suppose we wish to find the pdf of \( Y_1 = X_1 + X_2 \).

We can consider the two one-to-one transformations

\[
\begin{align*}
y_1 &= x_1 + x_2 = u_1(x_1, x_2) \\
y_2 &= x_2 = u_2(x_1, x_2)
\end{align*}
\]

which maps the points in \( R_X^2 \) onto

\[
R_Y^2 = \{(y_1, y_2) \mid y_1 = 0, 1, 2, \ldots; y_2 = 0, 1, 2, \ldots, y_1\}.
\]

Notice that \( Y_2 \) is a dummy transform, and it is very simple. We could have chosen any second transform (as it is of no interest), and so chose one that is simple.

The inverse functions are given by

\[
\begin{align*}
x_1 &= y_1 - y_2 = w_1(y_1, y_2) \\
x_2 &= y_2 = w_2(y_2)
\end{align*}
\]

by rearranging Equations (7.1) and (7.2). Then the joint pdf of \( Y_1 \) and \( Y_2 \) is

\[
p_{Y_1,Y_2}(y_1, y_2) = p_{X_1,X_2}(x_1, x_2)(w_1(y_1, y_2), w_2(y_1, y_2))
\]

\[
= \frac{\mu_1^{y_1-y_2} \mu_2^{y_2} \exp(-\mu_1 - \mu_2)}{(y_1 - y_2)!y_2!}
\]

for \((y_1, y_2) \in R_Y^2\).

This does not end the question, since we originally sought the pdf of just \( Y_1 \); so we need to find the marginal pdf of \( p_{Y_1,Y_2}(y_1, y_2) \).

Now the marginal pdf of \( Y_1 \) is given by

\[
p_{Y_1}(y_1) = \sum_{y_2=0}^{y_1} p_{Y_1,Y_2}(y_1, y_2) = \sum_{y_2=0}^{y_1} \frac{\mu_1^{y_1-y_2} \mu_2^{y_2} \exp(-\mu_1 - \mu_2)}{(y_1 - y_2)!y_2!}.
\]

which can be shown to be

\[
p_{Y_1}(y_1) = \begin{cases} 
(\mu_1 + \mu_2)^{y_1} \exp(-(\mu_1 + \mu_2)) / y_1! & \text{for } y_1 = 0, 1, 2, \ldots \\
0 & \text{otherwise}. 
\end{cases}
\]

(Showing this is difficult and the details are not given.) This is the pdf of a Poisson random variable with mean \( \mu_1 + \mu_2 \). Thus \( Y_1 \sim \text{Pois}(\lambda = \mu_1 + \mu_2) \).
7.2. The change of variable method

7.2.2 Continuous random variables

Univariate case

**Theorem 7.2** If \( X \) has pdf \( f_X(x) \) for \( x \in \mathbb{R} \) and \( u \) is a strictly increasing or decreasing function for \( x \in \mathbb{R} \) then the rv \( Y = u(X) \) has pdf

\[
f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|
\]

where the RHS is expressed as a function of \( y \).

**Proof** Let the inverse function be \( X = w(Y) \), that is \( w(y) = u^{-1}(x) \).

*Case 1: \( y = u(x) \) is a strictly increasing function.*

If \( a < y < b \) then \( w(a) < x < w(b) \) and \( P(a < Y < b) = P(w(a) < X < w(b)) \), so that,

\[
\int_a^b f_Y(y) \, dy = \int_{w(a)}^{w(b)} f_X(x) \, dx = \int_a^b f\big\{w(y)\big\} \frac{dx}{dy} \, dy
\]

Therefore, \( f_Y(y) = f_X(w(y)) \frac{dx}{dy} \), where \( w(y) = u^{-1}(x) \).

*Case 2: \( y = u(x) \) is a strictly decreasing function of \( x \).*

If \( a < y < b \) then \( w(b) < x < w(a) \) and \( P(a < Y < b) = P(w(b) < X < w(a)) \), so that,
Figure 7.2: Transformation: Monotone decreasing function

\[
\int_a^b f_Y(y) \, dy = \int_{w(b)}^{w(a)} f_X(x) \, dx \\
= \int_b^a f_X(x) \frac{dx}{dy} \, dy \\
= -\int_a^b f_X(x) \frac{dx}{dy} \, dy.
\]

Therefore \( f_Y(y) = -f_X(w(y)) \frac{dx}{dy} \). But \( dx/dy \) is negative in the case of a decreasing function, so that in general,

\[
f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|.
\]

The absolute value of \( w'(y) = dx/dy \) is called the Jacobian of the transformation.

**Example 7.6** Let the pdf of \( X \) be given by

\[
f_X(x) = 1 \quad \text{for} \quad 0 < x < 1
\]

Consider the transformation \( Y = -2 \log X \). (Here, \( \log \) refers to a logarithm to base \( e \), or a natural logarithm.)

First note that the transformation is one-to-one. The inverse transformation is

\[
X = \exp(-Y/2) = w(Y).
\]
The space \( R_X = \{ x \mid 0 < x < 1 \} \) is mapped to \( R_y = \{ y \mid 0 < y < \infty \} \). Then,

\[
    w'(y) = \frac{d}{dy} \exp(-y/2) = -\frac{1}{2} \exp(-y/2),
\]

and so the Jacobian of the transformation \( |w'(y)| = \exp(-y/2)/2 \).

Now the pdf of \( Y = -2 \log X \) becomes

\[
    f_Y(y) = f_X(w(y)) |w'(y)| = f_X(\exp(-y/2)) \exp(-y/2)/2 = \frac{1}{2} \exp(-y/2) \text{ for } y > 0
\]

Note that \( Y \sim \exp(2) \).

---

**Example 7.7** Given \( X \) has pdf \( f_X(x) = e^{-x}, x \geq 0 \), find the pdf of \( Y = \sqrt{X} \).

For \( x \geq 0 \), \( y = \sqrt{x} \) is a strictly increasing function.

The inverse relation is \( x = y^2 \); \( dx/dy = 2y \). So the pdf of \( Y \) is

\[
    f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = 2y e^{-y^2}, \quad y \in [0, \infty).
\]
Example 7.8 Let rv $X$ be uniformly distributed on $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Find the distribution of $Y = \tan X$.

To find $R_Y$ consider the mapping $y = \tan x$. Then $R_Y = \{y \mid -\infty < y < \infty\}$ and the mapping is 1:1, which means $x = \tan^{-1} y$, and $dx/dy = 1/(1 + y^2)$. Hence

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = \frac{1}{\pi} \frac{1}{1 + y^2}$$

This is known as the Cauchy distribution.

A case where the function $u$ is not 1:1 is considered by an example. A modification of Theorem 7.2 is used.

Example 7.9 Given a random variable $Z$ which is distributed $N(0, 1)$, find the probability distribution of $Y = \frac{1}{2} Z^2$.

The relationship $y = u(z) = \frac{1}{2} z^2$ is not strictly increasing or strictly decreasing in $(-\infty, \infty)$ so we can’t apply Theorem 7.2 directly. However, we can subdivide the range of $z$ and $y$ so that in each portion the relationship is monotonic. Now

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2}, \quad -\infty < z < \infty$$

The inverse relation, $z = u^{-1}(y)$ is $z = \pm \sqrt{2y}$. For a given value of $y$ there are 2 values of $z$. However, in the range $-\infty < z < 0$, $y$ and $z$ are monotonically related. Similarly, for $0 < z < \infty$, $y$ and $z$ are
monotonically related. Thus (see Figure 7.9),
\[ P(a < Y < b) = P\left(-\sqrt{2b} < Z < -\sqrt{2a}\right) + P\left(\sqrt{2a} < Z < \sqrt{2b}\right). \]
The two terms on the right are equal because the distribution of \( Z \) is symmetrical about 0.
Thus \( P(a < Y < b) = 2P\left(\sqrt{2a} < Z < \sqrt{2b}\right) \), and
\[
f_Y(y) = 2f_Z(z) \left| \frac{dz}{dy} \right| = 2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2} y^{-1/2} \]
\[ f_Y(y) = e^{-y y^{-1/2}/\sqrt{\pi}}, \quad 0 < y < \infty \]
This pdf is that of a gamma distribution with parameters \( \alpha = 1/2 \) and \( \beta = 1 \). It follows from this result that if \( X \) is \( N(\mu, \sigma^2) \), the pdf of \( Y = \frac{1}{2} (X - \mu)^2 / \sigma^2 \) is also Gamma(\( \alpha = 1/2, \beta = 1 \)) since then \( \frac{X - \mu}{\sigma} \) is distributed \( N(0, 1) \).

Note that the probability can only be doubled as in Example 7.9 if both \( Y = u(Z) \) and the pdf of \( Z \) are symmetrical about the same point.

### 7.3 The distribution function method

This method only works for continuous random variables.
Module 7. Transformations of random variables

Figure 7.6: The transformation $Y = X^2$ when $X$ is defined from 1 to 3. The thicker line corresponds to the region where the transformation applies. Note that for $Y < y$, $2 - \sqrt{y} - 1 < X < 2 + \sqrt{y} - 1$.

There are two basic steps. Firstly, the distribution function of the transformed variable is obtained. Secondly, the probability density function is derived by differentiation.

The procedure is best demonstrated using an example.

**Example 7.10** Consider the random variable $X$ with pdf

$$f_X(x) = \begin{cases} 
  x/4 & \text{for } 1 < x < 3 \\
  0 & \text{elsewhere}
\end{cases}$$

Suppose we seek the pdf of the random variable $Y$ where $Y = X^2$. Note that $1 < y < 9$.

The distribution function for $Y$ is

$$F_Y(y) = P(Y \leq y) \quad (\text{by definition})$$

$$= P(X^2 \leq y) \quad (\text{since } Y = X^2)$$

$$= P(X \leq \sqrt{y}).$$

This last step is not trivial, but it is critical. (The next example shows a situation where more care is needed.) In this case, there is a one-to-one relationship between $X$ and $Y$ over the region of which $X$ is defined (ie has a positive probability); see Figure 7.6.
7.3. The distribution function method

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

Then continue as follows:

\[ F_Y(y) = P(X \leq \sqrt{y}) \]
\[ = F_X(\sqrt{y}) \quad \text{(by definition of } F_X(x) \text{)} \]
\[ = \int_{1}^{\sqrt{y}} \left( x/4 \right) dx \]
\[ = (y - 1)/8 \]

for \( 1 < y < 9 \), and is zero elsewhere. Now recall this has found the distribution function of \( Y \). To find the pdf we must differentiate:

\[ f_Y(y) = \frac{d}{dy} \left( \frac{y - 1}{8} \right) = \begin{cases} 
1/8 & \text{for } 1 < y < 9 \\
0 & \text{elsewhere} 
\end{cases} \]

Note the range for which \( Y \) is defined; since \( 1 < x < 3 \), then \( 1 < y < 9 \).

**Example 7.11** Consider the same random variable \( X \) as in the previous example, but the transformation \( Y = (X - 2)^2 + 1 \) (see Figure 7.7).

In this case, the transformation is not a one-to-one transform. Proceed as before to find the distribution function of \( Y \).

\[ F_Y(y) = P(Y \leq y) \quad \text{(by definition)} \]
\[ = P\left( (X - 2)^2 + 1 \leq y \right) \quad \text{(since } Y = (X - 2)^2 + 1). \]
At this point care is needed. From Figure 7.7, whenever \((X - 2)^2 + 1 < y\) for some value \(y\), then \(X\) must be in the range \(2 - \sqrt{y - 1}\) to \(2 + \sqrt{y - 1}\). So we proceed

\[
F_Y(y) = P ((X - 2)^2 + 1 \leq y)
\]
\[
= P \left( 2 - \sqrt{y - 1} < X < 2 + \sqrt{y - 1} \right)
\]
\[
= \int_{2 - \sqrt{y - 1}}^{2 + \sqrt{y - 1}} \frac{x}{4} \, dx
\]
\[
= \frac{1}{8} \left. x^2 \right|_{2 - \sqrt{y - 1}}^{2 + \sqrt{y - 1}}
\]
\[
= \frac{1}{8} \left[ (2 + \sqrt{y - 1})^2 - (2 - \sqrt{y - 1})^2 \right]
\]
\[
= \sqrt{y - 1}.
\]

Again, this is the distribution function; so

\[
f_Y(y) = \begin{cases} 
\frac{1}{2\sqrt{y-1}} & \text{for } 1 < y < 2 \\
0 & \text{elsewhere}
\end{cases}
\]

---

**Example 7.12**  Example 7.9 is repeated here using the distribution function method.

Given \(Z\) is distributed \(N(0, 1)\) we seek the probability distribution of \(Y = \frac{1}{2}Z^2\).

\[
f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in (-\infty, \infty).
\]

Let \(Y\) have pdf \(f_Y(y)\) and df \(F_Y(y)\). Then

\[
F_Y(y) = P (Y \leq y) = P \left( \frac{1}{2}Z^2 \leq y \right) = P \left( Z^2 \leq 2y \right)
\]
\[
= P \left( -\sqrt{2y} \leq Z \leq \sqrt{2y} \right)
\]
\[
= F_Z(\sqrt{2y}) - F_Z(-\sqrt{2y})
\]
7.4. The moment generating function method

where $F_Z$ is the df of $Z$. Hence

$$f_Y(y) = F'_Y(y) = F'_Z(\sqrt{2y}) - F'_Z(-\sqrt{2y})$$

$$= \frac{\sqrt{2}}{2\sqrt{y}} f_Z(\sqrt{2y}) - \frac{\sqrt{2}}{2\sqrt{y}} f_Z(-\sqrt{2y})$$

$$= \frac{1}{\sqrt{2y}} [f_Z(\sqrt{2y}) + f_Z(-\sqrt{2y})]$$

$$= \frac{1}{2y} \left[ \frac{1}{\sqrt{2\pi}} e^{-y} + \frac{1}{\sqrt{2\pi}} e^{-y} \right]$$

$$= \frac{e^{-y} - \frac{1}{2}} {\sqrt{\pi}}$$

as before.

Obviously, care is needed to ensure the steps are followed logically. Drawing diagrams like Figure 7.6 and 7.7 are certainly encouraged. Note that the functions you finish with should be pdfs; you should check that they actually are.

This method can also be used when there is more than one variable of interest. See the texts for examples.

7.4 The moment generating function method

The moment generating function (mgf) of a random variable, if it exists, completely specifies the distribution of the random variable. Thus, if two random variables have the same mgf, then they must have identical distributions. The mgf method used here is very useful to find the distribution of a linear combination of $n$ independent random variables. The method essentially involves the computation of the mgf of the transformed variable $Y = u(X_1, X_2, \ldots, X_n)$ when the joint distribution of independent $X_1, X_2, \ldots, X_n$ is given.

We construct the method below for the transformation $Y = X_1 + X_2 + \cdots + X_n$, but the same principles can be applied for other linear combinations also.

Consider $n$ independent random variables $X_1, X_2, \ldots, X_n$ with mgfs $M_{X_1}(t)$, $M_{X_2}(t)$, \ldots, $M_{X_n}(t)$. We consider the transformation $Y = X_1 + X_2 + \cdots + X_n$. 

Reading 7.2 WMS, Section 6.5.
Since the $X_i$ are independent, $f_{X_1,X_2,...,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n)$. So, by definition of the mgf,

$$M_Y(t) = E(\exp(tY)) = E(\exp(t(X_1 + X_2 + \cdots + X_n)))$$

$$= \int \cdots \int \exp[t(x_1 + x_2 + \cdots + x_n)]f(x_1, x_2, \ldots, x_n) \, dx_1 \cdots dx_n \, dx_2 \, dx_1$$

$$= \int \cdots \int \exp(tx_1)f(x_1)\exp(tx_2)f(x_2)\cdots\exp(tx_n)f(x_n) \, dx_n \cdots dx_2 \, dx_1$$

$$= \int \exp(tx_1)f(x_1) \, dx_1 \int \exp(tx_2)f(x_2) \, dx_2 \cdots \int \exp(tx_n)f(x_n) \, dx_n$$

$$= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t)$$

$$= \prod_{i=1}^{n} M_{X_i}(t)$$

Hence we get the following result: If $X_1, X_2, \ldots, X_n$ are independent random variables and $Y = X_1 + X_2 + \cdots + X_n$, then the mgf of $Y$, $M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$ where $M_{X_i}(t)$ is the value of the mgf of $X_i$ at $t$ for $i = 1, 2, \ldots, n$.

Note that $\prod_{i=1}^{n} a_i = a_1 a_2 \cdots a_n$; $\prod$ is the symbol for a product in the same way that $\sum$ is the symbol for a summation.

The above result also holds for discrete variables; just replace the integrations by summations.

**Example 7.13** Suppose that $X_i \sim \text{Pois}(\lambda_i)$ for $i = 1, 2, \ldots, n$. What is the distribution of $Y = X_1 + X_2 + \cdots + X_n$?

**Solution** Since $X_i$ has a Poisson distribution with parameter $\lambda_i$, the mgf of $X_i$ is

$$M_{X_i}(t) = \exp[\lambda_i(e^t - 1)].$$

The mgf of $Y = X_1 + X_2 + \cdots + X_n$ is

$$M_Y(t) = \prod_{i=1}^{n} \exp[\lambda_i(e^t - 1)]$$

$$= \exp[\lambda_1(e^t - 1)] \exp[\lambda_2(e^t - 1)] \cdots \exp[\lambda_n(e^t - 1)]$$

$$= \exp \left( (e^t - 1) \sum_{i=1}^{n} \lambda_i \right).$$
7.5 The chi-square distribution

Using \( \Lambda = \sum_{i=1}^{n} \lambda_i \), the mgf of \( Y \) is

\[
M_Y(t) = \exp \left[(e^t - 1)\Lambda\right],
\]

which is the mgf of a Poisson distribution with mean \( \Lambda = \sum_{i=1}^{n} \lambda_i \).
This means that the sum of \( n \) independent Poisson distribution is also a Poisson distribution, whose mean is the sum of the individual Poisson means.

7.5 The chi-square distribution

Examples 7.9 and 7.12 give rise to the chi-square distribution, which is an important model in statistical theory.

**Definition 7.3** A continuous random variable \( X \) with probability density function

\[
f_X(x) = \frac{x^{(\nu/2)-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}, \quad x > 0 \tag{7.3}
\]

is said to have a chi-square distribution with parameter \( \nu (> 0) \). The parameter is known as the degrees of freedom. We write \( X \sim \chi^2(\nu) \).

The chi-square distribution is actually a special case of the gamma distribution. Comparison of Definitions 7.3 and 5.11 reveal that \( X \) has a \( \chi^2(\gamma) \) distribution if and only if \( X \) has a gamma distribution with parameters \( \beta = 2 \) and \( \alpha = \nu/2 \).

The basic properties of the chi-square follow directly from those of the gamma distribution.

**Theorem 7.4** If \( X \sim \chi^2(\nu) \) then

1. \( E(X) = \nu \)
2. \( \text{var}(X) = 2\nu \)
3. \( M_X(t) = (1 - 2t)^{-\nu/2} \)

The importance of the chi-square distribution is hinted at in Examples 7.9 and 7.12, which essentially prove the following theorem.

Theorem 7.5 If \( Z \sim N(0, 1) \) then \( Z^2 \) has a chi-square distribution with one degree of freedom.

Proof  Exercise—see Example 7.9.

A useful property of the chi-square distribution is that the sum of independent rvs, each with a chi-square distribution, also has a chi-square distribution. This property is exemplified in the following theorem, which will be made use of later.

Theorem 7.6 If \( Z_1, Z_2, \ldots, Z_n \) are independently and identically distributed as \( N(0, 1) \), then the sum of square \( S = \sum_i Z_i^2 \) has a \( \chi^2(n) \) distribution.

Proof  The elegant way to prove this result is using mgf’s. Since \( Z_i \sim \chi^2(1) \), from Theorem 7.4

\[
M_{Z_i}(t) = (1 - 2t)^{-1/2}
\]

It follows then from Theorem 3.18 that \( S = \sum_{i=1}^{n} Z_i^2 \) has mgf

\[
M_S(t) = \prod_{i=1}^{n} (1 - 2t)^{-1/2} = \left[ (1 - 2t)^{-1/2} \right]^n = (1 - 2t)^{-n/2}
\]

which is the mgf of \( \chi^2(n) \).

Chi-square probabilities cannot in general be calculated without resorting to numerical integration which is done in R and Rcmdr. The functions are \texttt{pchisq(q = , df = )} to calculate probabilities from quantiles, \texttt{qchisq(p = , df = )} to calculate quantiles from probabilities.
Example 7.14 The variable $X$ has a chi-square distribution with 12 df. Determine the value of $X$ below which lies 90% of the distribution.

We seek a value $c$ such that $P(X < c) = F_X(c) = 0.90$ where $X \sim \chi^2(12)$.

$qchisq(p=0.9, df=12)$

[1] 18.54935

This assumed the default of `lower.tail=TRUE`.

7.6 Self-assessment exercises

The following exercises are designed to provide practice at problem-solving based on the material in this module. Solutions are provided at the end of the module. Additional exercises are available in the next section and in the textbook.

Ex. 7.1 Suppose the pdf of $X$ is given by

$$f_X(x) = \begin{cases} 
\frac{x}{2} & 0 < x < 2 \\
0 & \text{otherwise}
\end{cases}$$

(a) Find the pdf of $Y = X^3$ using the distribution function method.
(b) Find the pdf of $Y = X^3$ using the change of variable method.
Ex. 7.2 The discrete bivariate random vector \((X_1, X_2)\) has the joint pf

\[
    f_{X_1, X_2}(x_1, x_2) = \begin{cases} 
    (2x_1 + x_2)/14 & \text{for } x_1 = 0, 1; x_2 = 0, 1 \\
    0 & \text{elsewhere}
    \end{cases}
\]

Consider the transformations

\[
    Y_1 = X_1 + X_2 \\
    Y_2 = X_2
\]

(a) Determine the joint pf of \((Y_1, Y_2)\).
(b) Deduce the distribution of \(Y_1\).

Ex. 7.3 Consider \(n\) random variables \(X_i\) such that \(X_i \sim \text{Gam}(\alpha_i, \beta)\). Determine the distribution of \(Y = \sum_{i=1}^{n} X_i\).

7.7 Exercises

Ex. 7.4 If \(X\) is a random variable with probability function

\[
    P(X = x) = \binom{4}{x} (0.2)^x (0.8)^{4-x}, \quad x = 0, 1, 2, 3, 4,
\]

find the probability function of the random variable defined by \(Y = \sqrt{X}\).

Ex. 7.5 Given the random variable \(X\) with probability function

\[
    P(X = x) = \frac{x^2}{30}, \quad x = 1, 2, 3, 4,
\]

find the probability function of \(Y = (X - 3)^2\).

Ex. 7.6 A random variable \(X\) has pdf given by

\[
    f_X(x) = 1, \quad 0 < x < 1
\]

Find the pdf of the random variable \(Y\) defined by \(Y = -2 \log_e(X)\).

Ex. 7.7 The random variable \(X\) has pdf

\[
    f_X(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < \infty
\]

Find the pdf of \(Y\) where \(Y = X^2\).
### Ex. 7.8
If $W$ has a chi-square distribution with $\nu$ degrees of freedom and has pdf

$$f_W(w) = \frac{e^{-\frac{1}{2}w}w^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma(\frac{1}{2}\nu)}, \quad w > 0$$

find (a) the mode (b) $E(W)$ (c) $\text{var}(W)$ (d) mgf

### Ex. 7.9
A random variable $X$ has distribution function

$$F_X(x) = \frac{2x + 1}{2}, \quad -0.5 < x < 0.5$$

(a) Find the distribution function, $F_Y(y)$, of the random variable $Y = 4 - X^2$.

(b) Hence find the pdf of $Y$, $f_Y(y)$.

### Ex. 7.10
Suppose a projectile is fired at an angle $\theta$ from the horizontal with a velocity $v$. The horizontal distance that the projectile travels, $R$, is given by

$$R = \frac{v^2}{g} \sin 2\theta,$$

where $g$ is the acceleration due to gravity (usually set at $g \approx 9.8 \text{ m/s}^2$).

(a) If $\theta$ is uniformly distributed over the range $(0, \pi/4)$, find the probability density function of $R$.

(b) Sketch the pdf of $R$ over a suitable range for $v = 12$ and using $g \approx 9.8$.

### Ex. 7.11
Most computers have facilities to generate continuous uniform random numbers between zero and one, say $X$. When exponential random numbers are needed, they are usually obtained from $X$ using the transformation $Y = -\alpha \ln X$.

(a) Show that $Y$ has an exponential distribution and determine its parameters.

(b) Deduce the mean and variance of $Y$.

### Ex. 7.12
Consider a random variable $W$ for which $P(W = 2) = 1/6$, $P(W = -2) = 1/3$ and $P(W = 0) = 1/2$.

(a) Plot the probability function of $W$.

(b) Find the mean and variance of $W$.

(c) Determine the distribution of $V = W^2$.

(d) Find the distribution function of $W$.  

7.8 Some answers and hints

7.1 Note that for $0 < x < 2$ where $X$ is defined, the transformation $Y = X^3$ is a one-to-one transformation. The inverse transform is $X = Y^{1/3}$ and so $Y$ is defined for $0 < y < 8$.

(a) $F_Y(y) = P(Y \leq y) = P(X^3 \leq y) = P(X \leq y^{1/3}) = F_X(y^{1/3}) = \int_{u=0}^{y^{1/3}} u/2 \, du = u^2/4 \bigg|_{u=0}^{u=y^{1/3}} = y^{2/3}/4$. The pdf is then found by differentiating: $\frac{d}{dy} y^{2/3}/4 = y^{-1/3}/6$. So the pdf of $Y$ is

$$f_Y(y) = \begin{cases} y^{-1/3}/6 & \text{for } 0 < y < 8 \\ 0 & \text{otherwise} \end{cases}$$

(b) Note that $w(y) = y^{1/3}$ so that $w'(y) = y^{-2/3}3$.

$$f_Y(y) = f_X(y) |J| = y^{1/3}/2 \times y^{-2/3}/3 = y^{-1/3}/6.$$ 

So again the pdf of $Y$ is

$$f_Y(y) = \begin{cases} y^{-1/3}/6 & \text{for } 0 < y < 8 \\ 0 & \text{otherwise} \end{cases}$$

7.2 This one is much harder than it looks, mainly since $Y_1$ takes the values 1 and 2 for $Y_2 = 0$; but $Y_1$ takes the values 2 and 3 for $Y_2 = 1$.

The joint pdf is

$$
\begin{array}{cccc}
Y_1 = 1 & Y_1 = 2 & Y_1 = 3 \\
Y_2 = 0 & 2/14 & 4/14 & 0 \\
Y_2 = 1 & 0 & 3/14 & 5/14
\end{array}
$$

$f_{Y_1}(y = 1) = 2/14$; $f_{Y_1}(y = 2) = 7/14$; $f_{Y_1}(y = 3) = 5/14$.

7.3 $Y \sim \text{Gam} (\sum \alpha, \beta)$.

7.4 See Example 7.1.

7.5 The transformation is not 1-1. See Example 7.2. $P(Y = 0) = 9/30$, $P(Y = 1) = 20/30$, $P(Y = 2) = 1/30$.

7.6 Use Theorem 7.2. $g(y) = \frac{1}{2}e^{-y/2}$, $0 < y < \infty$
7.7 Transformation is not 1-1. See Example 7.9. \( f_Y(y) = \frac{1}{\pi \sqrt{y(1+y)}} \), \( 0 < y < \infty \).

7.8 (a) \( \nu - 2 \) (b) \( \nu \) (c) \( 2\nu \) (d) \( (1-2t)^{-\nu/2} \)

7.9 (a) \( F_Y(y) = P(Y \leq y) = 1 - 2\sqrt{1-y}, 3.75 < y < 4 \) (b) \( f_Y(y) = \frac{(4-y)^{-1/2}}{(4-y)^{-1/2}} \)
Module 7. Transformations of random variables
Module 8

Sampling distributions

Contents

8.1 Introduction ........................................... 282
8.2 Sampling distributions ................................. 284
  8.2.1 Random sampling .................................. 284
  8.2.2 The sampling distribution of the mean ........ 286
8.3 Sampling distributions related to the normal distribution ........................................... 287
  8.3.1 The t distribution .................................. 292
  8.3.2 The F distribution ................................. 295
8.4 The Central Limit Theorem ............................ 297
8.5 The normal approximation to the binomial .......... 300
8.6 Self-assessment exercises ............................. 301
8.7 Exercises .............................................. 302
8.8 Some answers and hints ............................... 305

Module objectives

On completion of this module, students should be able to:
• use appropriate techniques to determine the sampling distributions of some statistics \((t, F, \chi^2)\) related to the normal distribution
• understand the concepts of the Central Limit Theorem and be able to apply the Central Limit Theorem in appropriate circumstances
• use the Central Limit Theorem to approximate binomial probabilities by normal probabilities in appropriate circumstances

8.1 Introduction

This module introduces the application of distribution theory to the discipline of statistics. Distribution theory is about describing random variables using the concept of probability. Statistics is about data collection and extracting information from data. A simple example demonstrates how these two things are related.

Example 8.1 A simple random sample (SRS) is taken from a population. Interest is in the mean of the sample.

Typically the population of potential values will be very large. For demonstration purposes however suppose the population contains only \(N = 5\) elements, 2, 3, 3, 6, 8, and assume that the sample is of size \(n = 2\) and sampling is without replacement.

By simple random sampling we mean all possible samples of size \(n\) are equally likely. But there are \(\binom{N}{n}\) samples of size \(n\) sampling without replacement from \(N\) objects. Hence in our example there are \(\binom{5}{2} = 10\) equally likely possible samples. For example, the sample \(\{2, 6\}\) has a probability of \(1/10\) of being selected. Further, the mean of this sample, is 4.0.

We can systematically list all 10 samples as shown in Table 8.1.

From this table we can obtain the distribution of the mean itself—see Table 8.2. This distribution is known as the sampling distribution of the mean in recognition that it’s based on sampling data.

Notice we can determine the mean and variance of the distribution as \(\mu_X = 4.4\) and \(\sigma^2_X = 1.89\).

It is instructive to repeat this example, this time sampling randomly with replacement. Then there are 25 equally-likely samples possible. You should show that the mean and variance of the sampling distribution in this case are respectively \(\mu_X = 4.4\) and \(\sigma^2_X = 2.52\).
8.1. Introduction

In principle we could obtain the distribution of any statistic, whether it be a mean, median, variance, range or whatever, using the approach in Example 8.1. In practice, the number of possible samples may be astronomical or infinite and the exact elements of the population difficult or impossible to list.

In terms of distribution theory we should see Table 8.2 as just the distribution of a random variable, since a sample mean is a random variable. From a statistical viewpoint this distribution tells us what we can expect for the mean when we randomly sample from a population. This sort of knowledge is essential in extracting information from samples. In particular notice that the mean of the sampling distribution in the example is the same as the mean of the population itself $\mu = (2 + 3 + 3 + 6 + 8)/5 = 4.4$ and the variance of the sampling distribution is smaller than the variance of the population ($\sigma^2 = 5.04$). (The exact relationship between $\sigma^2_X$ and $\sigma^2$ is expanded on in Theorem 8.4 below.)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,3</td>
<td>2.5</td>
<td>1/10</td>
</tr>
<tr>
<td>2,3</td>
<td>2.5</td>
<td>1/10</td>
</tr>
<tr>
<td>2,6</td>
<td>4.0</td>
<td>1/10</td>
</tr>
<tr>
<td>2,8</td>
<td>5.0</td>
<td>1/10</td>
</tr>
<tr>
<td>3,3</td>
<td>3.0</td>
<td>1/10</td>
</tr>
<tr>
<td>3,6</td>
<td>4.5</td>
<td>1/10</td>
</tr>
<tr>
<td>3,8</td>
<td>5.5</td>
<td>1/10</td>
</tr>
<tr>
<td>3,8</td>
<td>5.5</td>
<td>1/10</td>
</tr>
<tr>
<td>6,8</td>
<td>7.0</td>
<td>1/10</td>
</tr>
</tbody>
</table>

Table 8.1: List of samples in Example 8.1.

<table>
<thead>
<tr>
<th>Mean</th>
<th>2.5</th>
<th>3.0</th>
<th>4.0</th>
<th>4.5</th>
<th>5.0</th>
<th>5.5</th>
<th>7.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 8.2: Sampling distribution of the mean in Example 8.1.
8.2 Sampling distributions

**Definition 8.1** A statistic is a real-valued function of the observations in a sample.

In general terms, if \( X_1, X_2, \ldots, X_n \) represents a numerical sample of size \( n \) from some population, then \( T = g(X_1, X_2, \ldots, X_n) \) represents a statistic. Examples of statistics are the sample mean, standard deviation, variance and proportion. The individual observations in a sample are also statistics. Hence, for example, the minimum (usually denoted \( \text{min} \) or \( X_{[1]} \)), maximum, and the range \( \text{max} - \text{min} \) are statistics. In fact, so is the first observation \( X_1 \) or even just the number 2 (or any other number), which can be thought of as a degenerate function of the observations.

The point is, a rv is a numeric variable which is fully determined by the values of all, some or even no members of the sample.

**Definition 8.2** The sampling distribution of a statistic is the theoretical probability distribution of the statistic associated with all possible samples of a particular size drawn from a particular population.

Notice that this definition is not confined to random samples. In practically all applications though it is assumed that the sample involved is random in some sense. Random sampling imposes a probability structure on the possible values of the statistic which enables us to define the sampling distribution.

**8.2.1 Random sampling**

For our purposes a random sample is defined as follows:

**Definition 8.3** The set of random variables \( X_1, X_2, \ldots, X_n \) is said to be a random sample of size \( n \) from a population with df \( F_X(x) \) if the \( X_i \) are identically and independently distributed (iid) with df \( F_X(x) \).
In this definition, the distribution function has been used to cater for both
discrete and continuous rvs.

It should be pointed out that this definition of a random sample is stan-
dard in theoretical statistics. In the practice of statistics however there are
many sampling designs involving randomness; eg simple random sampling,
stratified sampling, systematic sampling. These are often said to give rise
to ‘random samples’ even though the samples produced don’t necessarily
satisfy Definition 8.3. It is usually clear from the context whether the strict
or loose meaning of ‘random sample’ is intended.

When a sample of size \( n \) is assumed to be chosen ‘at random’ without re-
placement from a population of size \( N \), does this sample constitute a random
sample in the sense of Definition 8.3?

The strict answer is no. Even though each member of the sample has the
same distribution as the population, the members of the sample are not
independent. For example, in Example 8.1, consider say the second member
of the sample, \( X_2 \). We have that \( P (X_2 = 2) = 0.2 \) (why is this?) but that
\( P (X_2 = 2 \mid X_1 = 2) = 0 \). Therefore \( P (X_2 = 2 \mid X_1 = 2) \neq P (X_2 = 2) \) and
so \( X_2 \) and \( X_1 \) are dependent.

Sampling at random without replacement however does approximate a ran-
dom sample if the sample size \( n \) is small compared to the population size \( N \).
In this case, the impact of removing some members of the population has
minimal impact on the distribution of the population remaining and we can
argue the observations are approximately independent. This typifies many
practical sampling situations.

Note also in Example 8.1, if the sampling is with replacement, we have that
\( P (X_2 = 2 \mid X_1 = 2) = P (X_2 = 2) = P (X_1 = 2) \) and a similar statement can
be made for all sample members and all values of the population. We can
conclude that in this case the sample is a random sample according to the
definition.

So sampling at random with replacement gives rise to a random sample and
sampling at random without replacement approximates a random sample
provided \( n \ll N \). Sampling without replacement when the sample size
is not much smaller than the population size does not give rise to a ran-
dom sample. This situation is often referred to as sampling from a finite
population.

Sampling at random with replacement is also known as repeated sampling. A
typical example is a sequence of independent trials (Bernoulli or otherwise)
such as occur when a coin or die is tossed repeatedly. Each trial involves
random sampling from a population, which in the case of a coin comprises
\( H \) and \( T \) with equal probability.
It should be noted that sampling from a continuous distribution is really a theoretical concept in which the population is assumed uncountably infinite. Consequently sampling at random in this situation is assumed to produce a random sample.

### 8.2.2 The sampling distribution of the mean

Some powerful results exist describing the sampling distribution of the mean of a random sample.

**Theorem 8.4** If \(X_1, X_2, \ldots, X_n\) is a random sample of size \(n\) from a population with mean \(\mu\) and variance \(\sigma^2\), then the sample mean \(\bar{X}\) has a sampling distribution with mean \(\mu\) and variance \(\sigma^2/n\).

**Proof** This is a direct application of Corollary 6.22 with \(a_i = 1/n\).

Theorem 8.4 applies to a random sample as defined in 8.3. The distribution of the population from which the sample is drawn is irrelevant.

It is interesting to note what happens if a sample is randomly selected without replacement from a finite population as described in 8.2.1.

It turns out in this situation that \(\mu_{\bar{X}}\) still equals \(\mu\) (because this result does not depend on the observations being independent) but

\[
\text{var}(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}
\]

which is smaller than \(\frac{\sigma^2}{n}\) by the factor \(\frac{N-n}{N-1}\), which is known as the finite population correction factor.

**Example 8.2** In Example 8.1 we have that \(\mu = 4.4\) and \(\sigma^2 = 5.04\) for the population.

If sampling is done without replacement, the sampling distribution of the mean has mean \(\mu_{\bar{X}} = \mu = 4.4\) and variance \(\text{var}(\bar{X}) = 1.89 = \frac{5.04}{2} \frac{5-2}{5-1}\) in agreement with the note following Theorem 8.4.

If sampling is done with replacement we see \(\mu_{\bar{X}} = \mu = 4.4\) and \(\text{var}(\bar{X}) = 2.52 = \frac{5.04}{2}\) in agreement with the theorem.
Theorem 8.4 describes the location and spread of the sampling distribution of the mean. Can we say anything about its shape? Wouldn’t that depend on the shape of the population distribution?

Well, yes, but surprisingly only when the sample is small. For large samples it turns out that it doesn’t as we see in Section 8.4.

8.3 Sampling distributions related to the normal distribution

The normal distribution is a model of many naturally occurring phenomena. Consequently it is particularly useful to study the sampling distributions of statistics of random samples from normal populations.

Firstly a useful result.

Theorem 8.5 Let \(X_1, X_2, \ldots, X_n\) be a set of independent rvs where \(X_i \sim N(\mu_i, \sigma_i^2)\). Define the linear combination \(Y\) as

\[Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n\]

Then \(Y\) is distributed \(N(\Sigma a_i\mu_i, \Sigma a_i^2\sigma_i^2)\).

Proof In Theorem 5.4 we showed the mgf of the rv \(X_i\) to be

\[M_{X_i}(t) = \exp\left\{\mu_i t + \frac{1}{2} \sigma_i^2 t^2\right\} \quad i = 1, 2, \ldots, n.\]

So, for a constant \(a_i\),

\[M_{a_iX_i}(t) = M_{X_i}(a_it) \quad \text{(using Theorem 3.3 with} \quad \beta = 0, \quad \alpha = a_i)\]

\[= \exp\left\{\mu_i a_it + \frac{1}{2} \sigma_i^2 a_i^2 t^2\right\}.\]

Since the mgf of a sum of independent rv’s is equal to the product of their mgf’s we have

\[M_Y(t) = \prod_{i=1}^{n} \exp\left\{\mu_i a_it + \frac{1}{2} \sigma_i^2 a_i^2 t^2\right\}\]

\[= \exp\left\{t\Sigma a_i\mu_i + \frac{1}{2} t^2 \Sigma a_i^2 \sigma_i^2\right\}.\]
This is of the form of the mgf of a normal rv with mean, \( E(Y) = \Sigma a_i \mu_i \) and variance, \( \text{var}(Y) = \Sigma a_i^2 \sigma_i^2 \).

The sampling distribution of the sum and mean of a random sample from a normal population follows directly from Theorem 8.5.

**Corollary 8.6** Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from \( N(\mu, \sigma^2) \). Define the sum \( S \) and mean respectively as

\[
S = X_1 + X_2 + \cdots + X_n \\
\overline{X} = (X_1 + X_2 + \cdots + X_n)/n
\]

Then \( S \sim N(n\mu, n\sigma^2) \) and \( \overline{X} \sim N(\mu, \sigma^2/n) \).

**Proof** The proof is left as an exercise.

Corollary 8.6 (Theorem 7.1 in WMS) relies only on the population from which the sample is drawn being normally distributed and on the properties of expectation. It forms the basis for inference about the population mean of a normal distribution with known variance.

**Example 8.3** Suppose that envelopes are counted out in packets of 25 by weighing them, and that the weight of an envelope is distributed normally with mean 3 gm and standard deviation 0.6 gm. Any weighed pile of envelopes is ‘counted’ as 25 if it weighs between 70 and 80 gm. What is the probability that a pile of 25 will not be counted as such?

Let \( X_i \) be the weight of the \( i \)th envelope. Then \( X_i \) is distributed \( N(3,0.36) \). Let \( Y = X_1 + X_2 + \cdots + X_{25} \). Then \( Y \) is distributed \( N(75,9) \). We require \( 1 - P(70 < Y < 80) = 0.096 \) where the probability was evaluated in R,

\[
> 1 - \text{diff( pnorm(q=c(70,80),mean=75,sd=3))} \\
[1] 0.0955807
\]
8.3. Sampling distributions related to the normal distribution

Example 8.4 The I.Q.’s for a large population of 10 year-old boys (assumed normally distributed) were determined and found to have a mean of 110 and a variance of 144. How large a sample would have to be taken in order to have a probability of 0.9 that the mean I.Q. of the sample would not differ from the expected value 110 by more than 5?

This example relies upon the standard normal. Let \( X_i \) be the I.Q. of the \( i \)th boy. Then \( X_i \) is distributed \( N(110, 144) \). Consider a sample of size \( n \) and let \( \bar{X} = \sum_{i=1}^{n} X_i / n \). Then \( \bar{X} \) is distributed \( N(110, 144/n) \).

We have

\[
P \left( \left| \bar{X} - 110 \right| \leq 5 \right) = .90.
\]

That is

\[
P \left( \frac{\left| \bar{X} - 110 \right| \sqrt{n}}{\sqrt{12}} \leq \frac{5 \sqrt{n}}{12} \right) = P \left( Z \leq \frac{5 \sqrt{n}}{12} \right) = .90.
\]

For the standard normal, the 5 and 95’tile are \(-1.64, 1.64\).

\[
> \text{qnorm(p=0.95)}
\]

[1] 1.644854

\[5 \sqrt{n}/12 = 1.645 \text{ so } n = 16.\]

Another strategy is to plot the probability for a grid of values of \( n \) and then identify the value of \( n \) for which \( P \left( \left| \bar{X} - 110 \right| \leq 5 \right) = .90\)

\[
\begin{align*}
px & \leftarrow \text{numeric(n)} \\
\text{for (i in seq(along=n))}{
px[i] & \leftarrow \text{diff( pnorm(q=c(105,115),mean=110,sd=12/sqrt(n[i]) ) )}
}
\end{align*}
\]

plot(px ~ n,type='l',las=1)
abline(h=0.9,lty=2)
Example 8.5  A certain product involves a plunger fitting into a cylindrical tube and the 2 items are manufactured in such a way that the diameter of the plunger can be considered a normal rv with mean 2.1 cm and s.d. 0.1 cm while the inside diameter of the cylindrical tube is a normal rv with mean 2.3 cm and s.d. 0.05 cm. For a plunger and tube chosen randomly from a day’s production run, find the probability that the plunger will not fit into the cylinder.

Let $X,Y$ be the diameter of the plunger and cylinder respectively. Then $X \sim N(2.1, 0.01)$, $Y \sim N(2.3, 0.0025)$ and we want $P(Y < X)$.

Using $E(X - Y) = 2.3 - 2.1 = 0.2$ and $\text{var}(X - Y) = 0.0025 + 0.01 = 0.0125$,

The distribution of $Y - X$ is $N(0.2, 0.0125)$ so that, $P(Y - X < 0) = 0.037$ where the probability was calculated using R,

```
> pnorm(q=0,mean=0.2,sd=sqrt(0.0125))
[1] 0.03681914
```

We now turn our interest to the sampling distribution of the variance in a normal population.

Firstly a preliminary result which is little more than a rewording of Theorem 7.6.

Theorem 8.7  Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from $N(\mu, \sigma^2)$. Then $\sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2$ has a chi-square distribution with $n$ degrees of freedom.

Proof  Putting $Z_i = \frac{X_i - \mu}{\sigma}$ we have

$$\sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^{n} Z_i^2$$

But $Z_i$ is the standardised version of $X_i$ and has a standard normal $N(0, 1)$ distribution. Also the rvs $Z_i$ are independent because the $X_i$ are independent ($i = 1, 2, \ldots, n$). The required result follows directly from Theorem 7.6.

♠
The key result describing the distribution of the sample variance
\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \] (8.2)
follows.

**Theorem 8.8** Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from \( N(\mu, \sigma^2) \).
Then
\[ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \]

**Partial proof** We have, using (8.2),
\[ \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \]
which looks much like the conditions for Theorem 8.7. The difference is that \( \bar{X} \) replaces \( \mu \). This difference is important. We can write
\[ \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X} + \mu - \mu}{\sigma} \right)^2 \]
\[ = \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 - n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 \]
(Check the algebra for yourself.) This expression can be rewritten
\[ T = U + V \]
where
\[ T = \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2, \quad U = n \left( \frac{\bar{X} - \mu}{\sigma^2} \right)^2, \quad V = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \]
Now \( T \sim \chi^2(n) \) (by Theorem 8.7) and \( V \sim \chi^2(1) \) because \( V \) is the square of a standard normal rv (since \( \bar{X} \sim N(\mu, \sigma^2/n) \)). Therefore the mgf of \( T \) is \( (1 - 2t)^{-n} \) and the mgf of \( V \) is \( (1 - 2t)^{-1} \). It follows that, provided \( T \) and \( V \) are independent, the mgf of \( T \) is the product of the mgfs of \( U \) and \( V \). The mgf of \( U \) is therefore
\[ M_U(t) = \frac{M_T(t)}{M_V(t)} = (1 - 2t)^{n-1} \]
which is the mgf of the chi-square distribution with \( n - 1 \) df.
(This is a partial proof because we have not proven the independence of \( U \) and \( V \).)
Example 8.6 A random sample of size 10 is selected from the $N(20, 5)$ distribution. What is the probability that the variance of this sample exceeds 10?

By Theorem 8.8,

$$\frac{(n-1)S^2}{\sigma^2} = \frac{9S^2}{5} \sim \chi^2(9)$$

Therefore

$$P(S^2 > 10) = P\left(\frac{9S^2}{5} > \frac{9 \times 10}{5}\right) = P(\chi^2 > 18)$$

where $\chi^2 \sim \chi^2(9)$. Using R we find $P(S^2 > 10) = 0.03517$.

$$> \text{pchisq}(q=18, df=9, lower.tail=F)$$

[1] 0.03517354

8.3.1 The $t$ distribution

The basis of the independence assumed in the proof of Theorem 8.8 is the rather surprising result that the mean and variance of a random sample from a normal population are independent rvs as stated in the following theorem.

Theorem 8.9 Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from $N(\mu, \sigma^2)$. Then the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

are independent.

Proof See DGS, Section 7.3 or WMS, Exercise 13.83. This proof is not examinable.

A further distribution that derives from sampling a normal population is the $t$-distribution.
8.3. Sampling distributions related to the normal distribution

**Definition 8.10** A continuous random variable $X$ with probability density function

$$f_X(x) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\pi\nu} \Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} \text{ for all } x$$

(8.3)

is said to have a $t$ distribution with parameter $\nu (> 0)$. The parameter is known as the degrees of freedom. We write $X \sim t(\nu)$.

The pdf of the $t$ distribution is very similar to the standard normal distribution, being bell-shaped and symmetric about zero. The variance is greater than one however and is dependent on $\nu$ as shown in the following theorem.

**Theorem 8.11** If $X \sim t(\nu)$ then

1. For $\nu > 1$, $E(X) = 0$
2. For $\nu > 2$, var($X$) = $\frac{\nu}{\nu - 2}$
3. The mgf does not exist because only the first $\nu - 1$ moments exist.

**Proof** See WMS, Exercise 7.12 or DGS, Exercise 1, p409. This proof is not examinable.

♠

It can be shown, although we won’t prove it, that as $\nu \to \infty$, the $t$ distribution converges to the standard normal.

The usefulness of the $t$-distribution derives from the following result.

**Theorem 8.12** Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from $N(\mu, \sigma^2)$. Then the rv

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

follows a $t(n - 1)$ distribution where $\overline{X}$ is the sample variance and $S^2$ is the sample variance as defined in (8.2).

**Partial proof** The statistic $T$ can be re-expressed as

$$T = \frac{X - \mu}{S/\sqrt{n}}$$

$$= \left(\frac{X - \mu}{\sigma/\sqrt{n}}\right) \frac{1}{\sqrt{\frac{(n-1)S^2}{(n-1)}}}$$

$$= \frac{Z}{Y/(n-1)}$$
where $Z$ is a $N(0, 1)$ variable and $Y$ is a chi-square variable with $(n - 1)$ df.
The derivation of the pdf of $\frac{Z}{\sqrt{Y/(n-1)}}$ is addressed in DGS, Section 7.4 and WMS, Exercise 7.56. This proof is not examinable.

Notice that $T$ represents a standardised version of the sample mean and because of this is an important statistic in statistical inference. You will have seen it’s use in examples such as the following where the behaviour of a sample mean from a normal population (or one that approximates it) is of interest.

**Example 8.7** A random sample $21, 18, 16, 24, 16$ is drawn from a normal population with mean of 20.

(a) What is the value of $T$ for this sample.

(b) In random samples from this population, what is the probability that $T$ is less than the value found in (a)?

(a) From the sample $\overline{x} = 19.0$ and $s^2 = 12.0$. Therefore $t = \frac{19.0 - 20}{\sqrt{12.0}/5} = -0.645$. (Notice we have used lower-case symbols for specific values of statistics and upper-case for the rvs, but don’t lose any sleep over the distinction!)

(b) Interest here is in $P(T < -0.645)$ where $T \sim t(4)$. From R

$$\begin{array}{l}	ext{pt(q=-0.645,df=4)} \\
\uparrow \uparrow \\
0.2770289 \\
\end{array}$$

$$\begin{array}{l}
y <- c(21,18,16,24,16) \\
y < t.test(y,mu=20,alternative="less") \\
data: y \\
t = -0.6455, \; df = 4, \; p-value = 0.2769 \\
\end{array}$$
8.3.2 The $F$ distribution

Definition 8.13 introduces a distribution which describes the ratio of two sample variances from normal populations and is used in inferences concerning the comparison of two variances. This distribution is also used in analysis of variance, a technique used to test the equality of several means.

Definition 8.13 A continuous random variable $X$ with probability density function

$$f_X(x) = \frac{\Gamma((\nu_1 + \nu_2)/2)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}x^{(\nu_1/2)-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(\nu_1x + \nu_2)^{\nu_1+\nu_2/2}} \text{ for } x > 0 \quad (8.4)$$

is said to have a $F$ distribution with parameters $\nu_1$ ($> 0$) and $\nu_2$ ($> 0$). The parameters are known respectively as the numerator and denominator degrees of freedom. We write $X \sim F(\nu_1, \nu_2)$.

The basic properties of the $F$ distribution are as follows.

**Theorem 8.14** If $X \sim F(\nu_1, \nu_2)$ then

1. For $\nu > 2$, $E(F) = \frac{\nu_2}{\nu_2 - 2}$

2. For $\nu_2 > 4$, $\text{var}(F) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$

3. The mgf does not exist.

**Proof** Not covered. ♠

**Theorem 8.15** Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample of size $n_1$ from $N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be an independent random sample of size $n_2$ from $N(\mu_2, \sigma_2^2)$. Then the rv

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

follows a $F(n_1 - 1, n_2 - 1)$ distribution.
Partial proof  The $F$ statistic can be rewritten as

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{U_1/\nu_1}{U_2/\nu_2}$$

where $U_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2}$, $\nu_1 = n_1 - 1$, $U_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$, and $\nu_2 = n_2 - 1$.

$U_1$ and $U_2$ have chi-square distributions and the pdf of $\frac{U_1/\nu_1}{U_2/\nu_2}$ is described in WMS, Exercise 7.57 or DGS, Section 8.7. This proof is not examinable.

$F$ probabilities are calculated in R with \texttt{pf(q= , df1 = , df2 = )}.

Example 8.8  Suppose $X \sim F(2, 10)$. Determine

(a) $P (X < 1)$
(b) $x$ such that $P (X > x) = 0.01$

Solution

(a) $P (X < 1) = 0.6$.

\[
> \text{pf(q=1,df1=2,df2=10,lower.tail=T)}
\]

[1] 0.5981224

(b) The upper 1% quantile is $x_{0.01} = 7.56$.

\[
> \text{qf(p=0.01,df1=2,df2=10,lower.tail=F)}
\]

[1] 7.559432
8.4 The Central Limit Theorem

Sampling distributions for various statistics of interest are well developed when sampling is from a normal distribution as shown in Section 8.3. Although these results are important, the question remains, what say we don’t know, as is usually the case, the distribution of the population from which the random sample is drawn?

In Section 8.2.2 general results are given describing the mean and variance of the sample mean which hold for any population distribution. Can we say more than this about the sampling distribution of the mean in general?

The answer is yes of course—why else would we mention it! The main result is contained in the following theorem which is so important that it is given the grand title, The Central Limit Theorem or CLT for short.

**Theorem 8.16** Let \( X_1, X_2, \ldots, X_n \) be a random sample from a distribution with mean \( \mu \) and variance \( \sigma^2 \). Then the rv \( Z_n \) defined by

\[
Z_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}
\]

converges in distribution to a standard normal variable.

**Proof** Let the rv \( X_i (i = 1, \ldots, n) \) have mgf \( M_{X_i}(t) \). Then

\[
M_{X_i}(t) = 1 + \mu_1 t + \mu_2 t^2 / 2! + \ldots.
\]

But

\[
\bar{X} = \frac{1}{n} X_1 + \cdots + \frac{1}{n} X_n
\]

so

\[
M_{\bar{X}}(t) = \prod_{i=1}^{n} M_{X_i/n}(t) = [M_{X_i}(t/n)]^n.
\]

Now \( Z_n = \frac{\sqrt{n} \bar{X}}{\sigma} - \frac{\sqrt{n} \mu}{\sigma} \), which is of the form \( Y = \alpha X + b \) so

\[
M_{Z_n}(t) = e^{-\sqrt{n} \mu t/\sigma} M_{\bar{X}}(\sqrt{n} t/\sigma)
\]

\[
= e^{-\sqrt{n} \mu t/\sigma} \left[ M_{X_i} \left( \frac{\sqrt{n} t}{\sigma n} \right) \right]^n
\]

\[
= e^{-\sqrt{n} \mu t/\sigma} \left[ M_{X_i}(t/\sigma \sqrt{n}) \right]^n
\]
Then
\[
\log M_{Z_n}(t) = \frac{-\sqrt{n} \mu t}{\sigma} + n \log \left[ 1 + \frac{\mu'_1 t}{\sigma \sqrt{n}} + \frac{\mu'_2 t^2}{2n \sigma^2} + \frac{\mu'_3 t^3}{3! n \sqrt{n} \sigma^3} + \ldots \right]
\]
\[
= \frac{-\sqrt{n} \mu t}{\sigma} + n \left[ \frac{\mu'_1 t}{\sigma \sqrt{n}} + \frac{\mu'_2 t^2}{2n \sigma^2} + \frac{\mu'_3 t^3}{6n \sqrt{n} \sigma^3} + \ldots \right] - \frac{n}{2} \left[ \frac{\mu'_1 t}{\sigma \sqrt{n}} + \ldots \right]^2 + \frac{n}{3} \left[ \ldots \right]^3 - \ldots
\]
\[
= \frac{\mu'_2 t^2}{2\sigma^2} - \frac{(\mu'_1)^2 t^2}{2\sigma^2} + \text{terms in } (1/\sqrt{n}), \text{ etc.}
\]
Now, \( \lim_{n \to \infty} \log M_{Z_n}(t) = \frac{(\mu'_2 - (\mu'_1)^2) t^2}{2\sigma^2} = \frac{t^2}{2} \) because the terms in \((1/\sqrt{n})\), etc, go to zero and \( \mu'_2 - (\mu'_1)^2 = \sigma^2 \). Thus \( \lim_{n \to \infty} M_{Z_n}(t) = e^{\frac{t^2}{2}} \), which is the mgf of a \( N(0,1) \) random variable. So \( Z_n \) converges in distribution to a standard normal.

Strictly this is a partial proof in that the concept of convergence in distribution has not been defined and it’s assumed that the mgf of \( X_i \) exists, which does not need to be the case. However, the arguments used in the proof are powerful and worth understanding. The ‘convergence in distribution’ bit is saying that \( Z_n \) approaches normality as \( n \to \infty \). So when \( n \) is ‘large’ we can expect \( Z_n \) to approximate a standard normal distribution and, transforming \( Z_n \) back to the sample mean, we can expect the \( \bar{X} \) to approximate a \( N(\mu, \sigma^2/n) \) distribution.

Example 8.9 A soft-drink vending machine is set so that the amount of drink dispensed is a random variable with a mean of 200 millilitres and a standard deviation of 15 millilitres. What is the probability that the average amount dispensed in a random sample of size 36 is at least 204 millilitres?

Let \( X \) be the amount of drink dispensed in ml. The distribution of \( X \) is not known. However, the mean of \( X \), \( \mu = 200 \), and standard deviation of \( X \), \( \sigma = 15 \), are known. The distribution of the sample mean (average), \( \bar{X} \), can be approximated by the normal distribution with mean of \( \mu = 200 \) and standard error of \( \sigma_{\bar{X}} = \sigma / \sqrt{n} = 15 / \sqrt{36} = 15/6 \), according to the CLT. That is, \( \bar{X} \sim N(200, (15/6)^2) \). Now
\[
P(\bar{X} \geq 204) \approx P_N(\bar{X} \geq 204) = 0.0548
\]
\((P_N(A) \) denotes the probability of event \( A \) involving a rv assumed to be normal in distribution.)
8.4. The Central Limit Theorem

```r
> pnorm(q=204, mean=200, sd=15/6, lower.tail=F)
[1] 0.05479929
```

**Example 8.10** Consider the experiment of throwing a fair die \( n \) times where we observe the sum of the faces showing. For \( n = 12 \), find the probability that the sum of the faces is at least 52.

Let the rv \( X_i \) be the number showing on the \( i \)th throw and define \( Y = X_1 + \cdots + X_{12} \). We want \( P(Y \geq 52) \).

In order to use Theorem 8.16 note that the event ‘\( Y \geq 52 \)’ is equivalent to ‘\( \bar{X} \geq 52/12 \)’ where \( \bar{X} = Y/12 \) is the mean number showing from the 12 tosses. Now the distribution of each \( X_i \) is rectangular with \( P(X_i = x) = 1/6, x = 1, 2, \ldots, 6 \) and \( E(X_i) = 7/2, \text{var}(X_i) = 35/12 \).

It follows that \( E(\bar{X}) = 7/2 \) and \( \text{var}(\bar{X}) = 35/(12 \times 12) \). Then from Theorem 8.16,

\[
P(Y \geq 52) \approx P(\bar{X} \geq 52/12) = 0.0455
\]

```r
> pnorm(q=52/12, mean=7/2, sd=sqrt(35)/12, lower.tail=F)
[1] 0.04548447
```

Example 8.10 can also be solved using a generalisation of the Central Limit Theorem. This generalisation indicates why the normal distribution plays such an important role in statistical theory. It says that a large class of rv’s converge in distribution to the standardized normal.

**Theorem 8.17** Let \( X_1, X_2, \ldots, X_n \) be a sequence of independent rv’s with \( E(X_i) = \mu_i, \text{var}(X_i) = \sigma_i^2 \). Define \( Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n \). Then under certain general conditions, including \( n \) large, \( Y \) is distributed approximately \( N(\sum_i a_i\mu_i, \sum_i a_i^2\sigma_i^2) \).

**Proof** Not part of the course. ♠

Note that if the \( X_i \) are normally distributed then \( Y = \sum_{i=1}^n a_iX_i \) will have a normal distribution for any \( n \), large or small.
Example 8.11  Suppose we have a number of independent noise voltages, say $V_i$, $i = 1, 2, \ldots, n$. Let $V$ be the sum of the voltages and suppose each $V_i$ is distributed $U(0, 10)$. For $n = 20$ find $P(V > 105)$.

This is clearly an example of Theorem 8.17 with each $a_i = 1$. In order to find $P(V > 105)$ we need to know the distribution of $V$. We have $E(V_i) = 5$ and $\text{var}(V_i) = 100/12$ so that from Theorem 8.17, we can say that $V$ is distributed approximately normal with mean $20 \times 5 = 100$ and variance $20 \times 100/12$. That is, $\frac{V - 100}{10\sqrt{5/3}}$ is distributed $N(0,1)$ approximately. So

$$P(V > 105) \simeq P_N(V > 105) = 0.35$$

> pnorm(q=105,mean=100,sd=sqrt(2000/12),lower.tail=F)
[1] 0.3492677

---

8.5 The normal approximation to the binomial

Reading 8.4  DGS, Section 5.8; WMS, Section 7.5.

The normal approximation to the binomial distribution has already been considered in Section 5.2.7; it is seen here again as an application of the Central Limit Theorem.

The essential point to recognise is that a sample proportion is a sample mean. Consider a sequence of independent Bernoulli trials resulting in the random sample $X_1, X_2, \ldots, X_n$ where

$$X_i = \begin{cases} 
0 & \text{if failure} \\
1 & \text{if success}
\end{cases}$$

denotes whether or not the $i$th trial is a success. Then the sum

$$Y = \sum_{i=1}^{n} X_i$$

represents the number of successes in the $n$ trials and

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{Y}{n}$$
is a sample mean representing the proportion or fraction of trials which are successful. In this context we usually denote \( \bar{X} \) by the sample proportion \( \hat{p} \).

Note that \( E(X_i) = p \) and \( \text{var}(X_i) = p(1 - p) \). Therefore

\[
E(Y) = np \quad \text{and} \quad \text{var}(Y) = np(1 - p)
\]

and

\[
E(\bar{X}) = p \quad \text{and} \quad \text{var}(\bar{X}) = \frac{p(1 - p)}{n}
\]

Theorems 8.16 and 8.17 are applicable to \( \bar{X} \) and \( Y \) respectively. Hence

\[
\bar{X} = \hat{p} \sim N\left(p, \frac{p(1 - p)}{n}\right) \text{ approx}
\]

and

\[
Y = n\hat{p} \sim N(np, np(1 - p)) \text{ approx}
\]

### 8.6 Self-assessment exercises

The following exercises are designed to provide practice at problem-solving based on the material in this module. Solutions are provided at the end of the module. Additional exercises are available in the next section and in the textbook.

**Ex. 8.1** A random sample of size 81 is taken from a population with mean 128 and standard deviation 6.3. What is the probability that the sample mean will not fall between 126.6 and 129.4?

**Ex. 8.2** Let \( Y_1, Y_2, \ldots, Y_n \) be \( n \) independent random variables, each with probability density function

\[
f_Y(y) = \begin{cases} 
3y^2 & \text{for } 0 \leq y \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

Determine the probability that the sample mean will be within one standard deviation of the population mean, using the Central Limit Theorem.

**Ex. 8.3** Suppose the weights of eggs in a dozen carton actually have a weight that is normally distributed with mean 59 and variance 0.7.

(a) Find the probability that the sample mean weight will exceed 59.5 grams.
(b) Find the probability that a sample of twelve eggs will produce a sample variance of greater than 1.

**Ex. 8.4** In a carton of a dozen eggs, the number that are broken has a Poisson distribution with mean 0.2. Find the probability that the sample mean of the number of broken eggs per carton is more than one.

**Ex. 8.5** The random variable $M$ has the following probability density function

$$f_M(m) = \begin{cases} 3m^2 & \text{for } 0 < m < 1 \\ 0 & \text{otherwise} \end{cases}$$

A random sample of size 9 is taken from the distribution, and the sample mean $\overline{M}$ is computed.

(a) Compute the mean of $M$.

(b) Compute the variance of $M$.

(c) State the approximate distribution of $\overline{M}$ including the parameters of the distribution.

(d) Compute the probability that the sample mean will be within 0.1 of the true mean.

**Ex. 8.6** (Computer exercise) Simulate a random samples $X_1, X_2, \ldots, X_9$, of size 9 from a $N(2, 36)$ distribution. Obtain the mean and variance of $\sum X_i$ and $\overline{X}$. Verify they are approximately as expected.

**8.7 Exercises**

**Ex. 8.7** Consider the random variable $A$, defined as

$$A = \frac{Z}{\sqrt{W/\nu}},$$

where $Z$ has a standard normal distribution independent of $W$ which has a $\chi^2$ distribution with $\nu$ degrees of freedom. Also, consider the random variable $B$, defined as

$$B = \frac{W_1/\nu_1}{W_2/\nu_2},$$

where $W_1$ and $W_2$ are independent $\chi^2$ variables with $\nu_1$ and $\nu_2$ degrees of freedom respectively.
8.7. Exercises

(a) Write down (do not derive) the distribution of the random variable $A$, including the parameters of the distribution.

(b) Write down (do not derive) the distribution of the random variable $B$, including the parameters of the distribution.

(c) Deduce that the distribution of $A^2$ is a special case of the distribution of $B$, and state the values of the parameters for which this is true.

Ex. 8.8 A manufacturer makes metal tubes which must fit into metal sleeves. The metal tubes have an average diameter of 1 cm with a standard deviation of 0.005 cm. The sleeves have an average interior diameter of 1.01 cm with a standard deviation of 0.004 cm. Both diameters are assumed to be normally distributed.

(a) What is the probability that a randomly selected tube will fit into a randomly selected sleeve?

(b) What is the probability that tube will fit and that the difference in diameters will be less than 0.015 cm?

(c) What is the probability that the average diameter of 10 randomly selected tubes will be less than 1.005 cm?

Ex. 8.9 A manufacturing plant produces 4 tonnes of its product on a particular day with a standard deviation of 0.2 tonnes per day. Find the probability that in a 20 day period the plant will produce less than 78 tonnes of product if daily productions can be assumed independent.

Ex. 8.10 (Computer exercise) Let $Y$ be the change in depth of a river from one day to the next measured (in cms) at a specific location. Assume $Y$ is uniformly distributed for $y \in [-70, 70]$.

(a) Find the probability that the mean change in depth for a period of 30 days will be greater than 10 cms.

(b) Use simulation to estimate the probability in (a).

Ex. 8.11 The number of deaths per year due to typhoid fever is assumed to have a Poisson distribution with rate $\lambda = 4.6$ per year.

(a) If deaths from year to year can be assumed to be independent what is the distribution over a 20 year period?

(b) Find the probability that there will be more than 110 deaths due to typhoid in period of 20 years.

Ex. 8.12 (Computer exercise) The probability that a cell is a lymphocyte is 0.2.
(a) Write down an exact expression for the probability that in a sample containing 150 cells that at least 40 are lymphocytes. Evaluate this expression using R.

(b) Write down an approximate expression for this probability and evaluate it.

Ex. 8.13 (Computer exercise) Suppose that the probability of a person aged 60–64 dying after receiving influenza vaccine is 0.006. In a sample of 200 persons aged 60–64 years:

(a) Write down an exact expression for the probability that more than 5 will die after vaccination for influenza. Evaluate this expression using R.

(b) Write down an approximate expression for this probability and evaluate it.

(c) Simulate (a) using R.

(d) If 4 persons died in the sample of 200 what conclusion would you make about the probability of dying after vaccination?

Ex. 8.14 (Computer exercise) Illustrate the Central Limit Theorem for the uniform distribution on $[-1, 1]$ by simulation and repeated sampling.

Ex. 8.15 (Computer exercise) Demonstrate the Central Limit Theorem using R.

(a) For the uniform distribution, where $-1 < x < 1$, (a symmetric distribution).

(b) For the exponential distribution with parameter 1, (a skewed distribution).

Ex. 8.16 (Computer exercise) Suppose we again have random samples, $X_1, X_2, \ldots, X_9$, of size 9 from a $N(2, 36)$ distribution. Estimate the mean and variance of

$$X_1 + 2X_2 - X_3 + 3X_4 - 2X_5 + X_6 - 4X_7 + 2X_8 - X_9.$$

Verify it is approximately as expected.

Ex. 8.17 (Computer exercise) Generate 100 random samples, $X_1, X_2, X_3, X_4$, of size 4 from a $N(2, 6^2)$ distribution.

(a) Determine the covariance of $(X_1 + X_2 + X_3 + X_4)$ and $(6X_1 + 3X_2 + 3X_3 + 4X_4)$.

(b) Also determine the covariance of $(X_1 + X_2 - X_3 - X_4)$ and $(X_1 - X_2 + X_3 - X_4)$. 
8.8. Some answers and hints

(c) How do the estimated values compare to the theoretical values?

Ex. 8.18 (Computer exercise) Simulate Example 8.3 and so estimate the probability that a pile of 25 envelopes will not be counted as such.

Ex. 8.19 (Computer exercise) Simulate Example 8.5 with R.

Ex. 8.20 (Computer exercise) Use simulation to estimate \( P(V > 105) \) in Example 8.11.

Ex. 8.21 (Computer exercise) Simulate the result of Example 8.10 with R.

8.8 Some answers and hints

8.1 Let the rv be \( X \) so that \( E(X) = 128 \) and \( \text{var}(X) = (6.3)^2 \). Then the distribution of the sample mean will be (by the Central Limit Theorem) \( X \sim N(128, (6.3)^2/81) \). Then

\[
P \left( \frac{126.6 - 128}{6.3/\sqrt{81}} < Z < \frac{129.4 - 128}{6.3/\sqrt{81}} \right) = P(-2 < Z < 2) = 0.9544
\]

So the required answer is

\[
P \left( X < 126.6 \text{ or } X > 129.4 \right) = 1 - 0.9544 = 0.0456.
\]

8.2 The mean of \( Y \) is

\[
E(Y) = \int_0^1 y(3y^2) \, dy = 3/4.
\]

Similarly, \( E(Y^2) = 3/5 \) so that \( \text{var}(Y) = 3/80 \). Hence, by the CLT, \( Y \sim N(3/4, 3/(80n)) \). Therefore

\[
P \left( \frac{3/4 - \sqrt{3/80}}{\sqrt{3/80}} < Y < \frac{3/4 + \sqrt{3/80}}{\sqrt{3/80}} \right) = P \left( 0.56 < Y < 0.94 \right)
\]

is required. So for a sample of size \( n \),

\[
P \left( \frac{0.56 - 0.75}{0.19/\sqrt{n}} < Z < \frac{0.56 - 0.75}{0.19/\sqrt{n}} \right) = P \left( -\sqrt{n} < Z < \sqrt{n} \right)
\]

which approaches one as \( n \to \infty \). For example, if \( n = 10 \),

\[
P \left( -\sqrt{10} < Z < \sqrt{10} \right) = 0.9984.
\]
8.3 (a) Let the weight be \( E \), so that \( E \sim N(59, 0.7) \). Then by the CLT, the sample means have the distribution \( \bar{E} \sim N(59, 0.7/12) \). So

\[
P(\bar{E} > 59.5) = P \left( Z > \frac{59.5 - 59}{\sqrt{0.7/12}} \right) = P(z > 2.07) \approx 0.019
\]

(b) We seek \( P(s^2 > 1) \). We know that

\[
\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}
\]

where \( n = 12 \) and \( \sigma^2 = 0.7 \). So

\[
P(s^2 > 1) = P \left( \frac{11s^2}{0.7} > \frac{11 \times 1}{0.7} \right)
= P(\chi^2_{11} > 15.714)
\approx 0.152
\]

8.4 Let the number broken be \( B \), so \( B \sim \text{Pois}(0.2) \). The sample mean number broken will have the distribution \( \bar{B} \sim N(0.2, 0.2/\sqrt{12}) \) approx. So

\[
P(B \geq 1) = P \left( Z > \frac{1 - 0.2}{\sqrt{0.2/\sqrt{12}}} \right) = P(Z > 6.196) = 0.
\]

In contrast, in any one carton, the probability of more than one broken egg is

\[
P(B > 1) = 1 - P(B = 0) = 1 - 0.8187 = 0.181
\]

using the Poisson distribution.

8.5 (a) \( E(M) = 3/4 \) (b) \( \text{var}(M) = 3/80 \) (c) \( M \sim N(3/4, 1/240) \) approx (d) 0.8788

8.6 Method: Generate 9 random samples of size 100 from a \( N(2, 36) \) distribution and put them in 9 columns of an array with 100 rows. We now consider each row to be a random sample of size 9 (so we have 100 random samples of size 9). Use the \texttt{apply} function to get the row sums and means.

\[
\begin{align*}
sim<-100 \\
samp<-9 \\
sim.mat<-matrix(0,nrow=nsim,ncol=nsamp)
for ( i in 1:9) {
  sim.mat[,i]<-rnorm(mean=2,sd=6,n=100)
}
row.totals<-apply(sim.mat,MARGIN=1,FUN=sum)
\end{align*}
\]
8.8. Some answers and hints

row.means<-apply(sim.mat,MARGIN=1,FUN=mean)
results<-array(0,dim=c(1,4))
results<-c(mean(row.totals),var(row.totals),mean(row.means),var(row.means))

8.8 (a) 0.94 (b) 0.7686 (This is a conditional probability!) (c) 0.9921

8.9 Let $Y = X_1 + X_2 + \cdots + X_{20}$ and use Theorem 8.17. Ans=.01255.

8.10 (a) Use Theorem 8.16  (b) Use Theorem 8.17

8.11 (a) A sum of Poisson variates is also Poisson. (b) 0.0303

8.12 (a) Use pbinom.  (b) Use an approximation to the binomial distribution. Ans = 0.0262

8.13 (a) Use pbinom.  (b) Use an approximation to the binomial distribution. Ans = 0.00143 (c) Use rbinom to simulate 1000 outcomes. Tabulate the results.

8.14 METHOD: Generate a random sample of size 100 from a distribution which is uniform on $[-1,1]$. Draw a density of the random sample and test it for normality. Generate a second random sample of size 100, sum the two and plot the density of the sum. Test the sum for normality. Continue the process, adding a third, fourth, ..., sample to the sum. At each step draw the density and test the sum for normality. Continue until it appears you have achieved convergence to a normal distribution. (Normality should be achieved reasonably quickly, stop if a satisfactory result has still not been obtained after generating 12 samples.)

Draw a histogram of the initial sample and the final sum.

To test for normality use the qqnorm function to get a normal probability plot. The graph should be an approximate straight line.

samsz<-100
nruns<-5
par(mfrow=c(1,2),oma=c(4,4,8,4))  # sets up the plotting window
u1<-runif(n=samsz,min=-1,max=1)
hist(u1)
qqnorm(u1)
locator(1)  # this means click the left button on the graph to proceed
u2<-u1
for ( i in 1:nruns){
u2<-u2+runif(n=samsz,min=-1,max=1)
hist(u2)

Use the same simulations as in Exercise 8.6. We get the weighted sum by matrix multiplication, e.g.

\[
\begin{pmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
\vdots \\
Y_{100}
\end{pmatrix} =
\begin{pmatrix}
X_{1,1} & X_{1,2} & X_{1,3} & X_{1,4} & X_{1,5} & X_{1,6} & X_{1,7} & X_{1,8} & X_{1,9} \\
X_{2,1} & X_{2,2} & X_{2,3} & X_{2,4} & X_{2,5} & X_{2,6} & X_{2,7} & X_{2,8} & X_{2,9} \\
X_{3,1} & X_{3,2} & X_{3,3} & X_{3,4} & X_{3,5} & X_{3,6} & X_{3,7} & X_{3,8} & X_{3,9} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
X_{100,1} & X_{100,2} & X_{100,3} & X_{100,4} & X_{100,5} & X_{100,6} & X_{100,7} & X_{100,8} & X_{100,9}
\end{pmatrix}
\]

\[
\begin{pmatrix}
W_1 \\
W_2 \\
W_3 \\
W_4 \\
W_5 \\
W_6 \\
W_7 \\
W_8 \\
W_9
\end{pmatrix}
\]

\[
\text{weights} \leftarrow \text{c}(1, 2, -1, 3, -2, 1, -4, 2, -1)
\]

\[
\text{weighted.sum} \leftarrow \text{sim.mat} \times \text{weights} \quad \# \text{matrix multiplication}
\]

\[
\text{print(mean(weighted.sum))}
\]

\[
\text{print(var(weighted.sum))}
\]

We apply the same computing strategies as in Exercises 8.6 and 8.16 but for each simulation, we are generating a matrix of normal rv’s.
for (j in 1:4) X[,j]<-rnorm(mean=2,sd=6,n=8)
#   matrix multiplication to make the weighted sums
Y1<-X %*% c(1,1,1,1)
Y2<-X %*% c(6,3,3,4)
Y3<-X %*% c(1,1,-1,-1)
Y4<-X %*% c(1,-1,1,-1)
R1<-cov(Y1,Y2)       # this is a 2 X 2 covariance matrix
R2<-cov(Y3,Y4)
cov.mat[i,1]<-R1      # row i corresponds to simulation i
cov.mat[i,2]<-R2
}
print(apply(cov.mat,MARGIN=2,FUN=mean))

8.18 nsim<-1000
excl<-0
for ( i in 1:nsim){ sumwt<-sum(rnorm(mean=3,sd=0.6,n=25))
   # use the OR operator - "||"
if (sumwt>=80 || sumwt<=70 ) excl<-excl+1   # too heavy or too light
}
Pexcl<-excl/nsim
print(Pexcl)

8.19 nsim<-500
plungers<-rnorm(mean=2.1,sd=0.1,n=nsim)
cylinders<-rnorm(mean=2.3,sd=0.05,n=nsim)
oversize<-cylinders-plungers
Pnotfit<-sum(oversize<=0)/nsim
print(Pnotfit)
There were 15 of the 500 plungers that did not fit, giving an estimate of 0.03 for the probability that a plunger would not fit.

8.20 nsim<-1000
exceed<-numeric(0)
for ( i in 1:nsim){
   sumVi<-sum(runif(n=20,min=0,max=10))
exceed<-c(exceed,sumVi>105)
}
Pexc<-sum(exceed)/nsim
print(Pexc)
My overall result: I used a sample size of $n = 1000$ by repeating the above 10 times and obtained 346 samples with a sum greater than 105. Hence my estimate of $P(V > 105) = 0.346$. 
8.21 nsim<-1000
nthrows<-12
exceed52<-numeric(0)
for ( i in 1:nsim){
faces<-sample(1:6,replace=T,size=nthrows)
exceed52<-c(exceed52,sum(faces)>=52) # a collection of 1's or 0's
}
Pexc<-sum(exceed52)/nsim
print(Pexc)
Bibliography


Bibliography


