MATH1063: Introduction to Statistics

Dr Helen Ogden, based on original notes by Prof. Sujit Sahu

Contents

Chapter 1

Introduction to Statistics

1.1 What is statistics?

1.1.1 Early and modern definitions

The word *statistics* has its roots in the Latin word *status* which means 'the state', and in the middle of the 18th century was intended to mean *collection, processing and use of data by the state*. With the rapid industrialisation of Europe in the first half of the 19th century, statistics became established as a discipline. This led to the formation of the Royal Statistical Society, the premier professional association of statisticians in the UK and world-wide, in 1834. During this 19th century growth period, statistics acquired a new meaning: the *interpretation of data or methods of extracting information from data for decision making*. Thus statistics has its modern meaning as the methods for *collection, analysis and interpretation of data*. Indeed, the Oxford English Dictionary defines *statistics* as

> The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Note that the word 'state' has gone from its definition. Instead, statistical methods are now essential for anyone wanting to answer questions using data.

For example, will it rain tomorrow? Is smoking harmful during pregnancy? What degree classification will I receive upon graduation? Will the stock market crash tomorrow?

1.1.2 Statistics tames uncertainty

We often have to make decisions under uncertainty. It is useful for us to know the extent of uncertainty, which allows us to minimise the frequency of wrong decisions, or which minimises the loss due to wrong decisions.

Thus we have the equation

> Uncertain knowledge + Knowledge of the extent of uncertainty in it = Usable knowledge.

1.1.3 Why should I study statistics as part of my degree?

Studying statistics will equip you with the basic skills in data analysis and doing science with data. A decent level of statistical knowledge is required no matter what branch of mathematics, engineering, science and social science you will be studying. Learning statistical theories gives you the opportunity

to practice your deductive mathematical skills on real life problems. In this way, you will improve at mathematical methods while studying statistical methods.

> All knowledge is, in final analysis, history.
> All sciences are, in the abstract, mathematics.
> All judgements are, in their rationale, statistics.
> – Prof. C. R. Rao

1.1.4 Lies, Damn Lies and Statistics?

Sometimes people say, "you can prove anything in statistics!" and many such jokes. Such remarks bear testimony to the fact that often statistics and statistical methods are mis-quoted without proper verification and robust justification. This is even more important in recent years of the global pandemic when every day we have been showered with a deluge of numbers.

Statistics can be very much mis-used and mis-interpreted, and as statisticians it is our duty to correctly apply statistical techniques to develop scientifically robust and strong arguments. In particular, we need to be careful about what assumptions we make.

1.1.5 What's in this module?

- **Chapter 1**: We will start with the basic statistics used in everyday life, e.g. mean, median, mode, standard deviation, etc. Statistical analysis and report writing will be discussed. We will also learn how to explore data using graphical methods. We will introduce the R programming language to help us with these tasks.
- **Chapter 2: Introduction to Probability**. We will define and interpret probability as a measure of uncertainty. We will learn the rules of probability and then explore examples.
- **Chapter 3: Probability Distributions**. We will learn about lots of probability distributions, which can be used to model the outcomes of different types of events.
- **Chapter 4: Statistical Inference**. We will discuss basic ideas of statistical inference, including techniques of point and interval estimation and hypothesis testing.

1.2 Example data sets

In this module, we will assume that we have data from n randomly selected sampling units, which we will conveniently denote by $x_1, x_2, \ldots, x_n$. We will assume that these values are numeric, either discrete like counts, e.g. number of road accidents, or continuous, e.g. heights of 4-year-olds, marks obtained in an examination.

We will use the following examples to demonstrate ideas throughout the module:

Example 1.1 (Fast food service time)**.** This dataset contains the service times (in seconds) of customers at a fast-food restaurant. The first row is for customers who were served from 9–10am and the second row is for customers who who were served from 2–3pm on the same day.

AM	38	100	64	43	63	59	107	52	86	77
PM	45	62	52	72	81	88	64	75	59	70

We would like to compare these AM and PM service times.

Example 1.2 (Computer failures)**.** This dataset contains weekly failures of a university computer system over a period of two years.

```
4 0 0 0 3 2 0 0 6 7 6 2 1 11 6 1 2 1 1 2 0 2 2 1 0 12 8 4 5 0 5 4 1 0 8 2 5 2 1
12 8 9 10 17 2 3 4 8 1 2 5 1 2 2 3 1 2 0 2 1 6 3 3 6 11 10 4 3 0 2 4 2 1 5 3 3
2 5 3 4 1 3 6 4 4 5 2 10 4 1 5 6 9 7 3 1 3 0 2 2 1 4 2 13
```

We would like to summarise this data and make predictions about future failures.

Example 1.3 (Weight gain)**.** Is it true that students tend to gain weight during their first year in college? Cornell Professor of Nutrition, David Levitsky, recruited students from two large sections of an introductory health course. Although they were volunteers, they appeared to match the rest of the class in terms of demographic variables such as sex and ethnicity. 68 students were weighed during the first week of the semester, then again 12 weeks later. The first 10 rows of the data are:

Student number	Initial weight (kg)	Final weight (kg)
1	77.56423	76.20346
2	49.89512	50.34871
3	60.78133	61.68851
4	52.16308	53.97745
5	68.03880	70.30676
6	47.17357	48.08075
7	64.41006	67.13162
8	54.43104	56.24541
9	65.31725	67.13162
10	70.76035	69.85317

We would like to explore the data graphically and to test whether the data support the hypothesis that students gain weight during their first year in college.

Example 1.4 (Billionaires)**.** Fortune magazine publishes a list of the world's billionaires each year. The 1992 list includes 225 individuals. Their wealth, age, and geographic location (Asia, Europe, Middle East, United States, and Other) are reported.

The variables in the data are:

- `wealth`: Wealth of family or individual in billions of dollars
- `age`: Age in years (for families it is the maximum age of family members)
- `region`: Region of the World (Asia, Europe, Middle East, United States and Other).

The first 10 rows of the data are:

wealth	age	region
37.0	50	Middle East
24.0	88	United States
14.0	64	Asia
13.0	63	United States
13.0	66	United States
11.7	72	Europe
10.0	71	Middle East
8.2	77	United States
8.1	68	United States
7.2	66	Europe

We will investigate differences in wealth of billionaires by age and region using many exploratory

graphical tools and statistical methods.

1.3 Introduction to R

R is a programming language for statistics. We will use R as a calculator, to summarise data, make exploratory plots, perform statistical analysis, illustrate theorems and calculate probabilities. You will get practice of using R yourself during the computer practical sessions.

R is freely available to download: search "download R" or go to: https://cran.r-project.org/. We will access R via the RStudio integrated development environment, which you can download from https://posit.co/download/rstudio-desktop/ which provides a nice editor for R code alongside a console. Details about using R and RStudio are given in the the practical sessions. If you want to complete the practical sessions on your own computer, please try to install R and RStudio before attending the practical sessions.

We will not give details of using R in these notes: this will all be covered by the practical sessions. In the notes, we provide simple commands used to calculate useful quantities in R, and show some of the graphical outputs we can get from R.

1.4 Summarising data sets

1.4.1 Summarising categorical data

We can summarise categorical (not numeric) data by tables. For example, we could summarise the results of 20 coin tosses as: 12 heads, 8 tails.

For the computer failure data (Example 1.2) we may summarise the count of failures per week:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	17
12	16	21	12	11	8	7	2	4	2	3	2	2	1	1

1.4.2 Measures of location

1.4.2.1 Choosing a representative value for the data

We are seeking a representative value for the data $x_1, x_2, \ldots, x_n$ which should be a function of the data. If a is that representative value then how much error is associated with it? The total error could be the sum of squares of the errors,

$$\text{SSE}(a) = \sum_{i=1}^{n} (x_i - a)^2$$

or the sum of the absolute errors

$$\text{SAE}(a) = \sum_{i=1}^{n} |x_i - a|.$$

What value of a will minimise the SSE or the SAE? For SSE the answer is the sample mean and for SAE the answer is the sample median.

1.4.2.2 The sample mean

The *sample mean* is

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

For the service time data from Example 1.1, the AM mean time is 68.9 seconds and the PM mean time is 66.8 seconds.

Theorem 1.1. *The sample mean $\bar{x}$ minimises the SSE.*

Proof. We have

$$\begin{aligned}
\text{SSE}(a) &= \sum_{i=1}^{n} (x_i - a)^2 \\
&= \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - a)^2 \quad \text{(Add and subtract } \bar{x}\text{)} \\
&= \sum_{i=1}^{n} \left\{ (x_i - \bar{x})^2 + 2\,(x_i - \bar{x})\,(\bar{x} - a) + (\bar{x} - a)^2 \right\} \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2(\bar{x} - a) \sum_{i=1}^{n} (x_i - \bar{x}) + \sum_{i=1}^{n} (\bar{x} - a)^2 \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2,
\end{aligned}$$

since $\sum_{i=1}^{n} (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$.

The first term is free of a and the second term is non-negative for any value of a. Hence the minimum occurs when the second term is zero, i.e. when $a = \bar{x}$. □

We have established the fact that the sum of (or mean) squares of the deviations from any number a is minimised when a is the mean. This justifies why we often use the mean as a representative value.

1.4.2.3 The sample median

The *sample median* is the middle value in the ordered list of observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. For the AM service time data

$$38 < 43 < 52 < 59 < 63 < 64 < 77 < 86 < 100 < 107.$$

If n is odd, there is a unique middle value. If n is even, there are two middle values (63 and 64 for the AM service time data). Any value between those two middle values is a sample median. By convention, we often use the mean of these values (63.5 for the the AM service time data).

Theorem 1.2. *The sample median minimises the SAE.*

Proof. If $a < x_{(1)}$, then

$$\text{SAE}(a) = \sum_{i=1}^{n} (x_i - a). \tag{1.1}$$

As a increases, each term of (1.1) decreases until a reaches $x_{(1)}$, so $\text{SAE}(x_{(1)}) < \text{SAE}(a)$ for all $a < x_{(1)}$. We conclude the minimiser of the SAE is at least $x_{(1)}$.

Now suppose $x_{(k)} \leq a < x_{(k+1)}$. Then

$$\begin{aligned}
\text{SAE}(a) &= \sum_{i=1}^{k} (a - x_{(i)}) + \sum_{i=k+1}^{n} (x_{(i)} - a) \\
&= (2k - n)a - \sum_{i=1}^{k} x_{(i)} + \sum_{i=k+1}^{n} x_{(i)}.
\end{aligned}$$

The term $-\sum_{i=1}^{k} x_{(i)} + \sum_{i=k+1}^{n} x_{(i)}$ is constant for each any a in the interval $[x_{(k)}, x_{(k+1)})$. So the SAE in this interval is a straight line, with slope $2k - n$. This slope is negative if $k < \frac{n}{2}$, zero if $k = \frac{n}{2}$, and positive if $k > \frac{n}{2}$.

For each $k < \frac{n}{2}$, the SAE is decreasing in the interval $[x_{(k)}, x_{(k+1)})$, and we conclude the minimiser of the SAE is at least $x_{(k+1)}$. Starting at $k = 1$, we continue increasing k by one and concluding the the minimiser of the SAE is at least $x_{(k+1)}$, until we reach a k such that $k \geq \frac{n}{2}$:

- If n is odd, this happens at $k = \frac{n+1}{2}$. In that case, since $k > \frac{n}{2}$, the SAE is increasing in the interval $[x_{(k)}, x_{(k+1)})$, so we conclude the SAE is minimised at $x_{(k)} = x_{\left(\frac{n+1}{2}\right)}$, the median point.
- If n is even, this happens at $k = \frac{n}{2}$. In that case the SAE is constant in the interval $[x_{(k)}, x_{(k+1)})$, so we conclude that the SAE is minimised at any a between the two middle points $x_{\left(\frac{n}{2}\right)}$ and $x_{\left(\frac{n}{2}+1\right)}$, i.e. any median point.

□

We have established the fact that the sum of (or mean) of the absolute deviations from any number a is minimised when a is the median. This justifies why median is also often used as a representative value.

The mean gets more affected by extreme observations than the median. For example for the AM service times, suppose the next observation is 190. The median will be 64 instead of 63.5 but the mean will shoot up to 79.9.

1.4.2.4 The sample mode

The mode or the most frequent (or the most likely) value in the data is taken as the most representative value if we consider a 0-1 error function instead of the SAE or SSE above. Here, one assumes that the error is 0 if our guess a is the correct answer and 1 if it is not. It can then be proved that the best guess a will be the mode of the data.

1.4.3 Measures of spread

A quick measure of the spread is the *range*, which is defined as the difference between the maximum and minimum observations. For the AM service times the range is 69 (107 − 38) seconds.

The *variance* is

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

We have

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

so we calculate variance as

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

Sometimes the variance is defined with the divisor n instead of $n - 1$. We have chosen $n - 1$ since this is the default in R. We will return to this in Chapter 4.

The *standard deviation* (sd) is the square root of variance

$$s = \mathrm{sd}(x) = \sqrt{\mathrm{Var}(x)}.$$

The standard deviation for the AM service times is 23.2 seconds. Note that it has the same unit as the observations.

The *interquartile range* (IQR) is the difference between the third, Q_3 and first, Q_1 quartiles, which are respectively the observations ranked $\frac{1}{4}(3n+1)$ and $\frac{1}{4}(n+3)$ in the ordered list. Note that the median is the second quartile, Q_2. When n is even, definitions of Q_3 and Q_1 are similar to that of the median, Q_2. The IQR for the AM service times is $83.75 - 53.75 = 30$ seconds.

1.4.4 Summarising data in R

We can calculate the mean, median, variance and standard deviation very easily in R.

As an example, suppose we have stored the weekly counts of computer failures from Example 1.2 in an object called `compfail` in R. You will find out how to read the data into R in the practical sessions.

Then we can calculate these quantities with:

```
mean(compfail)
```

```
## [1] 3.75
```

```
median(compfail)
```

```
## [1] 3
```

```
var(compfail)
```

```
## [1] 11.43204
```

```
sd(compfail)
```

```
## [1] 3.38113
```

R saves us a lot of effort here relative to computing these quantities by hand.

The `summary` command provides several using summary statistics:

```
summary(compfail)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    3.00    3.75    5.00   17.00
```

We can find a table summarising the counts by typing

```
table(compfail)
```

```
## compfail
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 17
## 12 16 21 12 11  8  7  2  4  2  3  2  2  1  1
```

1.5 Exploratory data plots

1.5.1 Introduction

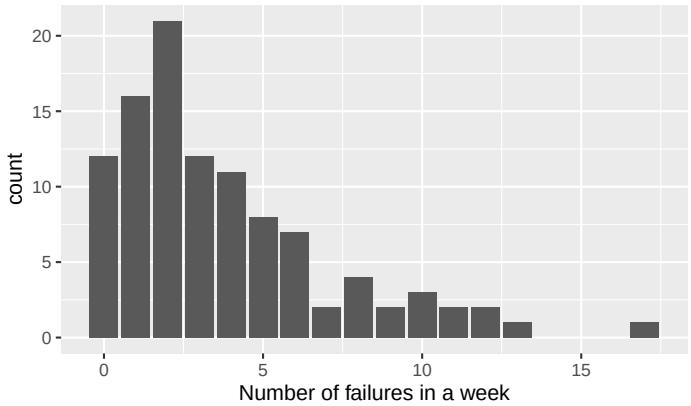
Often the clearest way to generate insight is to plot the data. Plots may be used to check the shape of the distribution of a variable, and to investigate the relationship between different variables in a dataset.

The type of plot which is most appropriate depends on whether we are looking at the distribution of a single variables or the relationship between two or more variables, and on whether the variables of interest are discrete or continuous.

Here, we will show some plots produced by R. You will find out how to make these plots in R in the practical sessions.

1.5.2 Distribution of a single discrete variable

To explore the distribution of a single discrete variable, we may use a bar plot. For instance, we may plot the weekly counts of computer failures from Example 1.2:

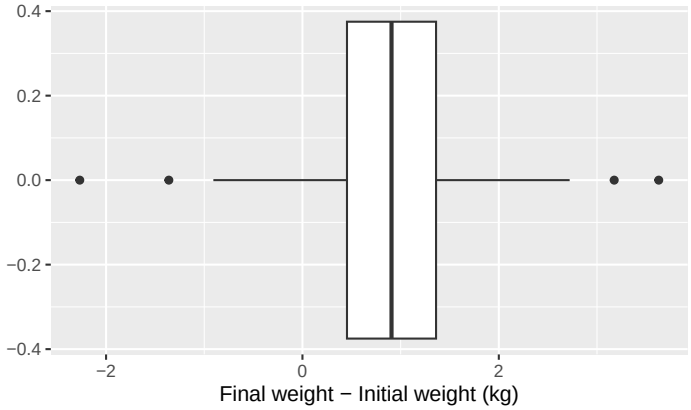


This plot shows the same information as in the summary table of counts, but it is easier to process the information visually.

1.5.3 Distribution of a single continuous variable

There are several different types of plots which may be used to summarise the shape of the distribution of a single continuous variables. For instance, suppose we are interested in the gain in weight in Example 1.3.

We can use a box-and-whiskers plot to summarise the distribution of the weight difference:

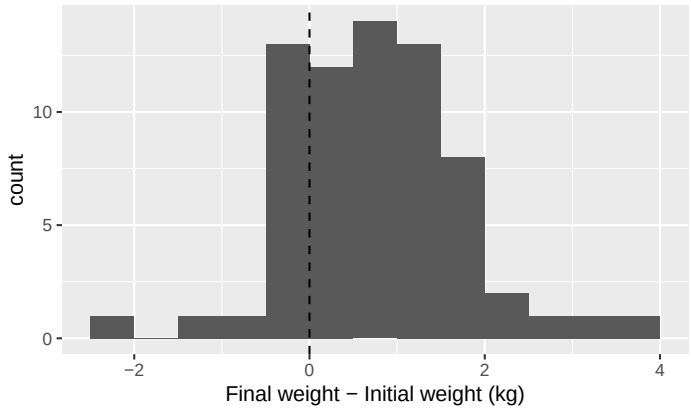


The line in the middle of the box is the median service weight difference, and the box runs from the first quartile to the third quartile. The lines (or whiskers) extend from the smallest value no further further than $1.5 \times$ IQR from the first quartile to the largest value no further than $1.5 \times$ IQR from the third quartile.

Any outliers, which are outside the range of the whiskers, are shown as separate points.

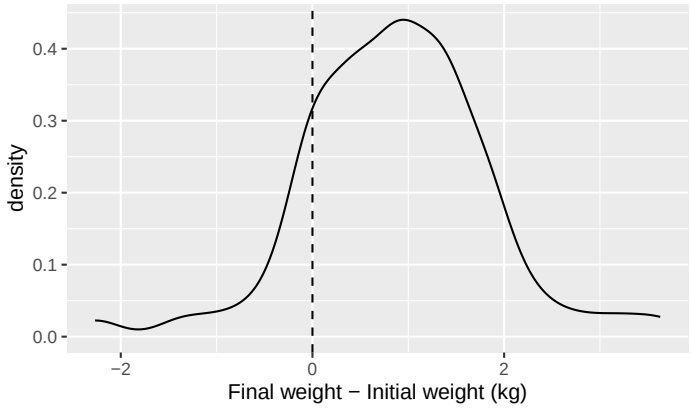
We can see the the majority of students in the sample do gain weight over the course of the study.

To get more information about the shape of the distribution, we could plot a histogram of the weight difference:



The histogram splits the x-axis into small intervals or "bins" (here each of width 0.5), and counts how many times the variable (here difference in weight) is in each bin. Histograms often look very different for different bin widths.

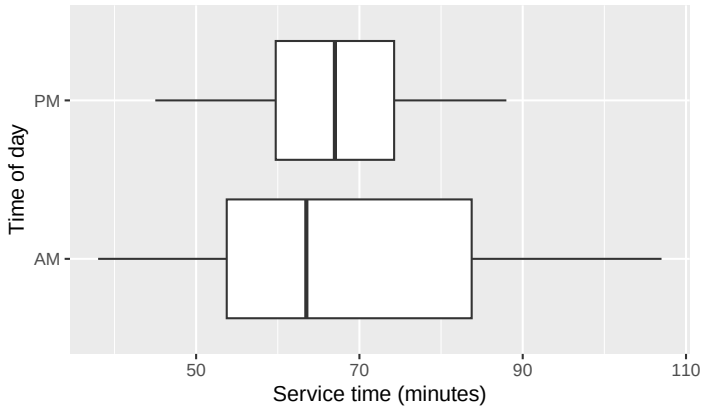
Alternatively, we could draw a density plot:



This plots an estimate of the probability density function, which we will learn about in Chapter 3.

1.5.4 Relationship between continuous and discrete variables

We can also use a box-and-whiskers plot to show how a continuous variable changes with a discrete one. For instance, for the service time data from Example 1.1, we are interested in how AM and PM service times compare:

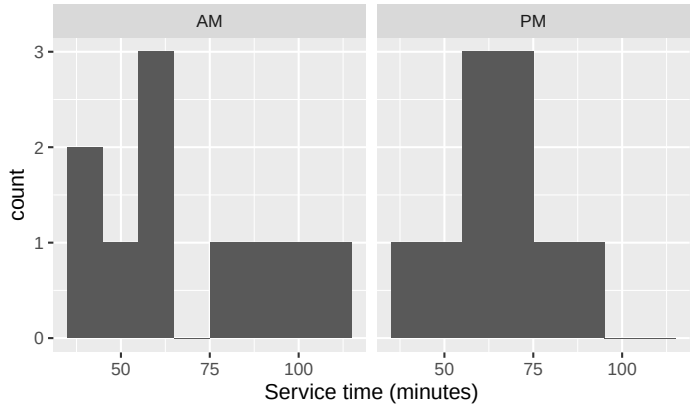


Since no outlying points are shown in this case, the whiskers show the range of the data.

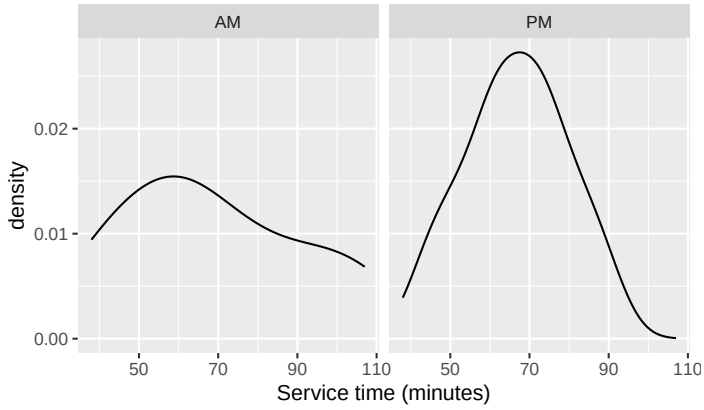
We see that the AM service times are on average shorter, but also more spread out than the PM service times.

We can also use any of the other methods for plotting a single continuous variable, and draw separate plots for each group in the discrete variable.

We can use histograms:

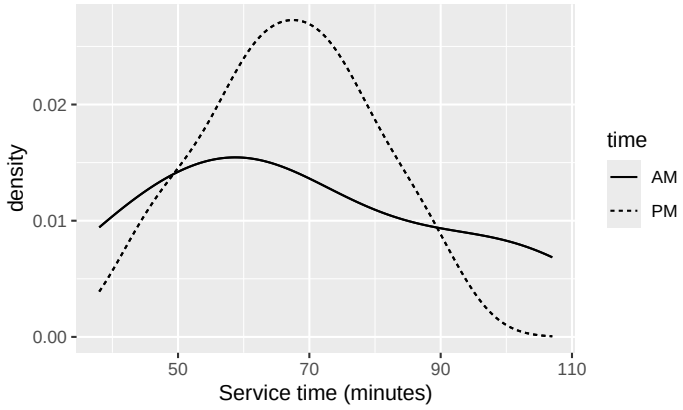


Or density plots:



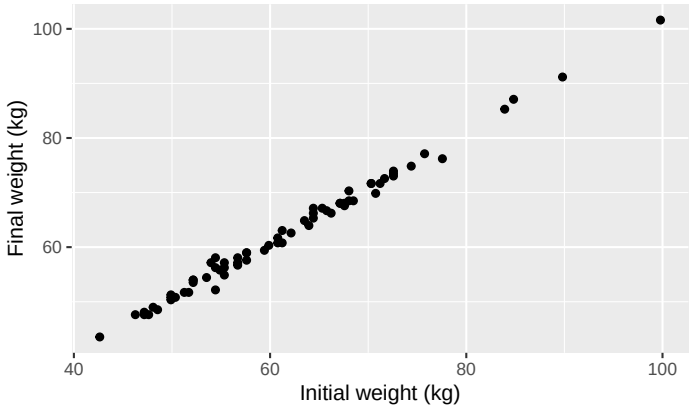
We could also plot the densities for the different groups on a single plot, using line type to distinguish

between the groups:

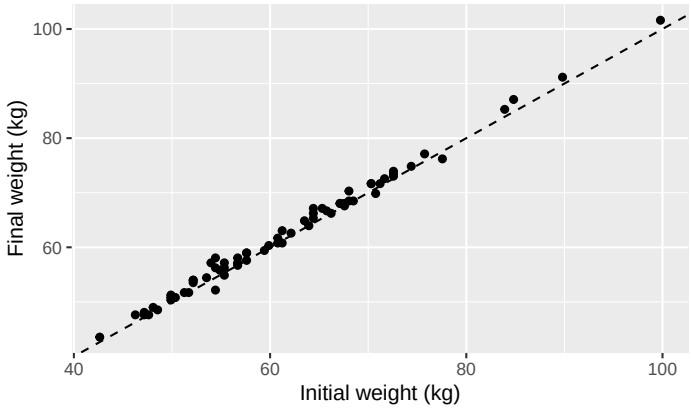


1.5.5 Relationship between two continuous variables

A scatter-plot is a good way to look at the relationship between two continuous variables. For instance, for the weight data from Example 1.3, we can plot the final weights against the initial weights:



We can add the straight line $y = x$ to the plot:

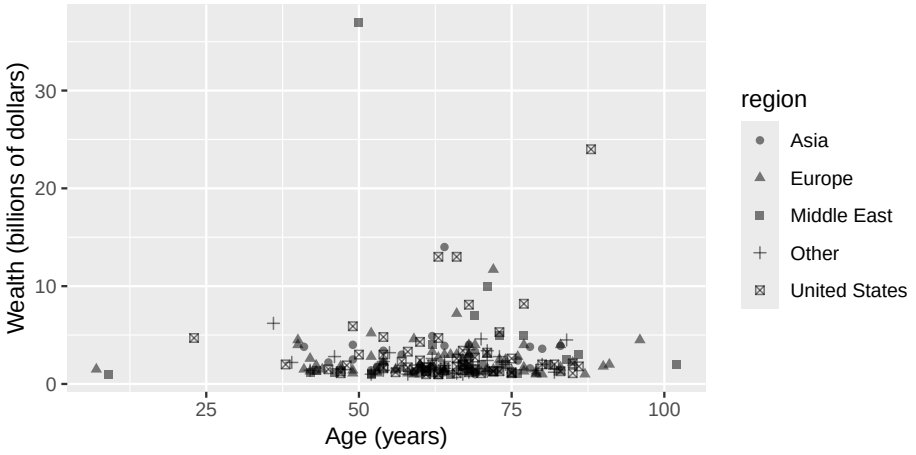


We can see that the majority of the points lie above this $y = x$ line, so the majority of students in the sample gain weight over the course of the study, as we have seen before. This appears to be the case irrespective of the initial weight of the student.

1.5.6 Relationships between more than two variables

We can combine the methods we have already seen to plot relationships between more than two variables.

For instance, in the billionaires data in Example 1.4, we can plot wealth against age as a scatter-plot, using the shape of the points to distinguish between different regions:



The plot lets us immediately read off some interesting aspects of the data. For instance, there were a few very young billionaires, the richest person in the world at this time was age 50 and from the Middle East, and the richest few people in the world had wealth many times higher than most people on the list.

Chapter 2

Introduction to Probability

2.1 Definitions of probability

Probabilities are often used to express the uncertainty of events of interest happening. For example, we may say that: (i) I think it is highly likely that Manchester City will retain the premiership title this season or to be more specific, I think there is more than an 80% chance that Manchester City will keep the title; (ii) the probability of a tossed fair coin landing heads is 0.5. As we have seen in Chapter 1, there is uncertainty everywhere, and probability is the language we use to quantify this uncertainty.

The two examples above, football and tossing a coin, convey two different interpretations of probability. The football probability is the commentator's own subjective belief. The commentator certainly has not performed a large experiment involving all the 20 teams over the whole (future) season under all playing conditions, players, managers and transfers. This notion is known as subjective probability. Subjective probability gives a measure of the plausibility of the proposition, to the person making it, in the light of past experience (e.g. Manchester City are the current champions) and other evidence (e.g. they spent the maximum amount of money buying players). There are plenty of other examples, e.g. I think there is a 70% chance that the FTSE 100 will rise tomorrow, or according to the Met Office there is a 40% chance that we will have a white Christmas this year in Southampton.

The second definition of probability comes from the long-term relative frequency of a result of a random experiment (e.g. coin tossing) which can be repeated an infinite number of times under essentially similar conditions.

Imagine we are able to repeat a random experiment under identical conditions and count how many of those repetitions result in the event A. The relative frequency of A, i.e. the ratio

$$\frac{\text{the number of repetitions resulting in } A}{\text{total number of repetitions}},$$

approaches a fixed limit value as the number of repetitions increases. This limit value is defined as $P\{A\}$.

As a simple example, in the experiment of tossing a particular coin, suppose we are interested in the event A of getting a 'head'. We can toss the coin 1000 times (i.e. do 1000 replications of the experiment) and record the number of heads out of the 1000 replications. Then the relative frequency of A out of the 1000 replications is the proportion of heads observed.

Sometimes, however, it is much easier to find $P\{A\}$ by using some 'common knowledge' about probability. For example, if the coin in the example above is fair (i.e. $P\{\text{head}\} = P\{\text{tail}\}$), then this information and the common knowledge that $P\{\text{head}\} + P\{\text{tail}\} = 1$ immediately imply that

$P\{\text{head}\} = 0.5$ and $P\{\text{tail}\} = 0.5$. The 'common knowledge' about probability will be formalised as the axioms of probability, which form the foundation of probability theory.

2.2 Some definitions

Before we can state and use the axioms of probability, we need introduce some terminology.

A *random experiment* is one in which we do not know exactly what outcome the experiment will give, even though we can write down all the possible outcomes. The set of all possible outcomes is called the *sample space* (S). For example, in a coin tossing experiment, $S = \{\text{head}, \text{tail}\}$. If we toss two coins together, $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ where H and T denote respectively the outcome head and tail from the toss of a single coin.

An *event* is a particular result of the random experiment. For example, HH (two heads) is an event when we toss two coins together. Similarly, at least one head e.g. $\{\text{HH}, \text{HT}, \text{TH}\}$ is an event as well. Events are denoted by capital letters $A, B, C, \ldots$ or A_1, B_1, A_2 etc., and a single outcome is called an *elementary event*, e.g. HH. An event which is a group of elementary events is called a *composite event*, e.g. at least one head. How to determine the probability of a given event A, $P\{A\}$, is the focus of probability theory.

Example 2.1 (Die throw)**.** Roll a six-faced die and observe the score on the uppermost face. Here $S = \{1, 2, 3, 4, 5, 6\}$, which is composed of six elementary events.

The *union* of two given events A and B, denoted as (A or B) or $A \cup B$, consists of the outcomes that are either in A or B or both. 'Event $A \cup B$ occurs' means 'either A or B occurs or both occur'.

For example, in Example 2.1, suppose A is the event that *an even number is observed*. This event consists of the set of outcomes 2, 4 and 6, i.e. $A = \{\text{an even number}\} = \{2, 4, 6\}$. Suppose B is the event that *a number larger than 3 is observed*. This event consists of the outcomes 4, 5 and 6, i.e. $B = \{\text{a number larger than 3}\} = \{4, 5, 6\}$. Hence the event $A \cup B = \{\text{an even number or a number larger than 3}\} = \{2, 4, 5, 6\}$. Clearly, when a 6 is observed, both A and B have occurred.

The *intersection* of two given events A and B, denoted as (A and B) or $A \cap B$, consists of the outcomes that are common to both A and B. 'Event $A \cap B$ occurs' means 'both A and B occur'. For example, in Example 2.1, $A \cap B = \{4, 6\}$. Additionally, if $C = \{\text{a number less than 6}\} = \{1, 2, 3, 4, 5\}$, the intersection of events A and C is the event $A \cap C = \{\text{an even number less than 6}\} = \{2, 4\}$.

The union and intersection of two events can be generalized in an obvious way to the union and intersection of more than two events.

Two events A and D are said to be *mutually exclusive* if $A \cap D = \emptyset$, where $\emptyset$ denotes the empty set, i.e. A and D have no outcomes in common. Intuitively, 'A and D are mutually exclusive' means 'A and D cannot occur simultaneously in the experiment'.

In Example 2.1, if $D = \{\text{an odd number}\} = \{1, 3, 5\}$, then $A \cap D = \emptyset$ and so A and D are mutually exclusive. As expected, A and D cannot occur simultaneously in the experiment.

For a given event A, the *complement* of A is the event that consists of all the outcomes not in A and is denoted by A' . Note that $A \cup A' = S$ and $A \cap A' = \emptyset$.

2.3 Axioms of probability

Here are the three axioms of probability:

A1 $P\{S\} = 1$.
A2 $0 \leq P\{A\} \leq 1$ for any event A.

A3 $P\{A \cup B\} = P\{A\} + P\{B\}$ provided that A and B are mutually exclusive events.

Here are some of the consequences of the axioms of probability:

(1) For any event $A, P\{A\} = 1 - P\{A'\}$.
(2) From (1) and Axiom A1, $P\{\emptyset\} = 1 - P\{S\} = 0$. Hence if A and B are mutually exclusive events, then $P\{A \cap B\} = 0$.
(3) If D is a subset of $E, D \subset E$, then $P\{E \cap D'\} = P\{E\} - P\{D\}$ which implies for arbitrary events A and $B, P\{A \cap B'\} = P\{A\} - P\{A \cap B\}$.
(4) It can be shown by mathematical induction that Axiom A3 holds for more than two mutually exclusive events:

$$P\{A_1 \cup A_2 \cup \cdots \cup A_k\} = P\{A_1\} + P\{A_2\} + \ldots + P\{A_k\}$$

provided that $A_1, \ldots, A_k$ are mutually exclusive events. Hence, the probability of an event A is the sum of the probabilities of the individual outcomes that make up the event.
(5) For the union of two arbitrary events, we have the General addition rule: For any two events A and B

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}.$$

Proof. We can write $A \cup B = (A \cap B') \cup (A \cap B) \cup (A' \cap B)$. All three of these are mutually exclusive events. Hence,

$$\begin{aligned} P\{A \cup B\} &= P\{A \cap B'\} + P\{A \cap B\} + P\{A' \cap B\} \\ &= P\{A\} - P\{A \cap B\} + P\{A \cap B\} + P\{B\} - P\{A \cap B\} \\ &= P\{A\} + P\{B\} - P\{A \cap B\}. \end{aligned}$$

□

(6) The sum of the probabilities of all the outcomes in the sample space S is 1 .

2.4 Using combinatorics to find probabilities

2.4.1 Experiments with equally likely outcomes

For an experiment with N equally likely possible outcomes, the axioms (and the consequences above) can be used to find $P\{A\}$ of any event A in the following way.

From consequence (4), we assign probability $1/N$ to each outcome.

For any event A, we find $P\{A\}$ by adding up $1/N$ for each of the outcomes in event A:

$$P\{A\} = \frac{\text{number of outcomes in } A}{\text{total number of possible outcomes of the experiment}}.$$

For experiments with equally likely outcomes, the task of calculating the probability of an event therefore reduces to counting the number of outcomes in the event and the total number of possible outcomes. In the following sections, will use ideas from combinatorics (the mathematics of counting) to help us complete these tasks.

Return to Example 2.1 where a six-faced die is rolled. Suppose that one wins a bet if a 6 is rolled. Then the probability of winning the bet is 1/6 as there are six possible outcomes in the sample space and exactly one of those, 6, wins the bet. Suppose A denotes the event that an even-numbered face is rolled. Then $P\{A\} = 3/6 = 1/2$ as we can expect.

Example 2.2 (Dice throw)**.** Roll 2 distinguishable dice and observe the scores. Here

$$S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),...,(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$$

which consists of 36 possible outcomes or elementary events, $A_1, \ldots, A_{36}$. What is the probability of the outcome 6 in both the dice? The required probability is 1/36. What is the probability that the sum of the two dice is greater than 6? How about the probability that the sum is less than any number, e.g. 8?

Hint: Write down the sum for each of the 36 outcomes and then find the probabilities asked just by inspection. Remember, each of the 36 outcomes has equal probability 1/36.

2.4.2 Multiplication rule of counting

To complete a specific task, suppose one has to complete $k(\geq 1)$ sub-tasks sequentially. If there are n_i different ways to complete the ith sub-task $(i = 1, \ldots, k)$ then there are $n_1 \times n_2 \times \cdots \times n_k$ different ways to complete the task.

Example 2.3 (Bus routes)**.** Suppose there are 7 routes to London from Southampton and then there are 5 routes to Cambridge out of London. How many ways can I travel to Cambridge from Southampton via London? The answer is obviously 35.

2.4.3 The number of permutations of k from n: $P(n,k)$

Suppose we are asked to select $k(\geq 1)$ from the $n(n \geq k)$ available people and sit the k selected people in k (different) chairs. How many different ways are there to complete the task?

By considering the ith sub-task as selecting a person to sit in the ith chair $(i = 1, \ldots, k)$, it follows directly from the multiplication rule in Section 2.4.2 that there are $n(n-1)\ldots(n-[k-1])$ ways to complete the task.

The number $n(n-1)\ldots(n-[k-1])$ is called the number of permutations of k from n and denoted by

$$P(n,k) = n(n-1)\ldots(n-[k-1]).$$

In particular, when $k = n$ we have $P(n,n) = n(n-1)\ldots 1$, which is called 'n factorial' and denoted as $n!$. Note that 0! is defined to be 1. We have

$$P(n,k) = n(n-1)\ldots(n-[k-1]) = \frac{n(n-1)\ldots(n-[k-1])\times(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}.$$

Example 2.4 (Football)**.** How many possible rankings are there for the 20 football teams in the premier league at the end of a season? This number is given by $P(20,20) = 20!$, which is a huge number! How many possible permutations are there for the top 4 positions who will qualify to play in Europe in the next season? This number is given by $P(20,4) = 20 \times 19 \times 18 \times 17$.

2.4.3.1 The number of combinations of k from n: $\binom{n}{k}$

Suppose we are asked to select $k(\geq 1)$ from the n $(n \geq k)$ available people. Note that this task does NOT involve sitting the k selected people in k (different) chairs. We want to find the number of possible ways to complete this task, which is denoted as $\binom{n}{k}$.

For this, let us reconsider the task of "selecting k from the n available people and sitting the k selected people in k (different) chairs", which we already know from the discussion above has $P(n,k)$ ways to complete. Alternatively, to complete this task, one has to complete two sub-tasks sequentially. The first sub-task is to select k from the n available people, which has $\binom{n}{k}$ ways. The second sub-task is

to sit the k selected people in k (different) chairs, which has $k!$ ways. It follows directly from the multiplication rule that there are $\binom{n}{k} \times k!$ ways to complete the task. Hence we have

$$P(n,k) = \binom{n}{k} \times k!$$

so

$$\binom{n}{k} = \frac{P(n,k)}{k!} = \frac{n!}{(n-k)!k!}.$$

Example 2.5 (Football)**.** How many possible ways are there to choose 3 teams for the bottom positions of the premier league table at the end of a season? This number is given by $\binom{20}{3} = 20 \times 19 \times 18/3!$, which does not take into consideration the rankings of the three bottom teams.

Example 2.6 (Microchips)**.** A box contains 12 microchips of which 4 are faulty. A sample of size 3 is drawn from the box without replacement.

- How many selections of 3 can be made? $\binom{12}{3}$
- How many samples have all 3 chips faulty? $\binom{4}{3}$.
- How many selections have exactly 2 faulty chips? $\binom{8}{1}\binom{4}{2}$
- How many samples of 3 have 2 or more faulty chips? $\binom{8}{1}\binom{4}{2} + \binom{4}{3}$

2.4.4 Calculation of probabilities of events under sampling 'at random'

For the experiment of 'selecting a sample of size n from a box of N items without replacement', a sample is said to be selected at random if all the possible samples of size n are equally likely to be selected. All the possible samples are then equally likely outcomes of the experiment and so assigned equal probabilities.

Example 2.7 (Microchips continued)**.** In Example 2.6 assume that 3 microchips are selected at random without replacement. Then

- each outcome (sample of size 3) has probability $1/\binom{12}{3}$.
- $P\{\text{all 3 selected microchips are faulty}\} = \binom{4}{3}/\binom{12}{3}$.
- $P\{\text{2 chips are faulty}\} = \binom{8}{1}\binom{4}{2}/\binom{12}{3}$.
- $P\{\text{2 or more chips are faulty}\} = \left(\binom{8}{1}\binom{4}{2} + \binom{4}{3}\right)/\binom{12}{3}$.

2.4.5 A general 'urn problem'

Example 2.6 is one particular case of the following general urn problem which can be solved by the same technique. A sample of size n is drawn at random without replacement from a box of N items containing a proportion p of defective items.

- How many defective items are in the box? Np. How many good items are there? $N(1-p)$. Assume these to be integers.
- The probability of exactly x defective items in the sample of n items is

$$\frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}.$$

- Which values of x (in terms of N, n and p) make this expression well defined? We'll see later that these values of x and the corresponding probabilities make up what is called the *hyper-geometric* distribution.

Example 2.8 (Selecting a committee)**.** There are 10 students available for a committee of which 4 are men and 6 are women. A random sample of 3 students are chosen to form the committee — what

is the probability that exactly one is a man? The total number of possible outcomes of the experiment is equal to the number of ways of selecting 3 students from 10 and is given by $\binom{10}{3}$. The number of outcomes in the event 'exactly one is a man' is equal to the number of ways of selecting 3 students from 10 with exactly one man, and given by $\binom{4}{1}\binom{6}{2}$ Hence

$$\begin{aligned} P\{\text{exactly one man}\} &= \frac{\text{number of ways of selecting one man and two women}}{\text{number of ways of selecting 3 students}} \\ &= \frac{\binom{4}{1}\binom{6}{2}}{\binom{10}{3}} = \frac{4 \times 15}{120} = \frac{1}{2} \end{aligned}$$

Example 2.9 (The National Lottery)**.** In Lotto, a winning ticket has six numbers from 1 to 59 matching those on the balls drawn on a Wednesday or Saturday evening. The 'experiment' consists of drawing the balls from a box containing 59 balls. The 'randomness', the equal chance of any set of six numbers being drawn, is ensured by the spinning machine, which rotates the balls during the selection process. What is the probability of winning the jackpot? The total number of possible selections of six balls/numbers is $\binom{59}{6}$ There is only 1 selection for winning the jackpot. Hence

$$P\{\text{jackpot}\} = \frac{1}{\binom{59}{6}} = 2.22 \times 10^{-8},$$

which is roughly 1 in 45 million.

There is one other way of win a very large prize, of £1 million, by using the bonus ball — matching 5 of the selected 6 balls plus matching the bonus ball. The probability of this is given by

$$P\{5 \text{ matches + bonus}\} = \frac{6}{\binom{59}{6}} = 1.33 \times 10^{-7}.$$

Other smaller prizes are given for fewer matches. The corresponding probabilities are:

$$P\{5 \text{ matches}\} = \frac{\binom{6}{5}\binom{53}{1}}{\binom{59}{6}} = 7.06 \times 10^{-6}.$$

$$P\{4 \text{ matches}\} = \frac{\binom{6}{4}\binom{53}{2}}{\binom{59}{6}} = 0.000459.$$

$$P\{3 \text{ matches}\} = \frac{\binom{6}{3}\binom{53}{3}}{\binom{59}{6}} = 0.0104.$$

2.5 Conditional probability and Bayes' Theorem

How can we use additional information, i.e. things that have already happened, in the calculation of probabilities? For example, a person may have a certain disease, e.g. diabetes or HIV/AIDS, whether or not they show any symptoms of it. Suppose a randomly selected person is found to have the symptom. Given this additional information, what is the probability that they have the disease? Note that having the symptom does not fully guarantee that the person has the disease.

Applications of conditional probability occur naturally in actuarial science and medical studies, where conditional probabilities such as "what is the probability that a person will survive for another 20 years given that they are still alive at the age of 40?" are calculated. In many real problems, one has to determine the probability of an event A when one already has some partial knowledge of the outcome of an experiment, i.e. another event B has already occurred. For this, one needs to find the conditional probability.

Example 2.10 (Die throw continued)**.** Return to the rolling of a fair die (Example 2.1). Let

$$A = \{\text{a number greater than } 3\} = \{4, 5, 6\}, B = \{\text{an even number}\} = \{2, 4, 6\}.$$

It is clear that $P\{B\} = 3/6 = 1/2$. This is the unconditional probability of the event B. It is sometimes called the *prior* probability of B.

However, suppose that we are told that the event A has already occurred. What is the probability of B now given that A has already happened?

The sample space of the experiment is $S = \{1, 2, 3, 4, 5, 6\}$, which contains $n = 6$ equally likely outcomes.

Given the partial knowledge that event A has occurred, only the $n_A = 3$ outcomes in $A = \{4, 5, 6\}$ could have occurred. However, only some of the outcomes in B among these n_A outcomes in A will make event B occur; the number of such outcomes is given by the number of outcomes $n_{A\cap B}$ in both A and B, i.e., $A \cap B$, and equal to 2. Hence the probability of B, given the partial knowledge that event A has occurred, is equal to

$$\frac{2}{3} = \frac{n_{A\cap B}}{n_A} = \frac{n_{A\cap B}/n}{n_A/n} = \frac{P\{A \cap B\}}{P\{A\}}$$

Hence we say that $P\{B|A\} = 2/3$, which is often interpreted as the *posterior* probability of B given A. The additional knowledge that A has already occurred has helped us to revise the prior probability of $1/2$ to $2/3$.

This simple example leads to the following general definition of conditional probability.

2.5.1 Definition of conditional probability

For events A and B with $P\{A\} > 0$, the conditional probability of event B, given that event A has occurred, is

$$P\{B|A\} = \frac{P\{A \cap B\}}{P\{A\}}.$$

Example 2.11. Of all individuals buying a mobile phone, 60% include a 64 GB hard disk in their purchase, 40% include a 16 MP camera and 30% include both. If a randomly selected purchase includes a 16 MP camera, what is the probability that a 64GB hard disk is also included? The conditional probability is given by

$$P\{\text{64 GB}|\text{16 MP}\} = \frac{P\{\text{64 GB} \cap \text{16 MP}\}}{P\{\text{16 MP}\}} = \frac{0.3}{0.4} = 0.75.$$

2.5.2 Multiplication rule of conditional probability

By rearranging the conditional probability definition, we obtain the multiplication rule of conditional probability:

$$P\{A \cap B\} = P\{A\}P\{B|A\}.$$

Clearly the roles of A and B could be interchanged:

$$P\{A \cap B\} = P\{B\}P\{A|B\}.$$

Hence the multiplication rule of conditional probability for two events is

$$P\{A \cap B\} = P\{B\}P\{A|B\} = P\{A\}P\{B|A\}.$$

It is straightforward to show by mathematical induction the following multiplication rule of conditional probability for $k(\geq 2)$ events $A_1, A_2, \ldots, A_k$:

$$P\{A_1 \cap A_2 \cap \ldots \cap A_k\} = P\{A_1\}P\{A_2|A_1\}P\{A_3|A_1 \cap A_2\} \ldots P\{A_k|A_1 \cap A_2 \cap \ldots \cap A_{k-1}\}.$$

Example 2.12 (Selecting a committee continued)**.** Return to the committee selection example (Example 2.8), where there are 4 men (M) and 6 women (W). We want to select a 2-person committee. Find:

(i) the probability that both are men,
(ii) the probability that one is a man and the other is a woman.

We have already dealt with this type of urn problem by using the combinatorial method. Here, the multiplication rule is used instead. Let M_i be the event that the ith person is a man, and W_i be the event that the ith person is a woman, $i = 1, 2$. Then

$$\text{Prob in (i)} = P\{M_1 \cap M_2\} = P\{M_1\}P\{M_2|M_1\} = \frac{4}{10} \times \frac{3}{9},$$

$$\begin{aligned}\text{Prob in (ii)} &= P\{M_1 \cap W_2\} + P\{W_1 \cap M_2\} \\ &= P\{M_1\}P\{W_2|M_1\} + P\{W_1\}P\{M_2|W_1\} \\ &= \frac{4}{10} \times \frac{6}{9} + \frac{6}{10} \times \frac{4}{9}\end{aligned}$$

You can find the probability that 'both are women' in a similar way.

2.5.3 Total probability formula

Example 2.13 (Phones)**.** Suppose that in our world there are only three phone manufacturing companies: A Pale, B Sung and C Windows, and their market shares are respectively 30, 40 and 30 percent. Suppose also that respectively 5, 8, and 10 percent of their phones become faulty within one year. If I buy a phone randomly (ignoring the manufacturer), what is the probability that my phone will develop a fault within one year? After finding the probability, suppose that my phone developed a fault in the first year — what is the probability that it was made by A Pale?

Company	Market share	Percent defective
A Pale	30%	5%
B Sung	40%	8%
C Windows	30%	10%

To answer this type of question, we derive two of the most useful results in probability theory: the total probability formula and Bayes' theorem. First, let us derive the total probability formula.

Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive, i.e.

$$B_i \cap B_j = \emptyset, \text{for all } 1 \leq i \neq j \leq k.$$

and exhaustive events, i.e.:

$$B_1 \cup B_2 \cup \ldots \cup B_k = S.$$

Now any event A can be represented by

$$A = A \cap S = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_k)$$

where $(A \cap B_1), (A \cap B_2), \ldots, (A \cap B_k)$ are mutually exclusive events. Hence the Axiom A3 of probability gives

$$\begin{aligned}P\{A\} &= P\{A \cap B_1\} + P\{A \cap B_2\} + ... + P\{A \cap B_k\} \\ &= P\{B_1\}P\{A|B_1\} + P\{B_2\}P\{A|B_2\} + ... + P\{B_k\}P\{A|B_k\}.\end{aligned}$$

This last expression is called the *total probability formula* for $P\{A\}$.

Example 2.14 (Phones continued)**.** We can now find the probability of the event, say A, that a randomly selected phone develops a fault within one year. Let B_1, B_2, B_3 be the events that the phone is manufactured respectively by companies A Pale, B Sung and C Windows. Then we have:

$$\begin{aligned} P\{A\} &= P\{B_1\}P\{A|B_1\} + P\{B_2\}P\{A|B_2\} + P\{B_3\}P\{A|B_3\} \\ &= 0.30 \times 0.05 + 0.40 \times 0.08 + 0.30 \times 0.10 \\ &= 0.077. \end{aligned}$$

Now suppose that my phone has developed a fault within one year. What is the probability that it was manufactured by A Pale? To answer this we need to introduce Bayes' Theorem.

2.5.4 Bayes' theorem

Theorem 2.1 (Bayes' Theorem)**.** *Let A and B be events. Then*

$$P\{B|A\} = \frac{P\{B\}P\{A|B\}}{P\{A\}}.$$

Proof. From the definition of conditional probability, we have

$$P\{B|A\} = \frac{P\{B \cap A\}}{P\{A\}} = \frac{P\{B\}P\{A|B\}}{P\{A\}}.$$

□

The probability $P\{B|A\}$ is called the posterior probability of B given A and $P\{B\}$ is called the prior probability. Bayes' theorem is the rule that converts the prior probability into the posterior probability by using the additional information that some other event, A above, has already occurred.

Example 2.15 (Phones continued)**.** The probability that my faulty phone was manufactured by A Pale is

$$P\{B_1|A\} = \frac{P\{B_1\}P\{A|B_1\}}{P\{A\}} = \frac{0.30 \times 0.05}{0.077} = 0.1948.$$

Similarly, the probability that the faulty phone was manufactured by B Sung is 0.4156, and the probability that it was manufactured by C Windows is $1 - 0.1948 - 0.4156 = 0.3896$.

As in this example, we usually need to use the total probability formula to calculate the denominator $P(A)$ in Bayes' theorem.

2.6 Independent events

2.6.1 Introduction and definition of independence

We have just seen that the probability of an event may change if we have additional information. However, in many situations the probabilities may not change. For example, the results of two coin tosses should not depend on each other. In this section we will learn about the probabilities of independent events. Much of statistical theory relies on the concept of independence.

We have seen examples where prior knowledge that an event A has occurred has changed the probability that event B occurs. There are many situations where this does not happen. The events are then said to be independent. Intuitively, events A and B are independent if the occurrence of one event does not affect the probability that the other event occurs. This is equivalent to saying that $P\{B|A\} = P\{B\}$, where $P\{A\} > 0$, and $P\{A|B\} = P\{A\}$, where $P\{B\} > 0$.

These give the following formal definition: A and B are *independent* events if $P\{A \cap B\} = P\{A\}P\{B\}$.

Example 2.16 (Die throw)**.** Throw a fair die. Let A be the event that "the result is even" and B be the event that "the result is greater than 3". We want to show that A and B are not independent.

For this, we have $P\{A \cap B\} = P\{\text{either a 4 or 6 thrown}\} = 1/3$, but $P\{A\} = 1/2$ and $P\{B\} = 1/2$, so that $P\{A\}P\{B\} = 1/4 \neq 1/3 = P\{A \cap B\}$. Therefore A and B are not independent events.

Independence is often assumed on physical grounds, although sometimes incorrectly. There are serious consequences for wrongly assuming independence, e.g. the financial crisis in 2008. However, when the events are independent then the simpler product formula for joint probability is then used.

Example 2.17 (Dice throw)**.** Two fair dice when shaken together are assumed to behave independently. Hence the probability of two sixes is $1/6 \times 1/6 = 1/36$.

Example 2.18 (Assessing risk in legal cases)**.** There have been some disastrous miscarriages of justice as a result of incorrect assumption of independence. Please read "Incorrect use of independence — Sally Clark Case" on Blackboard.

Theorem 2.2 (Independence of complementary events)**.** *If A and B are independent, so are A' and B'.*

Proof. Given that $P\{A \cap B\} = P\{A\}P\{B\}$, we need to show that $P\{A' \cap B'\} = P\{A'\}P\{B'\}$. We have

$$\begin{aligned} P\{A' \cap B'\} &= 1 - P\{A \cup B\} \\ &= 1 - [P\{A\} + P\{B\} - P\{A \cap B\}] \\ &= 1 - [P\{A\} + P\{B\} - P\{A\}P\{B\}] \\ &= [1 - P\{A\}] - P\{B\}[1 - P\{A\}] \\ &= [1 - P\{A\}][1 - P\{B\}] \\ &= P\{A'\}P\{B'\} \end{aligned}$$

□

2.6.2 Independence with three events

The ideas of conditional probability and independence can be extended to more than two events.

Three events A, B and C are defined to be independent if

$$P\{A \cap B\} = P\{A\}P\{B\},\ P\{A \cap C\} = P\{A\}P\{C\},\ P\{B \cap C\} = P\{B\}P\{C\}, \tag{2.1}$$

$$P\{A \cap B \cap C\} = P\{A\}P\{B\}P\{C\}. \tag{2.2}$$

Note that (2.1) does NOT imply (2.2) as shown by the next example. Hence, to show the independence of A, B and C, it is necessary to show that both (2.1) and (2.2) hold.

Example 2.19. A box contains eight tickets, each labelled with a binary number. Two are labelled with the binary number 111, two are labelled with 100, two with 010 and two with 001. An experiment consists of drawing one ticket at random from the box. Let A be the event "the first digit is 1", B the event "the second digit is 1" and C be the event "the third digit is 1". It is clear that $P\{A\} = P\{B\} = P\{C\} = 4/8 = 1/2$ and $P\{A \cap B\} = P\{A \cap C\} = P\{B \cap C\} = 1/4$, so the events are pairwise independent, i.e. (2.1) holds. However $P\{A \cap B \cap C\} = 2/8 \neq P\{A\}P\{B\}P\{C\} = 1/8$. So (2.2) does not hold and A, B and C are not independent.

Chapter 3

Probability Distributions

3.1 Introduction

Last chapter's combinatorial probabilities are difficult to find and very problem-specific. Instead, in this chapter we shall find easier ways to calculate probability in structured cases. The outcomes of random experiments will be represented as values of a variable which will be random since the outcomes are random. In so doing, we will make our life a lot easier in calculating probabilities in many stylised situations which represent reality.

3.2 Random variables

3.2.1 Introduction

In this section we will learn about the probability distribution of a random variable defined by its probability function. The probability function will be called the probability mass function for discrete random variables and the probability density function for continuous random variables.

A random variable defines a one-to-one mapping of the sample space consisting of all possible outcomes of a random experiment to the set of real numbers. For example, I toss a coin. Assuming the coin is fair, there are two possible equally likely outcomes: head or tail. These two outcomes must be mapped to real numbers. For convenience, I may define the mapping which assigns the value 1 if head turns up and 0 otherwise. Hence, we have the mapping

$$\text{Head} \to 1, \text{Tail} \to 0.$$

We can conveniently denote the random variable by X which is the number of heads obtained by tossing a single coin. The possible values of X are 0 and 1.

You will say that this is a trivial example. Indeed it is. But it is very easy to generalise the concept of random variables. Simply define a mapping of the outcomes of a random experiment to the real number space. For example, I toss the coin n times and count the number of heads and denote that to be X. X can take any real positive integer value between 0 and n. Among other examples, suppose I select a University of Southampton student at random and measure their height. The outcome in metres will probably be a number between one metre and two metres. But I can't exactly tell which value it will be since I do not know which student will be selected in the first place. However, when a student has been selected I can measure their height and get a value such as 1.432 metres.

We now introduce two notations: X (or in general the capital letters Y, Z etc.) to denote the random variable, e.g. height of a randomly selected student, and the corresponding lower case letter x (or y,

z) to denote a particular value, e.g. 1.432 metres. We will follow this convention throughout. For a random variable, say X, we will also adopt the notation $P(X \in A)$, read probability that X belongs to A, instead of the previous $P\{A\}$ for any event A.

3.2.2 Discrete and continuous random variables

If a random variable has a finite or countably infinite set of values it is called discrete. For example, the number of Apple computer users among 20 randomly selected students, or the number of credit cards a randomly selected person has in their wallet. When the random variable can take any value on the real line it is called a continuous random variable. For example, the height of a randomly selected student. A random variable can also take a mixture of discrete and continuous values, e.g. volume of precipitation collected in a day; some days it could be zero, on other days it could be a continuous measurement, e.g. 1.234 mm.

3.2.3 Probability distribution of a random variable

Recall the first axiom of probability ($P\{S\} = 1$), which means total probability equals 1. Since a random variable is merely a mapping from the outcome space to the real line, the combined probability of all possible values of the random variable must be equal to 1. A probability distribution distributes the total probability 1 among the possible values of the random variable.

Example 3.1. Returning to the coin-tossing experiment, if the probability of getting a head with a coin is p (and therefore the probability of getting a tail is $1 - p$), then the probability that $Y = 0$ is $1 - p$ and the probability that $Y = 1$ is p. This gives us the probability distribution of Y, and we say that Y has the probability function

$$P(Y = y) = \begin{cases} 1 - p & \text{for } y = 0 \\ p & \text{for } y = 1. \end{cases}$$

This is an example of the **Bernoulli distribution** with parameter p, the simplest discrete distribution.

Example 3.2. Suppose we consider tossing the coin twice and again defining the random variable X to be the number of heads obtained. The values that X can take are 0, 1 and 2 with probabilities $(1 - p)^2$, $2p(1 - p)$ and p^2, respectively. Here the probability function is

$$P(X = x) = \begin{cases} (1 - p)^2 & \text{for } x = 0 \\ 2p(1 - p) & \text{for } x = 1 \\ p^2 & \text{for } x = 2. \end{cases}$$

This is a particular case of the Binomial distribution. We will learn about it soon.

In general, for a discrete random variable we define a function $f(x)$ to denote $P(X = x)$ (or $f(y)$ to denote $P(Y = y)$) and call the function $f(x)$ the probability function (pf) or probability mass function (pmf) of the random variable X. Arbitrary functions cannot be a pmf since the total probability must be 1 and all probabilities are non-negative. Hence, for $f(x)$ to be the pmf of a random variable X, we require:

1. $f(x) \geq 0$ for all possible values of x.
2. $\sum_{\text{all } x} f(x) = 1$

In Example 3.2, we may rewrite the probability function in the general form

$$f(x) = \binom{2}{x} p^x (1 - p)^{2-x}, \text{for } x = 0, 1, 2,$$

where $f(x) = 0$ for any other value of x.

3.2.4 Continuous random variables

In many situations (both theoretical and practical) we often encounter random variables that are inherently continuous because they are measured on a continuum (such as time, length, weight) or can be conveniently well-approximated by considering them as continuous (such as the annual income of adults in a population, closing share prices).

For a continuous random variable, $P(X = x)$ is defined to be zero since we assume that the measurements are continuous and there is zero probability of observing a particular value, e.g. 1.2. The argument goes that a finer measuring instrument will give us an even more precise measurement than 1.2 and so on. Thus for a continuous random variable we adopt the convention that $P(X = x) = 0$ for any particular value x on the real line. But we define probabilities for positive length intervals, e.g. $P(1.2 < X < 1.9)$.

For a continuous random variable X we define its probability by using a continuous function $f(x)$ which we call its probability density function, abbreviated as its pdf. With the pdf we define probabilities as integrals, e.g.

$$P(a < X < b) = \int_a^b f(u)du,$$

which is naturally interpreted as the area under the curve $f(x)$ inside the interval (a, b). This is demonstrated in Figure 3.1. Recall that we do not use $f(x) = P(X = x)$ for any x as by convention we set $P(X = x) = 0$.

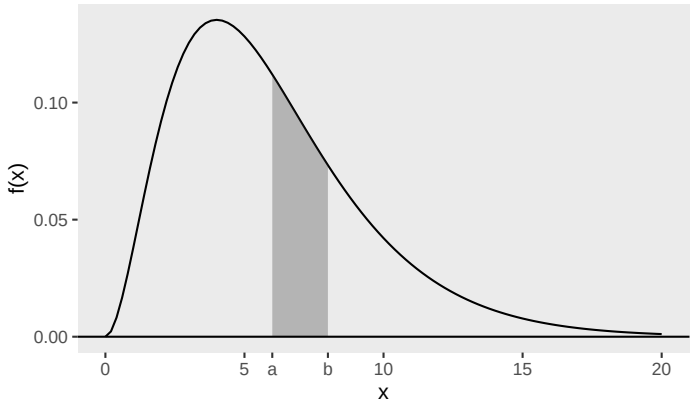


Figure 3.1: The shaded area is $P(a < X < b)$ if the pdf of X is the drawn curve.

Since we are dealing with probabilities which are always between 0 and 1, just any arbitrary function f (x) cannot be a pdf of some random variable. For $f(x)$ to be a pdf, as in the discrete case, we must have

1. $f(x) \geq 0$ for all possible values of x, i.e. $-\infty < x < \infty$,
2. $\int_{-\infty}^{\infty} f(u)du = 1$.

3.2.5 Cumulative distribution function (cdf)

Along with the pdf we also frequently make use of another function which is called the cumulative distribution function, abbreviated as the cdf. The cdf simply calculates the probability of the random variable up to its argument.

For a discrete random variable X, the cdf is the cumulative sum of the pmf $f(u)$ up to (and including)

$u = x$. That is,

$$P(X \leq x) \equiv F(x) = \sum_{u \leq x} f(u).$$

Example 3.3. Let X be the number of heads in the experiment of tossing two fair coins. Then the probability function is

$$P(X = 0) = 1/4,\ P(X = 1) = 1/2,\ P(X = 2) = 1/4.$$

From the definition, the CDF is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } 0 \leq x < 1 \\ 3/4 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

The cdf for a discrete random variable is a step function. The jump-points are the possible values of the random variable, and the height of a jump gives the probability of the random variable taking that value. It is clear that the probability mass function is uniquely determined by the cdf.

For a continuous random variable X, the cdf is defined as

$$P(X \leq x) \equiv F(x) = \int_{-\infty}^{x} f(u)du.$$

The fundamental theorem of calculus then tells us that

$$f(x) = \frac{dF(x)}{dx},$$

so for a continuous random variable the pdf is the derivative of the cdf. Also for any random variable X, $P(c < X \leq d) = F(d) - F(c)$. Let us consider an example.

Example 3.4 (Uniform distribution)**.** Suppose

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

In this case we say X has *uniform distribution* on the interval (a, b), which we will write as $X \sim U(a, b)$. We now have the cdf

$$F(x) = \int_a^x \frac{1}{b-a}du = \frac{x-a}{b-a},\ \ a < x < b.$$

A quick check confirms that $F'(x) = f(x)$. If $a = 0, b = 1$ then

$$P(0.5 < X < 0.75) = F(0.75) - F(0.5) = 0.25.$$

We shall see many more examples later.

3.3 Summaries of a random variable

3.3.1 Introduction

In Section 1.4, we defined various summaries of sample data $x_1, \ldots, x_n$, such as the mean and variance. A random variable X with either a pmf $f(x)$ or a pdf $f(x)$ may be summarised using similar measures.

3.3.2 Expectation

The mean of X is called an expectation since it is a value we can 'expect'! The expectation is defined as

$$E(X) = \begin{cases} \sum_{\text{all } x} x f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

when the sum or integral exists. They can't always be assumed to exist!

Thus, roughly speaking: the expected value is either sum or integral of value times probability. We use the $E(\cdot)$ notation to denote expectation. The argument is in upper case since it is the expected value of the random variable which is denoted by an upper case letter. We often use the Greek letter μ to denote $E(X)$.

Example 3.5. Consider the fair-die tossing experiment, with each of the six sides having a probability of 1/6 of landing face up. Let X be the number on the up-face of the die. Then

$$E(X) = \sum_{x=1}^{6} x P(X = x) = \sum_{x=1}^{6} x/6 = 3.5.$$

Example 3.6. Suppose $X \sim U(a, b)$, with pdf $f(x) = \frac{1}{b-a}$, $a < x < b$. Then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_{a}^{b} \frac{x}{b-a} dx \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}, \end{aligned}$$

the mid-point of the interval (a, b).

If $Y = g(X)$ for any function $g(\cdot)$, then Y is a random variable as well. To find $E(Y)$ we simply use the value times probability rule, i.e. the expected value of Y is either sum or integral of its value, $g(x)$, times probability $f(x)$:

$$E(Y) = E(g(X)) = \begin{cases} \sum_{\text{all } x} g(x) f(x) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

For example, if X is continuous, then $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx$. We prove an important property of expectation, namely expectation is a linear operator.

Theorem 3.1 (Linearity of expectation)**.** *Suppose $Y = g(X) = aX + b$; then $E(Y) = aE(X) + b$.*

Proof. The proof is given for the continuous case. In the discrete case replace integral ($\int$) by summation ($\sum$).

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} (ax + b) f(x) dx \\ &= a \int_{-\infty}^{\infty} x f(x) dx + b \int_{-\infty}^{\infty} f(x) dx \\ &= aE(X) + b, \end{aligned}$$

using the total probability is 1 property ($\int_{-\infty}^{\infty} f(x) dx = 1$) in the last integral. □

This is very convenient, e.g. suppose $E(X) = 5$ and $Y = -2X + 549$ then $E(Y) = 539$.

We will also prove an important property of expectation for symmetric random variables.

Theorem 3.2. *Suppose X is a random variable with a probability function or probability density function which is symmetric about some value c, so*

$$f(c+x) = f(c-x) \text{ for all } x > 0.$$

Then $E(X) = c$.

Proof. The proof is given for the continuous case. In the discrete case replace integral ($\int$) by summation ($\sum$).

First, let $Y = X - c$. Then Y is symmetric about 0, with probability density function $f(y) = f(-y)$ for all $y > 0$. Then

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} yf(y)dy \\
&= \int_{-\infty}^{0} yf(y)dy + \int_{0}^{\infty} yf(y)dy \\
&= \int_{0}^{\infty} -zf(-z)dz + \int_{0}^{\infty} yf(y)dy, \text{ substituting } z = -y \\
&= -\int_{0}^{\infty} zf(z)dz + \int_{0}^{\infty} yf(y)dy, \text{ since } f(-z) = f(z) \\
&= 0.
\end{aligned}$$

So, by the linearity of expectation, $E(X) = E(Y + c) = E(Y) + c = 0 + c = c$. □

This result makes it very easy to find the expectation of any symmetric random variable. The two examples we saw before, of a fair die and of a uniform random variable, were both symmetric, and they have expectation equal to the point of symmetry.

3.3.3 Variance

The variance measures the variability of a random variable and is defined by

$$\text{Var}(X) = E(X-\mu)^2 = \begin{cases} \sum_{\text{all } x}(x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous,} \end{cases}$$

where $\mu = E(X)$, and when the sum or integral exists. When the variance exists, it is the expectation of $(X - \mu)^2$ where μ is the mean of X. We now derive an easy formula to calculate the variance:

Theorem 3.3.

$$\text{Var}(X) = E(X-\mu)^2 = E\left(X^2\right) - \mu^2.$$

Proof. We have

$$\begin{aligned}
\text{Var}(X) &= E(X-\mu)^2 \\
&= E\left(X^2 - 2X\mu + \mu^2\right) \\
&= E\left(X^2\right) - 2\mu E(X) + \mu^2 \\
&= E\left(X^2\right) - 2\mu\mu + \mu^2 \\
&= E\left(X^2\right) - \mu^2.
\end{aligned}$$

□

Thus the variance of a random variable is the expected value of its square minus the square of its expected value.

We usually denote the variance by σ^2. The square is there to emphasise that the variance of any random variable is always non-negative. When can the variance be zero? When there is no variation at all in the random variable, i.e. it takes only a single value μ with probability 1. Hence, there is nothing random about the random variable — we can predict its outcome with certainty.

The square root of the variance is called the *standard deviation* of the random variable.

Example 3.7 (Uniform distribution)**.** Suppose $X \sim U(a, b)$, with pdf $f(x) = \frac{1}{b-a}$, $a < x < b$. Then

$$\begin{aligned} E\left(X^2\right) &= \int_a^b \frac{x^2}{b-a} dx \\ &= \frac{b^3 - a^3}{3(b-a)} \\ &= \frac{b^2 + ab + a^2}{3}. \end{aligned}$$

Hence

$$\mathrm{Var}(X) = \frac{b^2 + ab + a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12},$$

after simplification.

We prove one important property of the variance.

Theorem 3.4. *Suppose* $Y = aX + b$ *then* $\mathrm{Var}(Y) = a^2 \mathrm{Var}(X)$

Proof. Write $\mu = E(X)$. Then $E(Y) = a\mu + b$ and

$$\begin{aligned} \mathrm{Var}(Y) &= E\left[(Y - E(Y))^2\right] \\ &= E\left[(aX + b - a\mu - b)^2\right] \\ &= E\left[a^2(X-\mu)^2\right] \\ &= a^2 E\left[(X-\mu)^2\right] \\ &= a^2 \mathrm{Var}(X) \end{aligned}$$

□

This is a very useful result, e.g. suppose $\mathrm{Var}(X) = 25$ and $Y = -X + 5,000,000$; then $\mathrm{Var}(Y) = \mathrm{Var}(X) = 25$ and the standard deviation, $\sigma = 5$. In words a location shift, b, does not change variance but a multiplicative constant, a say, gets squared in variance, a^2.

3.3.4 Quantiles

For a given $0 < p < 1$, a pth *quantile* (or $100p$ *percentile*) of the random variable X with cdf $F(x)$ is defined to be a value q for which $F(q) = p$. If the cdf is invertible, we have $q = F^{-1}(p)$.

The 50th percentile is called the *median*. The 25th and 75th percentiles are called the *quartiles*.

Example 3.8 (Uniform distribution)**.** Suppose $X \sim U(a, b)$, with pdf $f(x) = \frac{1}{b-a}$, $a < x < b$. We have shown in Example 3.4 that the cdf is

$$F(x) = \frac{x-a}{b-a},\ a < x < b.$$

So for a given p, $F(q) = p$ implies

$$q = a + p(b - a).$$

The median of X is $\frac{b+a}{2}$ and the quartiles are $\frac{b+3a}{4}$ and $\frac{3b+a}{4}$.

The median of a symmetric random variable is the point of symmetry: ::: {.theorem} Suppose X is a random variable with a probability function or probability density function which is symmetric about some value c, so

$$f(c + x) = f(c - x) \text{ for all } x > 0.$$

Then the median of X is c. :::

3.4 Standard discrete distributions

3.4.1 Bernoulli distribution

A set of independent trials, where each trial has only two possible outcomes, conveniently called success (S) and failure (F), and the probability of success is the same in each trial are called a set of *Bernoulli trials*.

Suppose that we conduct one Bernoulli trial, where we get a success (S) or failure (F) with probabilities $P\{S\} = p$ and $P\{F\} = 1 - p$ respectively. Let X be an indicator of success:

$$X = \begin{cases} 1 & \text{if } S \\ 0 & \text{if } F. \end{cases}$$

Then X has Bernoulli distribution with parameter p, written as $X \sim \text{Bernoulli}(p)$.

The Bernoulli distribution has pmf

$$f(x) = p^x(1 - p)^{1-x}, x = 0, 1.$$

Hence

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p,$$
$$E\left(X^2\right) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p$$

and

$$\text{Var}(X) = E\left(X^2\right) - (E(X))^2 = p - p^2 = p(1 - p).$$

3.4.2 Binomial distribution

3.4.2.1 Introduction and definition

Suppose that we have a sequence of n Bernoulli trials such that we get a success with probability p. Let X be the number of successes in the n trials. Then X has binomial distribution with parameters n and p, written as $X \sim \text{Bin}(n, p)$.

An outcome of the experiment (of carrying out n such independent trials) is represented by a sequence of S's and F's (such as $SS \ldots FS \ldots SF$) that comprises x S 's, and $(n - x)$ F 's. The probability associated with this outcome is

$$P\{SS \ldots FS \ldots SF\} = pp \cdots (1 - p)p \cdots p(1 - p) = p^x(1 - p)^{n-x}.$$

For this sequence, $X = x$, but there are many other sequences which will also give $X = x$. In fact there are $\binom{n}{x}$ such sequences. Hence

$$P(X = x) = \binom{n}{x} p^x(1 - p)^{n-x}, x = 0, 1, \ldots, n.$$

This is the pmf of the binomial distribution with parameters n and p

How can we guarantee that $\sum_{x=0}^{n} P(X = x) = 1$? This guarantee is provided by the binomial theorem:

Theorem 3.5 (Binomial theorem)**.** *For any positive integer n and real numbers a and b,*

$$(a+b)^n = b^n + \binom{n}{1} ab^{n-1} + \cdots + \binom{n}{x} a^x b^{n-x} + \cdots + a^n.$$

To prove $\sum_{x=0}^{n} P(X = x) = 1$, i.e. $\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$, choose $a = p$ and $b = 1 - p$ in the binomial theorem.

Example 3.9. Suppose that widgets are manufactured in a mass production process with 1% defective. The widgets are packaged in bags of 10 with a money-back guarantee if more than 1 widget per bag is defective. For what proportion of bags would the company have to provide a refund?

First, we find the probability that a randomly selected bag has at most 1 defective widget. Let X be the number of defective widgets in a bag, then $X \sim \text{Bin}(n = 10, p = 0.01)$. So this probability is equal to

$$P(X = 0) + P(X = 1) = (0.99)^{10} + 10(0.01)^1(0.99)^9 = 0.9957.$$

Hence the probability that a refund is required is $1 - 0.9957 = 0.0043$, i.e. only just over 4 in 1000 bags will incur the refund on average.

Example 3.10. A binomial random variable can also be described using the urn model. Suppose we have an urn (population) containing N individuals, a proportion p of which are of type S and a proportion $1 - p$ of type F. If we select a sample of n individuals at random with replacement, then the number, X, of type S individuals in the sample follows the binomial distribution with parameters n and p.

3.4.2.2 Using R to calculate probabilities

Probabilities under all the standard distributions have been calculated in R and will be used throughout MATH1063. You will not be required to use any tables. For the binomial distribution the command

```
dbinom(x=3, size=5, prob=0.34)
```

calculates the pmf of $\text{Bin}(n = 5, p = 0.34)$ at $x = 3$, with value $P(X = 3) = \binom{5}{3}(0.34)^3(1 - 0.34)^{5-3}$. The command `pbinom` returns the cdf or the probability up to and including the argument. Thus

```
pbinom(q=3, size=5, prob=0.34)
```

will return the value of $P(X \leq 3)$ when $X \sim \text{Bin}(n = 5, p = 0.34)$. As a check, in Example 3.9, we may compute the probability that a randomly selected bag has at most 1 defective widget with the command

```
pbinom(q=1, size=10, prob=0.01)
```

```
## [1] 0.9957338
```

which matches our earlier calculations.

3.4.2.3 Expectation

Let $X \sim \text{Bin}(n, p)$. We have

$$E(X) = \sum_{x=0}^{n} xP(X = x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}.$$

Below we prove that $E(X) = np$. Recall that $k! = k(k-1)!$ for any $k > 0$.

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=1}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x} \\
&= np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-1-x+1)!}p^{x-1}(1-p)^{n-1-x+1} \\
&= np\sum_{y=0}^{n-1} \frac{(n-1)!}{y!(n-1-y)!}p^{y}(1-p)^{n-1-y} \\
&= np\sum_{y=0}^{n-1} \binom{n-1}{y}p^{y}(1-p)^{n-1-y} \\
&= np
\end{aligned}$$

where we used the substitution $y = x - 1$ and then conclude the last sum equals one as it is the sum of all probabilities in the $\text{Bin}(n-1, p)$ distribution.

3.4.2.4 Variance

Let $X \sim \text{Bin}(n, p)$. Then $\text{Var}(X) = np(1-p)$. It is difficult to find $E\left(X^2\right)$ directly, but the factorial structure allows us to find $E[X(X-1)]$. Recall that $k! = k(k-1)(k-2)!$ for any $k > 1$.

$$\begin{aligned}
E[(X(X-1)] &= \sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=2}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!}p^x(1-p)^{n-x} \\
&= n(n-1)p^2\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-2-x+2)!}p^{x-2}(1-p)^{n-2-x+2} \\
&= n(n-1)p^2\sum_{y=0}^{n-2} \frac{(n-2)!}{(y)!(n-2-y)!}p^{y}(1-p)^{n-2-y} \\
&= n(n-1)p^2
\end{aligned}$$

where we used the substitution $y = x-2$ and then conclude the last sum equals one as it is the sum of all probabilities in the $\text{Bin}(n-2, p)$ distribution. Now, $E\left(X^2\right) = E[X(X-1)] + E(X) = n(n-1)p^2 + np$. Hence,

$$\text{Var}(X) = E\left(X^2\right) - (E(X))^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

It is illuminating to see these direct proofs. Later on we shall apply statistical theory to directly prove these.

3.4.3 Geometric distribution

3.4.3.1 Introduction and definition

Suppose that we have the same situation as for the binomial distribution but we consider a different random variable X, which is defined as the number of trials that lead to the first success. The outcomes for this experiment are:

$$\begin{array}{rll} S & X=1, & P(X=1)=p \\ FS & X=2, & P(X=2)=(1-p)p \\ FFS & X=3, & P(X=3)=(1-p)^2p \\ FFFS & X=4, & P(X=4)=(1-p)^3p \\ \vdots & \vdots & \end{array}$$

In general we have

$$P(X=x)=(1-p)^{x-1}p,\ x=1,2,\ldots$$

This is called the **geometric** distribution, and it has a (countably) infinite domain starting at 1 rather than 0. We write $X \sim \text{Geo}(p)$.

Example 3.11. In a board game that uses a single fair die, a player cannot start until they have rolled a six. Let X be the number of rolls needed until they get a six. Then X is a Geometric random variable with success probability $p = 1/6$.

In order to check the probability function sums to one, we will need to use the general result on the geometric series:

Theorem 3.6 (Geometric series)**.** *For any real numbers a and r such that $|r| < 1$,*

$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}.$$

We now check the probability function sums to one:

$$\begin{aligned} \sum_{x=1}^{\infty} P(X=x) &= \sum_{x=1}^{\infty}(1-p)^{x-1}p \\ &= \sum_{y=0}^{\infty}(1-p)^{y}p \quad \text{(substitute } y=x-1\text{)} \\ &= \frac{p}{1-(1-p)} \quad \text{(geometric series, } a=p,\ r=1-p\text{)} \\ &= 1 \end{aligned}$$

We can also find the probability that $X > k$ for some given positive integer k :

$$\begin{aligned} \sum_{x=k+1}^{\infty} P(X=x) &= \sum_{x=k+1}^{\infty}(1-p)^{x-1}p \\ &= p\left[(1-p)^{k+1-1}+(1-p)^{k+2-1}+(1-p)^{k+3-1}+\ldots\right] \\ &= p(1-p)^k\sum_{y=0}^{\infty}(1-p)^{y} \\ &= (1-p)^k \end{aligned}$$

3.4.3.2 Memoryless property

Let X follow the geometric distribution and suppose that s and k are positive integers. We then have

$$P(X > s + k \mid X > k) = P(X > s).$$

The proof is given below. In practice this means that the random variable does not remember its age (denoted by k) to determine how long more (denoted by s) it will survive! The proof below uses the definition of conditional probability

$$P\{A \mid B\} = \frac{P\{A \cap B\}}{P\{B\}}.$$

Now the proof,

$$\begin{aligned} P(X > s + k \mid X > k) &= \frac{P(X > s + k, X > k)}{P(X > k)} \\ &= \frac{P(X > s + k)}{P(X > k)} \\ &= \frac{(1-p)^{s+k}}{(1-p)^k} \\ &= (1-p)^s, \end{aligned}$$

which does not depend on k. Note that the event $X > s + k$ and $X > k$ implies and is implied by $X > s + k$ since $s > 0$.

3.4.3.3 Expectation and variance

Let $X \sim \text{Geo}(p)$. We can show that $E(X) = 1/p$ using the negative binomial series:

Theorem 3.7 (Negative binomial series)**.** *For any positive integer n and real number x such that $|x| < 1$*

$$(1-x)^{-n} = 1 + nx + \frac{1}{2}n(n+1)x^2 + \frac{1}{6}n(n+1)(n+2)x^3 + \cdots + \frac{n(n+1)(n+2)\cdots(n+k-1)}{k!}x^k + \cdots$$

We have

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} xP(X = x) \\ &= \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\ &= p\left[1 + 2(1-p) + 3(1-p)^2 + 4(1-p)^3 + \ldots\right] \end{aligned}$$

The series in the square brackets is the negative binomial series with $n = 2$ and $x = 1 - p$. Thus $E(X) = p(1 - 1 + p)^{-2} = 1/p$. It can be shown that $\text{Var}(X) = (1-p)/p^2$ using negative binomial series. But this is more complicated and is not required here. The second-year module MATH2011 will provide an alternative proof.

3.4.4 Hypergeometric distribution

Suppose we have an urn (population) containing N individuals, a proportion p of which are of type S and a proportion $1-p$ of type F. If we select a sample of n individuals at random without replacement,

then the number, X, of type S individuals in the sample has the hypergeometric distribution, with pmf

$$P(X=x)=\frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}},\quad x=0,1,\ldots,n,$$

assuming that $x \le Np$ and $n-x \le N(1-p)$ so that the above combinations are well defined. The mean and variance of the hypergeometric distribution are given by

$$E(X)=np,\quad \mathrm{Var}(X)=np(1-p)\frac{N-n}{N-1}.$$

3.4.5 Negative binomial distribution

Still in the Bernoulli trials set-up, we define the random variable X to be the total number of trials until the rth success occurs, where r is a given positive integer. This is known as the negative binomial distribution with parameters p and r. [Note: if $r=1$, the negative binomial distribution is just the geometric distribution.] Firstly we need to identify the possible values of X. Possible values for X are $x=r,r+1,r+2,\ldots$. The probability mass function is

$$\begin{aligned}P(X=x)&=\binom{x-1}{r-1}p^{r-1}(1-p)^{(x-1)-(r-1)}\times p\\&=\binom{x-1}{r-1}p^{r}(1-p)^{x-r},\quad x=r,r+1,\ldots\end{aligned}$$

Example 3.12. A man plays roulette, betting on red each time. He decides to keep playing until he achieves his second win. The success probability for each game is 18/37 and the results of games are independent. Let X be the number of games played until he gets his second win. Then X is a negative binomial random variable with $r=2$ and p = 18/37.

What is the probability he plays more than 3 games? We have

$$P(X>3)=1-P(X=2)-P(X=3)=1-p^2-2p^2(1-p)=0.520.$$

Derivation of the mean and variance of the negative binomial distribution involves complicated negative binomial series and will be skipped for now, but will be proved in Section 3.7. For completeness we note down the mean and variance:

$$E(X)=\frac{r}{p},\quad \mathrm{Var}(X)=r\,\frac{1-p}{p^2}.$$

Thus when r = 1, the mean and variance of the negative binomial distribution are equal to those of the geometric distribution.

3.4.6 Poisson distribution

3.4.6.1 Introduction and definition

The Poisson distribution can be obtained as the limit of the binomial distribution with parameters n and p when $n\to\infty$ and $p\to 0$ simultaneously, but the product $\lambda=np$ remains finite. In practice this means that the Poisson distribution counts rare events (since $p\to 0$) in an infinite population (since $n\to\infty$). Theoretically, a random variable following the Poisson distribution can take any integer value from 0 to ∞. Examples of the Poisson distribution include: the number of breast cancer patients in Southampton; the number of text messages sent (or received) per day by a randomly selected first-year student; the number of credit cards a randomly selected person has in their wallet.

Let us find the pmf of the Poisson distribution as the limit of the pmf of the binomial distribution. Recall that if $X \sim \text{Bin}(n,p)$ then $P(X=x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$. Now:

$$\begin{aligned} P(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \frac{n^n}{n^n} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{n^x x!} (np)^x (n(1-p))^{n-x} \frac{1}{n^{n-x}} \\ &= \frac{n}{n}\frac{(n-1)}{n}\cdots\frac{(n-x+1)}{n}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n}{n}\frac{(n-1)}{n}\cdots\frac{(n-x+1)}{n}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-x}. \end{aligned}$$

Now it is easy to see that the above tends to

$$e^{-\lambda}\frac{\lambda^x}{x!}$$

as $n \to \infty$ for any fixed value of x in the range $0, 1, 2, \ldots$. Note that we have used the exponential limit:

$$e^{-\lambda} = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n,$$

and

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1$$

and

$$\lim_{n\to\infty}\frac{n}{n}\frac{(n-1)}{n}\cdots\frac{(n-x+1)}{n} = 1.$$

A random variable X has the Poisson distribution with parameter λ if it has the pmf:

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

We write $X \sim \text{Poisson}(\lambda)$. It is easy to show $\sum_{x=0}^{\infty} P(X=x) = 1$, i.e. $\sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = 1$. The identity you need is simply the expansion of e^{λ}.

3.4.6.2 Expectation

Let $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned}
E(X) &= \sum_{x=0}^{\infty} xP(X=x) \\
&= \sum_{x=0}^{\infty} xe^{-\lambda}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=1}^{\infty} \frac{\lambda\cdot\lambda^{(x-1)}}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{x=1}^{\infty} \frac{\lambda^{(x-1)}}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \quad (y=x-1) \\
&= \lambda e^{-\lambda}e^{\lambda} \quad \text{using the expansion of } e^{\lambda} \\
&= \lambda.
\end{aligned}$$

3.4.6.3 Variance

Let $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned}
E[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1)P(X=x) \\
&= \sum_{x=0}^{\infty} x(x-1)e^{-\lambda}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=2}^{\infty} x(x-1)\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=2}^{\infty} \lambda^2\frac{\lambda^{x-2}}{(x-2)!} \\
&= \lambda^2 e^{-\lambda}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \quad (y=x-2) \\
&= \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2 \quad \text{using the expansion of } e^{\lambda}.
\end{aligned}$$

Now, $E\left(X^2\right) = E[X(X-1)] + E(X) = \lambda^2 + \lambda$. Hence,

$$\text{Var}(X) = E\left(X^2\right) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Hence, the mean and variance are the same for the Poisson distribution.

3.4.6.4 Using R to calculate probabilities

For the Poisson distribution the command

```
dpois(x=3, lambda=5)
```

calculates the pmf of Poisson($\lambda = 5$) at $x = 3$. That is, the command will return the value $P(X = 3) = e^{-5}\frac{5^3}{3!}$. The command `ppois` returns the cdf or the probability up to and including the argument. Thus

```
ppois(q=3, lambda=5)
```

will return the value of $P(X \leq 3)$ when $X \sim$ Poisson($\lambda = 5$).

3.5 Standard continuous distributions

3.5.1 Uniform distribution

3.5.1.1 Definition and properties

A continuous random variable X is said to follow the uniform distribution if its pdf is of the form:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

where $a < b$ are parameters. We write $X \sim U(a, b)$.

We have already derived various properties the uniform distribution:

- **cumulative distribution function**:

  $$F(x) = \frac{x-a}{b-a},\ a < x < b$$

  from Example 3.4.
- **expectation**:

  $$E(X) = \frac{b+a}{2}$$

  from Example 3.6.
- **variance**:

  $$\text{Var}(x) = \frac{(b-a)^2}{12}$$

  from Example 3.7.
- **quantiles**: The pth quantile is $a + p(b-a)$ from Example 3.8. The median is $\frac{b+a}{2}$.

3.5.1.2 Using R to calculate probabilities

For $X \sim U(a = -1, b = 1)$, the command

```
dunif(x = 0.5, min = -1, max = 1)
```

calculates the pdf at $x = 0.5$. We specify a with the `min` argument and b with `max` argument. The command `punif` returns the cdf or the probability up to and including the argument. Thus

```
punif(q = 0.5, min = -1, max = 1)
```

will return the value of $P(X \leq 0.5)$.

The command `qunif` can be used to calculate quantiles. Thus

```
qunif(p = 0.5, min = -1, max = 1)
```

finds the median (the 0.5 quantile).

3.5.2 Exponential distribution

3.5.2.1 Introduction and definition

A continuous random variable X is said to follow the exponential distribution if its pdf is of the form:

$$f(x) = \begin{cases} \theta e^{-\theta x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

where $\theta > 0$ is a parameter. We write $X \sim \text{Exponential}(\theta)$. The distribution only resides in the positive half of the real line, and the tail goes down to zero exponentially as $x \to \infty$. The rate at which that happens is the parameter θ. Hence θ is known as the rate parameter.

It is easy to prove that $\int_0^\infty f(x)dx = 1$. This is left as an exercise. To find the mean and variance of the distribution we need to introduce the *gamma function*:

The gamma function $\Gamma(.)$ is defined for any positive number a as

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

We have the following facts:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}; \quad \Gamma(1) = 1; \quad \Gamma(a) = (a-1)\Gamma(a-1) \text{ if } a > 1$$

These last two facts imply that $\Gamma(k) = (k-1)!$ when k is a positive integer. Find $\Gamma\left(\frac{3}{2}\right)$.

3.5.2.2 Expectation and variance

By definition,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_0^\infty x\theta e^{-\theta x}dx \\ &= \int_0^\infty ye^{-y}\frac{dy}{\theta} \quad \text{(substitute } y = \theta x\text{)} \\ &= \frac{1}{\theta}\int_0^\infty y^{2-1}e^{-y}dy \\ &= \frac{1}{\theta}\Gamma(2) \\ &= \frac{1}{\theta} \quad \text{since } \Gamma(2) = 1! = 1. \end{aligned}$$

Now,

$$\begin{aligned} E\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 f(x) dx \\ &= \int_0^{\infty} x^2 \theta e^{-\theta x} dx \\ &= \theta \int_0^{\infty} \left(\frac{y}{\theta}\right)^2 e^{-y} \frac{dy}{\theta} \quad \text{(substitute } y = \theta x\text{)} \\ &= \frac{1}{\theta^2} \int_0^{\infty} y^{3-1} e^{-y} dy \\ &= \frac{1}{\theta^2} \Gamma(3) \\ &= \frac{2}{\theta^2} \quad \text{since } \Gamma(3) = 2! = 2, \end{aligned}$$

and so

$$\text{Var}(X) = E\left(X^2\right) - [E(X)]^2 = 2/\theta^2 - 1/\theta^2 = 1/\theta^2.$$

For this random variable the mean is equal to the standard deviation.

3.5.2.3 Using R to calculate probabilities

For $X \sim$ Exponential($\theta = 0.5$), the command

```
dexp(x = 3, rate = 1/2)
```

calculates the pdf at $x = 3$. The rate parameter to be supplied is the θ parameter here. The command `pexp` returns the cdf or the probability up to and including the argument. Thus

```
pexp(q = 3, rate = 1/2)
```

will return the value of $P(X \leq 3)$.

3.5.2.4 Cumulative distribution function and quantiles

It is easy to find the cdf of the exponential distribution. For $x > 0$,

$$F(x) = P(X \leq x) = \int_0^x \theta e^{-\theta u} du = 1 - e^{-\theta x}.$$

We have $F(0) = 0$ and $F(x) \to 1$ when $x \to \infty$ and $F(x)$ is non-decreasing in x. The cdf can be used to solve many problems. A few examples follow.

Example 3.13 (Mobile phone)**.** Suppose that the lifetime of a phone (e.g. the time until the phone does not function even after repairs), denoted by X, manufactured by the company A Pale, is exponentially distributed with mean 550 days. 1. Find the probability that a randomly selected phone will still function after two years, i.e. $X > 730$? (Assume there is no leap year in the two years.) 2. What are the times by which we expect $25\%, 50\%, 75\%$ and 90% of the manufactured phones to have failed?

Here the mean $1/\theta = 550$. Hence $\theta = 1/550$ is the rate parameter. The solution to the first problem is

$$P(X > 730) = 1 - P(X \leq 730) = 1 - \left(1 - e^{-730/550}\right) = e^{-730/550} = 0.2652.$$

Alternatively, we can do the calculation in R:

```
1 - pexp(q = 730, rate = 1 / 550)
```

```
## [1] 0.2651995
```

For the second problem we are given the probabilities of failure (0.25, 0.50, etc.). We will have to invert the probabilities to find the value of the random variable. In other words, we will have to find a q such that $F(q) = p$, where p is the given probability: the pth quantile of X.

The cdf of the exponential distribution is $F(q) = 1 - e^{-\theta q}$, so to find the pth quantile we must solve $p = 1 - e^{-\theta q}$ for q.

$$\begin{aligned} p &= 1 - e^{-\theta q} \\ \Rightarrow e^{-\theta q} &= 1 - p \\ \Rightarrow -\theta q &= \log(1-p) \\ \Rightarrow q &= \frac{-\log(1-p)}{\theta}. \end{aligned}$$

In Example 3.13, $\theta = 1/550$, so we have

$$q = -550 \times \log(1-p) = \begin{cases} 158 & \text{for } p = 0.25 \\ 381 & \text{for } p = 0.50 \\ 762 & \text{for } p = 0.75 \\ 1266 & \text{for } p = 0.90. \end{cases}$$

which gives the time in days until we expect $25\%, 50\%, 75\%$ and 90% of the manufactured phones to have failed.

In R you can find these values by

```
qexp(p = 0.25, rate = 1/550)
qexp(p = 0.50, rate = 1/550)
```

and so on. The function

```
qexp(p, rate)
```

calculates the pth quantile of the exponential distribution with parameter `rate`.

Example 3.14 (Survival function)**.** The exponential distribution is sometimes used to model the survival times in different experiments. For example, an exponential random variable T may be assumed to model the number of days a cancer patient survives after chemotherapy. In such a situation, the function $S(t) = 1 - F(t) = e^{-\theta t}$ is called the survival function. See Figure 3.2 for an example plot.

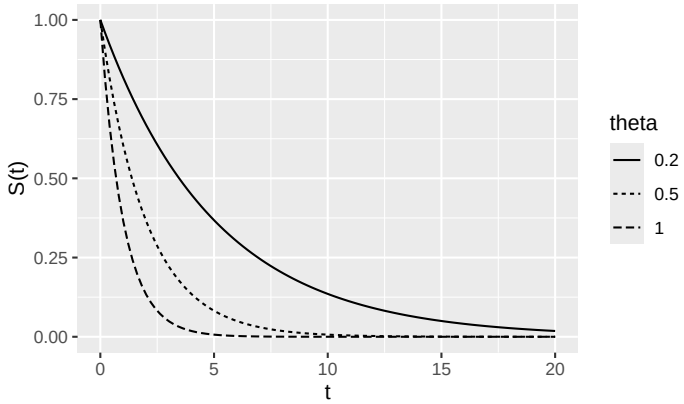


Figure 3.2: $S(t)$ for $\theta = 0.2, 0.5, 1$.

Assuming the mean survival time to be 100 days for a fatal late detected cancer, we can expect that half of the patients survive 69.3 days after chemo since

```
qexp(0.50, rate=1/100)
```

```
## [1] 69.31472
```

You will learn more about this in a third-year module, MATH3085: Survival Models, which is important in the actuarial profession.

3.5.2.5 Memoryless property

Like the geometric distribution, the exponential distribution also has the memoryless property. In simple terms, it means that the probability that the system will survive an additional period $s > 0$ given that it has survived up to time t is the same as the probability that the system survives the period s to begin with. That is, it forgets that it has survived up to a particular time when it is thinking of its future remaining life time.

The proof is exactly as in the case of the geometric distribution, reproduced below. Recall the definition of conditional probability:

$$P\{A \mid B\} = \frac{P\{A \cap B\}}{P\{B\}}.$$

Now the proof,

$$\begin{aligned} P(X > s+t \mid X > t) &= \frac{P(X > s+t, X > t)}{P(X > t)} \\ &= \frac{P(X > s+t)}{P(X > t)} \\ &= \frac{e^{-\theta(s+t)}}{e^{-\theta t}} \\ &= e^{-\theta s} \\ &= P(X > s). \end{aligned}$$

Note that the event $X > s+t$ and $X > t$ implies and is implied by $X > s+t$ since $s > 0$.

Example 3.15. Suppose the time T between any two successive arrivals in a hospital emergency department has exponential distribution, $T \sim \text{Exponential}(\lambda)$. Historically, the mean of these inter-arrival times is 5 minutes. Estimate λ, and hence estimate

(i) $P(0 < T < 5)$,
(ii) $P(T < 10 \mid T > 5)$.

An estimate of $E(T)$ is 5. As $E(T) = \frac{1}{\lambda}$ we take $\frac{1}{5}$ as the estimate of λ.

(i) $P(0 < T < 5) = \int_0^5 \frac{1}{5} e^{-t/5} dt = \left[-e^{-t/5}\right]_0^5 = 1 - e^{-1} = 0.63212.$
(ii) We have

$$\begin{aligned} P(T < 10 \mid T > 5) &= \frac{P(5 < T < 10)}{P(T > 5)} \\ &= \frac{\int_5^{10} \frac{1}{5} e^{-t/5} dt}{\int_5^{\infty} \frac{1}{5} e^{-t/5} dt} = \frac{\left[-e^{-t/5}\right]_5^{10}}{\left[-e^{-t/5}\right]_5^{\infty}} \\ &= 1 - e^{-1} = 0.63212. \end{aligned}$$

3.5.3 Normal distribution

3.5.3.1 Definition

A random variable X is said to have the normal distribution with parameters μ and σ^2 if it has pdf of the form

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},\ -\infty < x < \infty \tag{3.1}$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are two given constants. We write $X \sim N\left(\mu, \sigma^2\right)$.

We will show later that $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

The density (pdf) is much easier to remember and work with when the mean $\mu = 0$ and variance $\sigma^2 = 1$. This special case is called the *standard* normal distribution. In this case, we simply write:

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

We often use Z to denote a random variable with standard normal distribution.

It is easy to see that $f(x) > 0$ for all x. Next, we show $\int_{-\infty}^{\infty} f(x)dx = 1$ or total probability equals 1, so that $f(x)$ defines a valid pdf:

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz \quad \text{(substitute } z = \frac{x-\mu}{\sigma} \text{ so that } dx = \sigma dz\text{)} \\
&= \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz \quad \text{(since the integrand is an even function)} \\
&= \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{\infty} \exp\{-u\} \frac{du}{\sqrt{2u}} \text{ (substitute } u = \frac{z^2}{2} \text{ so that } z = \sqrt{2u} \text{ and } dz = \frac{du}{\sqrt{2u}}\text{)} \\
&= \frac{1}{2\sqrt{\pi}} 2 \int_{0}^{\infty} u^{\frac{1}{2}-1} \exp\{-u\} du \quad \text{(rearrange the terms)} \\
&= \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right) \quad \text{(recall the definition of the Gamma function)} \\
&= \frac{1}{\sqrt{\pi}} \sqrt{\pi} = 1 \quad \text{as } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.
\end{aligned}$$

3.5.3.2 Linear transformations

Theorem 3.8. *Suppose $X \sim N\left(\mu, \sigma^2\right)$ and a and b are constants. Then the distribution of $Y = aX + b$ is $N\left(a\mu + b, a^2\sigma^2\right)$.*

Proof. The cumulative distribution function of Y is

$$\begin{aligned}
F_Y(y) = P(Y \le y) = P(aX + b \le y) &= P\left(X \le \frac{y-b}{a}\right) \\
&= \int_{-\infty}^{\frac{y-b}{a}} f_X(x)dx, \text{ where } f_X(x) \text{ is the pdf of } X \\
&= \int_{-\infty}^{y} \frac{1}{a} f_X\left(\frac{u-b}{a}\right) du, \text{ substituting } u = ax + b.
\end{aligned}$$

So Y has probability density function

$$\begin{aligned} f_Y(y) &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{a}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\left(\frac{y-b}{a}-\mu\right)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi a^2\sigma^2}}\exp\left\{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}\right\}, \end{aligned}$$

which is the $N(a\mu+b, a^2\sigma^2)$ probability density function. So $Y \sim N(a\mu+b, a^2\sigma^2)$. □

An important consequence is that if $X \sim N(\mu, \sigma^2)$, then $Z = (X-\mu)/\sigma \sim N(0,1)$. We can "standardise" any normal random variable by subtracting the mean μ then dividing by the standard deviation σ.

3.5.3.3 Expectation and variance

We claimed that $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$, and now will prove these results.

$E(X) = \mu$ because $f(x)$ is symmetric about μ.

To prove $\text{Var}(X) = \sigma^2$, we show that $\text{Var}(Z) = 1$ where $Z = \frac{X-\mu}{\sigma}$. Once we have shown that, we will have $\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2$.

Since $E(Z) = 0, \text{Var}(Z) = E\left(Z^2\right)$, which is calculated below:

$$\begin{aligned} E\left(Z^2\right) &= \int_{-\infty}^{\infty} z^2 f(z)dz \\ &= \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2}\right\}dz \\ &= \frac{2}{\sqrt{2\pi}}\int_0^{\infty} z^2 \exp\left\{-\frac{z^2}{2}\right\}dz \quad \text{(since the integrand is an even function)} \\ &= \frac{2}{\sqrt{2\pi}}\int_0^{\infty} 2u\exp\{-u\}\frac{du}{\sqrt{2u}} \quad \text{(substitute } u = \frac{z^2}{2} \text{ so that } z = \sqrt{2u} \text{ and } dz = \frac{du}{\sqrt{2u}}) \\ &= \frac{4}{2\sqrt{\pi}}\int_0^{\infty} u^{\frac{1}{2}}\exp\{-u\}du \\ &= \frac{2}{\sqrt{\pi}}\int_0^{\infty} u^{\frac{3}{2}-1}\exp\{-u\}du \\ &= \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) \quad \text{(definition of the gamma function)} \\ &= \frac{2}{\sqrt{\pi}}\left(\frac{3}{2}-1\right)\Gamma\left(\frac{3}{2}-1\right) \quad \text{(reduction property of the gamma function)} \\ &= \frac{2}{\sqrt{\pi}}\frac{1}{2}\sqrt{\pi} \quad (\text{since } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}) \\ &= 1, \end{aligned}$$

as we hoped for! This proves $\text{Var}(X) = \sigma^2$.

3.5.3.4 Calculating probabilities

Suppose $X \sim N\left(\mu, \sigma^2\right)$ and we are interested in finding $P(a \le X \le b)$ for two constants a and b. To do this, we can use the fact that $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ and rewrite the probability of interest in terms

of standard normal probabilities:

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{b-\mu}{\sigma}\right) - P\left(Z \leq \frac{a-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right), \end{aligned}$$

where we use the notation $\Phi(\cdot)$ to denote the cdf of the standard normal distribution, i.e.

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} du.$$

This result allows us to find the probabilities about a normal random variable X of any mean μ and variance σ^2 through the probabilities of the standard normal random variable Z. For this reason, only $\Phi(z)$ is tabulated. Further more, due to the symmetry of the pdf of $Z, \Phi(z)$ is tabulated only for positive z values. Suppose $a > 0$, then

$$\begin{aligned} \Phi(-a) = P(Z \leq -a) &= P(Z > a) \\ &= 1 - P(Z \leq a) \\ &= 1 - \Phi(a). \end{aligned}$$

In R, we use the function `pnorm` to calculate the probabilities. `pnorm` which has arguments `mean` (to specify μ) and `sd` (to specify σ, the standard deviation). By default `mean = 0` and `sd = 1`, so by default `pnorm` calculates the standard normal cdf. So, we use the command

```
pnorm(1)
```

```
## [1] 0.8413447
```

to calculate $\Phi(1) = P(Z \leq 1)$. We can also use the command

```
pnorm(15, mean=10, sd=2)
```

```
## [1] 0.9937903
```

to calculate $P(X \leq 15)$ when $X \sim N\left(\mu = 10, \sigma^2 = 4\right)$ directly.

1. $P(-1 < Z < 1) = \Phi(1) - \Phi(-1) = 0.6827$. This means that 68.27% of the probability lies within 1 standard deviation of the mean.
2. $P(-2 < Z < 2) = \Phi(2) - \Phi(-2) = 0.9545$. This means that 95.45% of the probability lies within 2 standard deviations of the mean.
3. $P(-3 < Z < 3) = \Phi(3) - \Phi(-3) = 0.9973$. This means that 99.73% of the probability lies within 3 standard deviations of the mean.

We are often interested in the quantiles (inverse-cdf of probability), $\Phi^{-1}(\cdot)$ of the normal distribution for various reasons. We find the pth quantile by issuing the R command `qnorm(p)`

1. `qnorm(0.95)` $= \Phi^{-1}(0.95) = 1.645$. This means that the 95th percentile of the standard normal distribution is 1.645. This also means that $P(-1.645 < Z < 1.645) = \Phi(1.645) - \Phi(-1.645) = 0.90$.
2. `qnorm(0.975)` $= \Phi^{-1}(0.975) = 1.96$. This means that the 97.5th percentile of the standard normal distribution is 1.96. This also means that $P(-1.96 < Z < 1.96) = \Phi(1.96) - \Phi(-1.96) = 0.95$.

Example 3.16. Suppose the marks in MATH1063 follow the normal distribution with mean 58 and standard deviation 32.25.

1. What percentage of students will fail (i.e. score less than 40) in MATH1063?
   Answer: `pnorm(40, mean=58, sd=32.25)` = 28.84%.
2. What percentage of students will get an A result (score greater than 70)?
   Answer: `1 - pnorm(70, mean=58, sd=32.25)` = 35.49%.
3. What is the probability that a randomly selected student will score more than 90?
   Answer: `1 - pnorm(90, mean=58, sd=32.25)` = 0.1605.
4. What is the probability that a randomly selected student will score less than 25?
   Answer: `pnorm(25, mean=58, sd=32.25) = 0.1531`. Ouch!
5. What is the probability that a randomly selected student scores a 2:1 (i.e. a mark between 60 and 70)? Left as an exercise.

Example 3.17. A lecturer set and marked an examination and found that the distribution of marks was $N(42, 14^2)$. The school's policy is to present scaled marks whose distribution is $N(50, 15^2)$. What linear transformation should the lecturer apply to the raw marks to accomplish this and what would the raw mark of 40 be transformed to?

Let X be the raw mark and Y the scaled mark. We have $X \sim N(\mu_x = 42, \sigma_x^2 = 14^2)$ and aim to define the scaling such that $Y \sim N(\mu_y = 50, \sigma_y^2 = 15^2)$. If we standardise both variables, they should each have standard normal distribution, so we choose Y such that

$$\frac{X - \mu_x}{\sigma_x} = \frac{Y - \mu_y}{\sigma_y}$$

giving

$$Y = \mu_y + \frac{\sigma_y}{\sigma_x}(X - \mu_x) = 50 + \frac{15}{14}(X - 42).$$

Now at raw mark $X = 40$, the transformed mark would be

$$Y = 50 + \frac{15}{14}(40 - 42) = 47.86.$$

3.5.3.5 Log-normal distribution

If $X \sim N(\mu, \sigma^2)$ then the random variable $Y = exp(X)$ is called a log-normal random variable and its distribution is called a log-normal distribution with parameters μ and σ^2.

The mean of the random variable Y is given by

$$\begin{aligned}
E(Y) &= E[\exp(X)] \\
&= \int_{-\infty}^{\infty} \exp(x) \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \exp\left\{-\frac{\mu^2 - \left(\mu + \sigma^2\right)^2}{2\sigma^2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{x^2 - 2\left(\mu + \sigma^2\right)x + \left(\mu + \sigma^2\right)^2}{2\sigma^2}\right\} dx \\
&= \exp\left\{-\frac{\mu^2 - \left(\mu + \sigma^2\right)^2}{2\sigma^2}\right\} \quad \text{(integrating a } N(\mu + \sigma^2, \sigma^2) \text{ random variable over its domain)} \\
&= \exp\left\{\mu + \sigma^2/2\right\}.
\end{aligned}$$

Similarly, one can show that

$$\begin{aligned}
E(Y^2) &= E[\exp(2X)] \\
&= \int_{-\infty}^{\infty} \exp(2x) \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \cdots \\
&= \exp\left\{2\mu + 2\sigma^2\right\}.
\end{aligned}$$

Hence, the variance is given by

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = \exp\left\{2\mu + 2\sigma^2\right\} - \exp\left\{2\mu + \sigma^2\right\}.$$

The log-normal distribution is often used in practice for modelling economic variables of interest in business and finance, e.g. volume of sales, income of individuals. You do not need to remember the mean and variance of the log-normal distribution.

3.6 Joint distributions

3.6.1 Introduction

Often we need to study more than one random variable, e.g. height and weight, simultaneously, so that we can exploit the relationship between them to make inferences about their properties. Multiple random variables are studied through their joint probability distribution. In this section we will study covariance and correlation and then discuss when random variables are independent.

3.6.2 Joint distribution of discrete random variables

If X and Y are discrete, the quantity $f(x, y) = P(X = x \cap Y = y)$ is called the joint probability mass function (joint pmf) of X and Y. To be a joint pmf, $f(x, y)$ needs to satisfy two conditions:

$$f(x, y) \geq 0 \quad \text{for all } x \text{ and } y$$

and

$$\sum_{\text{all } x} \sum_{\text{all } y} f(x, y) = 1.$$

The marginal probability mass functions (marginal pmfs) of X and Y are respectively

$$f_X(x) = \sum_y f(x, y), \quad f_Y(y) = \sum_x f(x, y)$$

Use the identity $\sum_x \sum_y f(x, y) = 1$ to prove that $f_X(x)$ and $f_Y(y)$ are really pmfs.

Example 3.18. Suppose that two fair dice are tossed independently one after the other. Let

$$X = \begin{cases} -1 & \text{if the result from die 1 is larger} \\ 0 & \text{if the results are equal} \\ 1 & \text{if the result from die 1 is smaller} \end{cases}$$

and let $Y = |\text{difference between the two dice}|$. Find the joint probability pmf for X and Y.

There are 36 possible outcomes for the results of the dice rolls, and each gives a pair of values (x, y) for X and Y.

	1	2	3	4	5	6
1	(0, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)
2	(-1, 1)	(0, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)
3	(-1, 2)	(-1, 1)	(0, 0)	(1, 1)	(1, 2)	(1, 3)
4	(-1, 3)	(-1, 2)	(-1, 1)	(0, 0)	(1, 1)	(1, 2)
5	(-1, 4)	(-1, 3)	(-1, 2)	(-1, 1)	(0, 0)	(1, 1)
6	(-1, 5)	(-1, 4)	(-1, 3)	(-1, 2)	(-1, 1)	(0, 0)

			y			
x	0	1	2	3	4	5
-1	$\frac{0}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
0	$\frac{6}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$
1	$\frac{0}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Table 3.1: The joint probabilities for X and Y

The joint pmf is given in Table 3.1.

The marginal pmf for X is

$$f_X(x) = \begin{cases} \frac{15}{36} & \text{if } x = -1 \\ \frac{6}{36} & \text{if } x = 0 \\ \frac{15}{36} & \text{if } x = 1. \end{cases}$$

Exercise: Write down the marginal distribution of Y and hence find the mean and variance of Y.

3.6.3 Joint distribution of continuous random variables

If X and Y are continuous, a non-negative real-valued function $f(x, y)$ is called the joint probability density function (joint pdf) of X and Y if

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx\, dy = 1$$

The marginal pdfs of X and Y are respectively

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

Example 3.19. Define a joint pdf by

$$f(x, y) = \begin{cases} 6xy^2 & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

How can we show that the above is a pdf? It is non-negative for all x and y values. But does it integrate to 1? We are going to use the following rule.

Result: Suppose that a real-valued function $f(x, y)$ is continuous in a region A, where $A = \{(x, y) \text{ such that } a < x < b \text{ and } c < y < d\}$. Then

$$\int_A f(x, y) dx dy = \int_c^d \int_a^b f(x, y) dx\, dy.$$

The same result holds if a and b depend upon y, but c and d should be free of x and y. When we evaluate the inner integral $\int_a^b f(x, y) dx$, we treat y as constant.

Notes: To evaluate a bivariate integral over a region A we:

- Draw a picture of A whenever possible.
- Rewrite the region A as an intersection of two one-dimensional intervals. The first interval is obtained by treating one variable as constant.

- Perform two one-dimensional integrals.

Example 3.20. Continuing Example 3.19,

$$\begin{aligned}\int_0^1\int_0^1 f(x,y)dxdy &= \int_0^1\int_0^1 6xy^2dxdy\\ &= 6\int_0^1 y^2dy\int_0^1 xdx\\ &= 3\int_0^1 y^2dy\left[\text{ as }\int_0^1 xdx=\frac{1}{2}\right]\\ &= 1.\left[\text{ as }\int_0^1 y^2dy=\frac{1}{3}\right]\end{aligned}$$

Now we can find the marginal pdfs as well.

$$f_X(x)=2x,\ 0<x<1,\quad f_Y(y)=3y^2,\ 0<y<1.$$

The probability of any event in the two-dimensional space can be found by integration and again more details will be provided in the second-year module MATH2011, Statistical Distribution Theory. You will come across multivariate integrals in the second semester module MATH1060, Multivariable Calculus.

3.6.4 Covariance and correlation

We first define the expectation of a real-valued scalar function $g(X,Y)$ of X and Y :

$$E[g(X,Y)]=\begin{cases}\sum_x\sum_y g(x,y)f(x,y) & \text{if } X \text{ and } Y \text{ are discrete}\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy & \text{if } X \text{ and } Y \text{ are continuous.}\end{cases}$$

Example 3.21. Continuing Example 3.18, let $g(x,y)=xy$.

$$E(XY)=(-1)(0)0+(-1)(1)\frac{5}{36}+\cdots+(1)(5)\frac{1}{36}=0.$$

Exercises: Try $g(x,y)=x$. It will be the same thing as $E(X)=\sum_x xf_X(x)$.

We will not consider any continuous examples as the second-year module MATH2011 will study them in detail.

Suppose that two random variables X and Y have joint pmf or pdf $f(x,y)$ and let $E(X)=\mu_x$ and $E(Y)=\mu_y$. The covariance between X and Y is defined by

$$\text{Cov}(X,Y)=E\left[(X-\mu_x)(Y-\mu_y)\right]=E(XY)-\mu_x\mu_y.$$

Let $\sigma_x^2=\text{Var}(X)=E\left(X^2\right)-\mu_x^2$ and $\sigma_y^2=\text{Var}(Y)=E\left(Y^2\right)-\mu_y^2$. The correlation coefficient between X and Y is defined by:

$$\text{Corr}(X,Y)=\frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}=\frac{E(XY)-\mu_x\mu_y}{\sigma_x\sigma_y}.$$

It can be proved that for any two random variables, $-1\le\text{Corr}(X,Y)\le 1$. The correlation $\text{Corr}(X,Y)$ is a measure of linear dependency between two random variables X and Y, and it is free of the measuring units of X and Y as the units cancel in the ratio.

3.6.5 Independence

Independence is an important concept. Recall that we say two events A and B are independent if $P(A \cap B) = P(A) \times P(B)$. We use the same idea here. Two random variables X and Y having the joint pdf or pmf $f(x, y)$ are said to be independent if and only if $f(x, y) = f_X(x) \times f_Y(y)$ for *all* x and y.

In the discrete case X and Y are independent if each cell probability, $f(x, y)$, is the product of the corresponding row and column totals. In Example 3.18 X and Y are not independent.

Example 3.22. Suppose X and Y have joint pdf given by the probability table:

		y			
		1	2	3	Total
x	0	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{3}$
	1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
	2	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{6}$
	Total	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1

Verify that in the following example X and Y are independent. We need to check all 9 cells.

Example 3.23. Let $f(x, y) = 6xy^2, 0 < x < 1, 0 < y < 1$. Check that X and Y are independent.

Example 3.24. Let $f(x, y) = 2x, 0 \leq x \leq 1, 0 \leq y \leq 1$. Check that X and Y are independent.

Sometimes the joint pdf may look like something you can factorise, but X and Y may not be independent because they may be related in the domain. For instance,

1. $f(x, y) = \frac{21}{4}x^2y, x^2 \leq y \leq 1$. Not independent!
2. $f(x, y) = e^{-y}, 0 < x < y < \infty$. Not independent!

Here are some useful consequences of independence:

- Suppose that X and Y are independent random variables. Then

  $$P(X \in A, Y \in B) = P(X \in A) \times P(Y \in B)$$

  for any events A and B. That is, the joint probability can be obtained as the product of the marginal probabilities. We will use this result in the next section. For example, suppose Jack and Jess are two randomly selected students. Let X denote the height of Jack and Y denote the height of Jess. Then

  $$P(X < 182 \text{ and } Y > 165) = P(X < 182) \times P(Y > 165).$$

  This is true for any numbers other than the example numbers 182 and 165, and for any inequalities.
- Let $g(x)$ be a function of x only and $h(y)$ be a function of y only. Then, if X and Y are independent,

  $$E[g(X)h(Y)] = E[g(X)] \times E[h(Y)].$$

  As a special case, let $g(x) = x$ and $h(y) = y$. Then we have

  $$E(XY) = E(X) \times E(Y).$$

  Consequently, for independent random variables X and $Y, \text{Cov}(X, Y) = 0$ and $\text{Corr}(X, Y) = 0$. But the converse is not true in general. That is, merely having $\text{Corr}(X, Y) = 0$ does not imply that X and Y are independent random variables.

3.7 Sums of random variables

In this section we consider sums of random variables, which arise frequently in both practice and theoretical results. For example, the mark achieved in an exam is the sum of the marks for each question, and the sample mean is proportional to the sum of the sample values.

Suppose we have obtained a random sample from a distribution with pmf or pdf $f(x)$, so that X can either be a discrete or a continuous random variable. We will learn more about random sampling in the next chapter. Let $X_1, \ldots, X_n$ denote the random sample of size n where n is a positive integer. We use upper case letters since each member of the random sample is a random variable. For example, I toss a fair coin n times and let X_i take the value 1 if a head appears in the i th trial and 0 otherwise. Now I have a random sample $X_1, \ldots, X_n$ from the Bernoulli distribution with probability of success equal to 0.5 since the coin is assumed to be fair.

We can get a random sample from a continuous random variable as well. Suppose it is known that the distribution of the heights of first-year students is normal with mean 175 centimetres and standard deviation 8 centimetres. I can randomly select a number of first-year students and record each student's height.

Suppose $X_1, \ldots, X_n$ is a random sample from a population with distribution $f(x)$. Then it can be shown that the random variables $X_1, \ldots, X_n$ are mutually independent, i.e.

$$P\left(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\right) = P\left(X_1 \in A_1\right) \times P\left(X_2 \in A_2\right) \times \cdots P\left(X_n \in A_n\right)$$

for any set of events, $A_1, A_2, \ldots A_n$. That is, the joint probability can be obtained as the product of individual probabilities. An example of this for $n = 2$ was given in Section 3.6.5.

Example 3.25 (Distribution of the sum of independent binomial random variables)**.** Suppose $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, p)$ independently. Note that p is the same in both distributions. Using the above fact that joint probability is the multiplication of individual probabilities, we can conclude that $Z = X + Y$ has the binomial distribution. It is intuitively clear that this should happen since X comes from m Bernoulli trials and Y comes from n Bernoulli trials independently, so Z comes from $m + n$ Bernoulli trials with common success probability p.

Next we will prove the result mathematically, by finding the probability mass function of $Z = X + Y$ directly and observing that it is of the appropriate form. In our proof, we will need to use the fact that

$$\sum_{x+y=z} \binom{m}{x}\binom{n}{y} = \binom{m+n}{z}$$

where the above sum is also over all possible integer values of x and y such that $0 \leq x \leq m$ and $0 \leq y \leq n$. This fact may be proved by using the binomial theorem, but we state it here without proof.

Note that

$$P(Z = z) = P(X = x, Y = y)$$

subject to the constraint that $x + y = z$, $0 \leq x \leq m$, $0 \leq y \leq n$. Thus,

$$\begin{aligned}
P(Z = z) &= \sum_{x+y=z} P(X = x, Y = y) \\
&= \sum_{x+y=z} \binom{m}{x} p^x (1-p)^{m-x} \binom{n}{y} p^y (1-p)^{n-y} \\
&= \sum_{x+y=z} \binom{m}{x} \binom{n}{y} p^z (1-p)^{m+n-z} \\
&= p^z (1-p)^{m+n-z} \sum_{x+y=z} \binom{m}{x} \binom{n}{y} \\
&= \binom{m+n}{z} p^z (1-p)^{m+n-z}, \text{ using the fact above.}
\end{aligned}$$

Thus, we have proved that the sum of independent binomial random variables with common probability is binomial as well. This is called the reproductive property of random variables.

Now we will state two main results without proof. The proofs will presented in the second-year distribution theory module MATH2011. Suppose that $X_1, \ldots, X_n$ is a random sample from a population distribution with finite variance, and suppose that $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. Define a new random variable

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then:

1. $E(Y) = \mu_1 + \mu_2 + \cdots + \mu_n$.
2. $\text{Var}(Y) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$.

That is:

- The expectation of the sum of independent random variables is the sum of the expectations of the individual random variables
- the variance of the sum of independent random variables is the sum of the variances of the individual random variables.

The second result is *only* true for independent random variables, e.g. random samples. Now we will consider many examples.

Example 3.26 (Mean and variance of binomial distribution)**.** Suppose $Y \sim \text{Bin}(n, p)$. Then we can write:

$$Y = X_1 + X_2 + \ldots + X_n$$

where each X_i is an independent Bernoulli trial with success probability p. We have shown before that, $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$ by direct calculation. Now the above two results imply that:

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = p+p+\ldots+p = np. \text{Var}(Y) = \text{Var}(X_1)+\cdots+\text{Var}(X_n) = p(1-p)+\ldots+p(1-p) = np(1-p).$$

Thus we avoided the complicated sums used to derive $E(X)$ and $\text{Var}(X)$ in Sections 3.4.2.3 and 3.4.2.4.

Example 3.27 (Mean and variance of negative binomial distribution)**.** Recall that the negative binomial random variable Y is the number of trials needed to obtain the rth success in a sequence of independent Bernoulli trials, each with success probability p. Let X_i be the number of trials needed after the $(i-1)$th success to obtain the ith success. Each X_i is a geometric random variable and $Y = X_1 + \cdots + X_r$. Hence

$$E(Y) = E(X_1) + \cdots + E(X_r) = 1/p + \cdots + 1/p = r/p$$

and

$$\text{Var}(Y) = \text{Var}(X_) + \cdots + \text{Var}(X_r) = (1-p)/p^2 + \cdots + (1-p)/p^2 = r(1-p)/p^2.$$

Example 3.28 (Sum of independent normal random variables)**.** Suppose that $X_i \sim N\left(\mu_i, \sigma_i^2\right), i = 1, 2, \ldots, k$ are independent random variables. Suppose that

$$Y = a_1 X_1 + \cdots + a_k X_k.$$

Then we can prove that:

$$Y \sim N\left(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2\right).$$

It is clear that $E(Y) = \sum_{i=1}^{k} \mu_i$ and $\text{Var}(Y) = \sum_{i=1}^{k} \sigma_i^2$. But that Y has the normal distribution cannot yet be proved with the theory we know. This proof will be provided in the second-year distribution theory module MATH2011.

As a consequence of the stated result we can easily see the following. Suppose X_1 and X_2 are independent $N\left(\mu, \sigma^2\right)$ random variables. Then $2X_1 \sim N\left(2\mu, 4\sigma^2\right), X_1 + X_2 \sim N\left(2\mu, 2\sigma^2\right)$, and $X_1 - X_2 \sim N\left(0, 2\sigma^2\right)$. Note that $2X_1$ and $X_1 + X_2$ have different distributions. Suppose that $X_i \sim N\left(\mu, \sigma^2\right), i = 1, \ldots, n$ are independent. Then

$$X_1 + \cdots + X_n \sim N\left(n\mu, n\sigma^2\right),$$

and consequently

$$\bar{X} = \frac{1}{n}\left(X_1 + \cdots + X_n\right) \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

3.8 The Central Limit Theorem

3.8.1 Introduction

The sum (and average) of independent random variables show a remarkable behaviour in practice which is captured by the Central Limit Theorem (CLT). These random variables do not even have to be continuous, all we require is that they are independent and each of them has a finite mean and a finite variance. A version of the CLT follows.

3.8.2 Statement of the Central Limit Theorem (CLT)

Let $X_1, \ldots, X_n$ be independent random variables with finite $E\left(X_i\right) = \mu_i$ and finite $\text{Var}\left(X_i\right) = \sigma_i^2$. Define $Y = \sum_{i=1}^{n} X_i$. Then, for a sufficiently large n, the central limit theorem states that Y is approximately normally distributed with

$$E(Y) = \sum_{i=1}^{n} \mu_i, \quad \text{Var}(Y) = \sum_{i=1}^{n} \sigma_i^2.$$

This also implies that $\bar{X} = \frac{1}{n} Y$ also follows the normal distribution approximately, as the sample size $n \to \infty$. In particular, if $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, i.e. all means are equal and all variances are equal, then the CLT states that, as $n \to \infty$,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Equivalently,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

as $n \to \infty$. The notion of convergence is explained by the convergence of distribution of $\bar{X}$ to that of the normal distribution with the appropriate mean and variance. It means that the cdf of the left hand side, $\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma}$, converges to the cdf of the standard normal random variable, $\Phi(\cdot)$. In other words,

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \leq z\right) = \Phi(z), \quad -\infty < z < \infty.$$

So for "large samples", we can use $N(0,1)$ as an approximation to the sampling distribution of $\sqrt{n}(\bar{X}-\mu)/\sigma$. This result is 'exact', i.e. no approximation is required, if the distribution of the X_i 's are normal in the first place — this was discussed in the previous lecture.

How large does n have to be before this approximation becomes usable? There is no definitive answer to this, as it depends on how "close to normal" the distribution of X is. However, it is often a pretty good approximation for sample sizes as small as 20, or even smaller. It also depends on the skewness of the distribution of X; if the X-variables are highly skewed, then n will usually need to be larger than for corresponding symmetric X-variables for the approximation to be good. We investigate in one example in Figure 3.3.

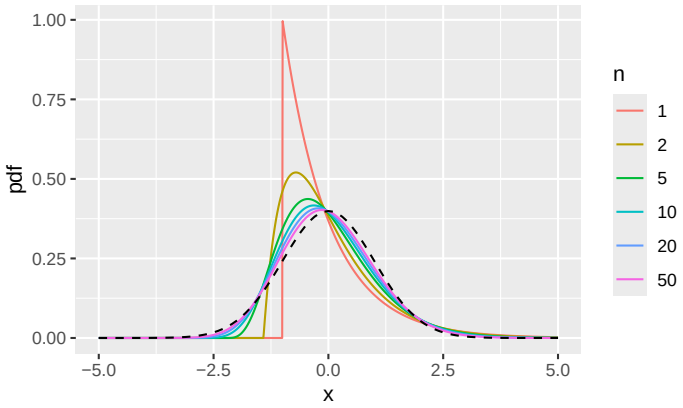


Figure 3.3: Distribution of normalised sample means for samples of different sizes. Initially very skew (original distribution, $n = 1$) becoming rapidly closer to standard normal (dashed line) with increasing n.

3.8.3 Application of CLT to binomial distribution

We know that a binomial random variable Y with parameters n and p is the number of successes in a set of n independent Bernoulli trials, each with success probability p. We may write

$$Y = X_1 + X_2 + \cdots + X_n,$$

where $X_1, \ldots, X_n$ are independent Bernoulli random variables with success probability p. It follows from the CLT that, for a sufficiently large n, Y is approximately normally distributed with expectation $E(Y) = np$ and variance $\text{Var}(Y) = np(1-p)$.

Hence, for given integers y_1 and y_2 between 0 and n and a suitably large n, we have

$$\begin{aligned} P(y_1 \leq Y \leq y_2) &= P\left\{\frac{y_1 - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{y_2 - np}{\sqrt{np(1-p)}}\right\} \\ &\approx P\left\{\frac{y_1 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{y_2 - np}{\sqrt{np(1-p)}}\right\} \end{aligned}$$

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

We should take account of the fact that the binomial random variable Y is integer-valued, and so $P(y_1 \leq Y \leq y_2) = P(y_1 - f_1 \leq Y \leq y_2 + f_2)$ for any two fractions $0 < f_1, f_2 < 1$. This is called continuity correction and we take $f_1 = f_2 = 0.5$ in practice.

$$\begin{aligned} P\left(y_1 \leq Y \leq y_2\right) &= P\left(y_1 - 0.5 \leq Y \leq y_2 + 0.5\right) \\ &= P\left\{\frac{y_1 - 0.5 - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{y_2 + 0.5 - np}{\sqrt{np(1-p)}}\right\} \\ &\approx P\left\{\frac{y_1 - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{y_2 + 0.5 - np}{\sqrt{np(1-p)}}\right\}. \end{aligned}$$

What do we mean by a suitably large n? A commonly-used guideline is that the approximation is adequate if $np \geq 5$ and $n(1-p) \geq 5$.

Example 3.29. A producer of natural yoghurt believed that the market share of their brand was 10%. To investigate this, a survey of 2500 yoghurt consumers was carried out. It was observed that only 205 of the people surveyed expressed a preference for their brand. Should the producer be concerned that they might be losing market share?

Assume that the conjecture about market share is true. Then the number of people Y who prefer this product follows a binomial distribution with $p = 0.1$ and $n = 2500$. So the mean is $np = 250$, the variance is $np(1-p) = 225$, and the standard deviation is 15 . The exact probability of observing $(Y \leq 205)$ is given by the sum of the binomial probabilities up to and including 205, which is difficult to compute. However, this can be approximated by using the CLT:

$$\begin{aligned} P(Y \leq 205) &= P(Y \leq 205.5) \\ &= P\left\{\frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{205.5 - np}{\sqrt{np(1-p)}}\right\} \\ &\approx P\left\{Z \leq \frac{205.5 - np}{\sqrt{np(1-p)}}\right\} \\ &= P\left\{Z \leq \frac{205.5 - 250}{15}\right\} \\ &= \Phi(-2.967) = 0.0015. \end{aligned}$$

This probability is so small that it casts doubt on the validity of the assumption that the market share is 10%.

Although the exact binomial probabilities are difficult to compute by hand, in this case we may compute them in R. Recall $Y \sim \text{Bin}(n = 2500, p = 0.1)$, so $P(Y \leq 205)$ is

```
pbinom(205, size = 2500, prob = 0.1)
```

```
## [1] 0.001173725
```

In this case the normal approximation was good enough to correctly conclude that this probability is very small (of the order of 0.1%), which was all we needed to answer the question of interest here.

Chapter 4

Statistical Inference

4.1 Statistical modelling

4.1.1 Introduction

Statistical analysis (or inference) involves drawing conclusions, and making predictions and decisions, using the evidence provided to us by observed data. To do this we use probability distributions, often called statistical models, to describe the process by which the observed data were generated. For example, we may suppose that the true proportion of mature students is p, $0 < p < 1$, and if we have selected n students at random, that each of those students gives rise to a Bernoulli distribution which takes the value 1 if the student is a mature student and 0 otherwise. The success probability of the Bernoulli distribution will be the unknown p. The underlying statistical model is then the Bernoulli distribution.

To illustrate with another example, suppose we have observed fast food waiting times in the morning and afternoon, as in Example 1.1. If we treat time as continuous then the waiting time for each customer could potentially be modelled as a normal random variable.

In general:

- The form of the assumed model helps us to understand the real-world process by which the data were generated.
- If the model explains the observed data well, then it should also inform us about future (or unobserved) data, and hence help us to make predictions (and decisions contingent on unobserved data).
- The use of statistical models, together with a carefully constructed methodology for their analysis, also allows us to quantify the uncertainty associated with any conclusions, predictions or decisions we make.

We will use the notation $x_1, x_2, \ldots, x_n$ to denote n observations of the random variables $X_1, X_2, \ldots, X_n$ (corresponding capital letters). For the fast food waiting time example, we have $n = 20$, $x_1 = 38, x_2 = 100, \ldots, x_{20} = 70$, and X_i is the waiting time for the ith person in the sample.

4.1.2 Statistical models

Suppose we denote the complete data by the vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and use $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ for the corresponding random variables. A statistical model specifies a probability distribution for the random variables $\boldsymbol{X}$ corresponding to the data observations $\boldsymbol{x}$. Providing a specification for the distribution of n jointly varying random variables can be a daunting task, particularly if n is large. However, this task is made much easier if we can make some simplifying assumptions, such as

1. $X_1, X_2, \ldots, X_n$ are independent random variables,
2. $X_1, X_2, \ldots, X_n$ have the same probability distribution (so $x_1, x_2, \ldots, x_n$ are observations of a single random variable X).

Assumption 1 depends on the sampling mechanism and is very common in practice. If we are to make this assumption for the Southampton student sampling experiment, we need to select randomly among all possible students. The assumption will be violated when samples are correlated either in time or in space, e.g. the daily air pollution level in Southampton for the last year or the air pollution levels in two nearby locations in Southampton. In this module we will only consider data sets where Assumption 1 is reasonable.

Assumption 2 is not always appropriate, but is often reasonable when we are modelling a single variable. In the fast food waiting time example, we must assume that there are no differences between the AM and PM waiting times for Assumption 2 to hold.

If Assumption 1 and 2 both hold, we say that $X_1, \ldots, X_n$ are independent and identically distributed (or i.i.d. for short).

4.1.3 A fully specified model

Sometimes a model completely specifies the probability distribution of $X_1, X_2, \ldots, X_n$. For example, if we assume that the waiting time $X \sim N(\mu, \sigma^2)$ where $\mu = 100$, and $\sigma^2 = 100$, then this is a fully specified model. In this case, there is no need to collect any data as there is no need to make any inference about any unknown quantities, although we may use the data to judge the plausibility of the model. A fully specified model might be appropriate when there is some external (to the data) theory as to why the model (in particular the values of μ and σ^2) was appropriate. Fully specified models such as this are uncommon as we rarely have external theory which allows us to specify a model so precisely.

4.1.4 A parametric statistical model

A parametric statistical model specifies a probability distribution for a random sample apart from the value of a number of parameters in that distribution. This could be confusing in the first instance – a parametric model does not specify parameters! Here the word parametric signifies the fact that the probability distribution is completely specified by a few parameters in the first place. For example, the Poisson distribution is parameterised by the parameter λ which happens to be the mean of the distribution; the normal distribution is parameterised by two parameters, the mean μ and the variance σ^2. When a parametric statistical model is assumed with some unknown parameters, statistical inference methods use data to estimate the unknown parameters, e.g. λ, μ, σ^2. Estimation will be discussed in more detail in the following sections.

4.1.5 A nonparametric statistical model

Sometimes it is not appropriate, or we want to avoid, making a precise specification for the distribution which generated $X_1, X_2, \ldots, X_n$. For example, when the data histogram does not show a bell-shaped distribution, it would be wrong to assume a normal distribution for the data. In such a case, although we can attempt to use some other non-bell-shaped parametric model, we can decide altogether to abandon parametric models. We may then still assume that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, but from a nonparametric statistical model which cannot be written down, having a probability function which only depends on a finite number of parameters. Such analysis approaches are also called distribution-free methods.

Example 4.1 (Computer failures)**.** Let X denote the count of computer failures per week from Example 1.2, summarised in the following table:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	17
12	16	21	12	11	8	7	2	4	2	3	2	2	1	1

We want to estimate how often will the computer system fail at least once per week in the next year? The answer is $52 \times (1 - P(X = 0))$. But how would you estimate $P(X = 0)$? Consider two approaches.

1. **Nonparametric**. Estimate $P(X = 0)$ by the relative frequency of number of zeros in the above sample, which is 12 out of 104. Thus our estimate of $P(X = 0)$ is 12/104. Hence, our estimate of the number of weeks when there will be at least one computer failure is $52 \times (1 - 12/104) = 46$.
2. **Parametric**. Suppose we assume that X follows the Poisson distribution with parameter λ. Then the answer to the above question is

$$52 \times (1 - P(X = 0)) = 52 \times \left(1 - e^{-\lambda}\frac{\lambda^0}{0!}\right) = 52 \times (1 - e^{-\lambda})$$

which involves the unknown parameter λ. For the Poisson distribution we know that $E(X) = \lambda$. Hence we could use the sample mean $\bar{X}$ to estimate $E(X) = \lambda$. Thus we estimate $\hat{\lambda} = \bar{x} = 3.75$. This type of estimator is called a moment estimator. Now our estimate of the number of weeks when there will be at least one computer failure is $52 \times (1 - e^{-3.75}) = 50.78 \approx 51$, which is very different from our answer of 46 from the nonparametric approach.

4.1.6 Should we prefer parametric or nonparametric and why?

The parametric approach should be preferred if the assumption of the Poisson distribution can be justified for the data. For example, we can look at the data histogram or compare the fitted probabilities of different values of X, i.e.

$$\hat{P}(X = x) = e^{-\hat{\lambda}}\frac{\hat{\lambda}}{x!},$$

with the relative frequencies from the sample. In general, often model-based analysis is preferred because it is more precise and accurate, and we can find estimates of uncertainty in such analysis based on the structure of the model. We shall see this later.

The nonparametric approach should be preferred if the model cannot be justified for the data, as in that case the parametric approach will provide incorrect answers.

4.2 Estimation

4.2.1 Introduction

Once we have collected data and proposed a statistical model for our data, the initial statistical analysis usually involves estimation.

- For a parametric model, we need to estimate the unknown (unspecified) parameter λ. For example, if our model for the computer failure data is that they are i.i.d. Poisson, we need to estimate the mean (λ) of the Poisson distribution.
- For a nonparametric model, we may want to estimate the properties of the data-generating distribution. For example, if our model for the computer failure data is that they are i.i.d., following the distribution of an unspecified common random variable X, then we may want to estimate $\mu = E(X)$ or $\sigma^2 = \mathrm{Var}(X)$.

In the following, we use the generic notation θ to denote the *estimand* (what we want to estimate or the parameter). For example, θ is the parameter λ in the first example, and θ may be either μ or σ^2 or both in the second example.

4.2.2 Population and sample

Recall that a statistical model specifies a probability distribution for the random variables $\boldsymbol{X}$ corresponding to the data observations $\boldsymbol{x}$.

- The observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ are called the sample, and quantities derived from the sample are sample quantities. For example, as in Chapter 1, we call

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

  the sample mean.
- The probability distribution for X specified in our model represents all possible observations which might have been observed in our sample, and is therefore sometimes referred to as the population. Quantities derived from this distribution are population quantities. For example, if our model is that $X_1, \ldots, X_n$ are i.i.d., following the common distribution of a random variable X, then we call $E(X)$ the *population mean.*

4.2.3 Statistic and estimator

A statistic $T(\boldsymbol{x})$ is any function of the observed data $x_1, \ldots, x_n$ alone (and therefore does not depend on any parameters or other unknowns).

An estimate of θ is any statistic which is used to estimate θ under a particular statistical model. We will use $\tilde{\theta}(\boldsymbol{x})$ (sometimes shortened to $\tilde{\theta}$) to denote an estimate of θ.

An estimate $\tilde{\theta}(\boldsymbol{x})$ is an observation of a corresponding random variable $\tilde{\theta}(\boldsymbol{X})$ which is called an estimator. Thus an estimate is a particular observed value, e.g. 1.2, but an estimator is a random variable which can take values which are called estimates.

An estimate is a particular numerical value, e.g. $\bar{x}$; an estimator is a random variable, e.g. $\bar{X}$.

The probability distribution of any estimator $\tilde{\theta}(\boldsymbol{X})$ is called its sampling distribution. The estimate $\tilde{\theta}(\boldsymbol{x})$ is an observed value (a number), and is a single observation from the sampling distribution of $\tilde{\theta}(\boldsymbol{X})$.

Example 4.2. Suppose that we have a random sample $X_1, \ldots, X_n$ from the uniform distribution on the interval $(0, \theta)$ where $\theta > 0$ is unknown. Suppose that $n = 5$ and we have the sample observations $x_1 = 2.3, x_2 = 3.6, x_3 = 20.2, x_4 = 0.9, x_5 = 17.2$. Our objective is to estimate θ. How can we proceed?

Here the pdf $f(x) = \frac{1}{\theta}$ for $0 \leq x \leq \theta$ and 0 otherwise. Hence $E(X) = \int_0^\theta \frac{1}{\theta} x dx = \frac{\theta}{2}$. There are many possible estimators for θ, e.g. $\hat{\theta}_1(\boldsymbol{X}) = 2\bar{X}$, which is motivated by the method of moments because $\theta = 2E(X)$. A second estimator is $\hat{\theta}_2(\boldsymbol{X}) = \max\{X_1, X_2, \ldots, X_n\}$, which is intuitive since θ must be greater than or equal to all observed values and thus the maximum of the sample value will be closest to θ.

How could we choose between the two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$? This is where we need to learn the sampling distribution of an estimator to determine which estimator will be unbiased, i.e. correct on average, and which will have minimum variability. We will formally define these in a minute, but first let us derive the sampling distribution, i.e. the pdf, of $\hat{\theta}_2$. Note that $\hat{\theta}_2$ is a random variable since the sample $X_1, \ldots, X_n$ is random. We will first find its cdf and then differentiate the cdf to get the pdf. For ease of notation, suppose $Y = \hat{\theta}_2(\boldsymbol{X}) = \max\{X_1, X_2, \ldots, X_n\}$. For any $0 < y < \theta$, the cdf of

$Y, F(y)$ is given by

$$\begin{aligned} P(Y \leq y) &= P\left(\max\left\{X_1, X_2, \ldots, X_n\right\} \leq y\right) \\ &= P\left(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y)\right) \quad \text{since max} \leq y \text{ if and only if each} \leq y] \\ &= P\left(X_1 \leq y\right) P\left(X_2 \leq y\right) \cdots P\left(X_n \leq y\right) \quad \text{since the } X_i \text{ are independent} \\ &= \frac{y}{\theta} \times \frac{y}{\theta} \times \cdots \times \frac{y}{\theta} \\ &= \left(\frac{y}{\theta}\right)^n. \end{aligned}$$

Now the pdf of Y is

$$f(y) = \frac{dF(y)}{dy} = n\frac{y^{n-1}}{\theta^n}, \quad 0 \leq y \leq \theta.$$

Using this pdf, it can be shown that $E\left(\hat{\theta}_2\right) = E(Y) = \frac{n}{n+1}\theta$, and

$$\text{Var}\left(\hat{\theta}_2\right) = \frac{n\theta^2}{(n+2)(n+1)^2}.$$

4.2.4 Bias and mean square error

In the uniform distribution example we saw that the estimator $\hat{\theta}_2 = Y = \max\left\{X_1, X_2, \ldots, X_n\right\}$ is a random variable and its pdf is given by $f(y) = n\frac{y^{n-1}}{\theta^n}$ for $0 \leq y \leq \theta$. This probability distribution is called the sampling distribution of $\hat{\theta}_2$. From this we have seen that $E\left(\hat{\theta}_2\right) = \frac{n}{n+1}\theta$.

In general, we define the bias of an estimator $\tilde{\theta}(\boldsymbol{X})$ of θ to be

$$\text{bias}(\tilde{\theta}) = E(\tilde{\theta}) - \theta.$$

An estimator $\tilde{\theta}(\boldsymbol{X})$ is said to be unbiased if

$$\text{bias}(\tilde{\theta}) = 0, \text{ i.e. if } E(\tilde{\theta}) = \theta.$$

So an estimator is unbiased if the expectation of its sampling distribution is equal to the quantity we are trying to estimate. Unbiased means "getting it right on average", i.e. under repeated sampling (relative frequency interpretation of probability).

Thus for the uniform distribution example, $\hat{\theta}_2$ is a biased estimator of θ and

$$\text{bias}\left(\hat{\theta}_2\right) = E\left(\hat{\theta}_2\right) - \theta = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta,$$

which goes to zero as $n \to \infty$. However, $\hat{\theta}_1 = 2\bar{X}$ is unbiased since $E\left(\hat{\theta}_1\right) = 2E(\bar{X}) = 2\frac{\theta}{2} = \theta$.

Unbiased estimators are "correct on average", but that does not mean that they are guaranteed to provide estimates which are close to the estimand θ. A better measure of the quality of an estimator than bias is the mean squared error (or MSE), defined as

$$\text{MSE}(\tilde{\theta}) = E\left[(\tilde{\theta} - \theta)^2\right]$$

Therefore, if $\tilde{\theta}$ is unbiased for θ, i.e. if $E(\tilde{\theta}) = \theta$, then MSE $(\tilde{\theta}) = \text{Var}(\tilde{\theta})$. In general, we have the following result:

Theorem 4.1.

$$\mathrm{MSE}(\tilde{\theta}) = \mathrm{Var}(\tilde{\theta}) + \mathrm{bias}(\tilde{\theta})^2$$

The proof is similar to the proof of Theorem 1.1.

Proof.

$$\begin{aligned}
\mathrm{MSE}(\tilde{\theta}) &= E\left[(\tilde{\theta}-\theta)^2\right] \\
&= E\left[(\tilde{\theta}-E(\tilde{\theta})+E(\tilde{\theta})-\theta)^2\right] \\
&= E\left[(\tilde{\theta}-E(\tilde{\theta}))^2+(E(\tilde{\theta})-\theta)^2+2(\tilde{\theta}-E(\tilde{\theta}))(E(\tilde{\theta})-\theta)\right] \\
&= E[\tilde{\theta}-E(\tilde{\theta})]^2+E[E(\tilde{\theta})-\theta]^2+2E[(\tilde{\theta}-E(\tilde{\theta}))(E(\tilde{\theta})-\theta)] \\
&= \mathrm{Var}(\tilde{\theta})+[E(\tilde{\theta})-\theta]^2+2(E(\tilde{\theta})-\theta)E[(\tilde{\theta}-E(\tilde{\theta}))] \\
&= \mathrm{Var}(\tilde{\theta})+\mathrm{bias}(\tilde{\theta})^2+2(E(\tilde{\theta})-\theta)[E(\tilde{\theta})-E(\tilde{\theta})] \\
&= \mathrm{Var}(\tilde{\theta})+\mathrm{bias}(\tilde{\theta})^2.
\end{aligned}$$

□

Example 4.3. Continuing with the uniform distribution $U(0,\theta)$ example, we have seen that $\hat{\theta}_1 = 2\bar{X}$ is unbiased for θ but bias $\left(\hat{\theta}_2\right) = -\frac{1}{n+1}\theta$. How do these estimators compare with respect to the MSE? Since $\hat{\theta}_1$ is unbiased, its MSE is its variance. Later, we will prove that for random sampling from any population

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n},$$

where $\mathrm{Var}(X)$ is the variance of the population sampled from. Returning to our example, we know that if $X \sim U(0,\theta)$ then $\mathrm{Var}(X) = \frac{\theta^2}{12}$. Therefore we have

$$\mathrm{MSE}\left(\hat{\theta}_1\right) = \mathrm{Var}\left(\hat{\theta}_1\right) = \mathrm{Var}(2\bar{X}) = 4\,\mathrm{Var}(\bar{X}) = 4\frac{\theta^2}{12n} = \frac{\theta^2}{3n}.$$

Now, for $\hat{\theta}_2$ we know that:

1. $\mathrm{Var}\left(\hat{\theta}_2\right) = \frac{n\theta^2}{(n+2)(n+1)^2}$
2. $\mathrm{bias}\left(\hat{\theta}_2\right) = -\frac{1}{n+1}\theta$.

Now

$$\begin{aligned}
\mathrm{MSE}\left(\hat{\theta}_2\right) &= \mathrm{Var}\left(\hat{\theta}_2\right) + \mathrm{bias}\left(\hat{\theta}_2\right)^2 \\
&= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{\theta^2}{(n+1)^2}\left(\frac{n}{n+2}+1\right) \\
&= \frac{\theta^2}{(n+1)^2}\frac{2n+2}{n+2}.
\end{aligned}$$

The MSE of $\hat{\theta}_2$ is an order of magnitude smaller than the MSE of $\hat{\theta}_1$ (of order $1/n^2$ rather than $1/n$), providing justification for the preference of $\hat{\theta}_2 = \max\{X_1, X_2, \ldots, X_n\}$ as an estimator of θ.

4.3 Estimating the population mean

4.3.1 Introduction

Often, one of the main tasks of a statistician is to estimate a population average or mean. However the estimates, using whatever procedure, will not be usable or scientifically meaningful if we do not know their associated uncertainties. For example, a statement such as: "the Arctic ocean will be completely ice-free in the summer in the next few decades" provides little information as it does not communicate the extent or the nature of the uncertainty in it. Perhaps a more precise statement could be: "the Arctic ocean will be completely ice-free in the summer some time in the next 20-30 years". This last statement not only gives a numerical value for the number of years for complete ice-melt in the summer, but also acknowledges the uncertainty of ± 5 years in the estimate. A statistician's main job is to estimate such uncertainties. In this lecture, we will get started with estimating uncertainties when we estimate a population mean. We will introduce the standard error of an estimator.

4.3.2 Estimation of a population mean

Suppose that $X_1, \ldots, X_n$ is a random sample from any probability distribution $f(x)$, which may be discrete or continuous. Suppose that we want to estimate the unknown population mean $E(X) = \mu$ and variance, $\mathrm{Var}(X) = \sigma^2$. In order to do this, it is not necessary to make any assumptions about $f(x)$, so this may be thought of as nonparametric inference.

We have the following results:

Theorem 4.2. *Suppose $X_1, \ldots, X_n$ is a random sample, with $E(X) = \mu$. The sample mean*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is an unbiased estimator of $\mu = E(X)$, i.e. $E(\bar{X}) = \mu$.

In other words, the sample mean is an unbiased estimator of the population mean.

Proof. We have

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X) = E(X)$$

so $\bar{X}$ is an unbiased estimator of $E(X)$. □

Theorem 4.3. *Suppose $X_1, \ldots, X_n$ is a random sample, with $\mathrm{Var}(X) = \sigma^2$. Then*

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Proof. We use the result that for independent random variables the variance of the sum is the sum of the variances from Section 3.7 . Thus,

$$\mathrm{Var}[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X) = \frac{n}{n^2} \mathrm{Var}(X) = \frac{\sigma^2}{n},$$

□

Together, Theorems 4.2 and 4.3 imply that the MSE of $\bar{X}$ is $\mathrm{Var}(X)/n$.

Theorem 4.4. *Suppose $X_1, \ldots, X_n$ is a random sample, with $\mathrm{Var}(X) = \sigma^2$. The sample variance with divisor $n-1$*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

is an unbiased estimator of σ^2, i.e. $E\left(S^2\right) = \sigma^2$.

In other words, the sample variance is an unbiased estimator of the population variance.

Proof. We need to show $E\left(S^2\right) = \sigma^2$. We have

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right].$$

To evaluate the expectation of the above, we need $E\left(X_i^2\right)$ and $E\left(\bar{X}^2\right)$. In general, we know for any random variable,

$$\mathrm{Var}(Y) = E\left(Y^2\right) - (E(Y))^2 \quad \Rightarrow E\left(Y^2\right) = \mathrm{Var}(Y) + (E(Y))^2.$$

Thus, we have

$$E\left(X_i^2\right) = \mathrm{Var}\left(X_i\right) + \left(E\left(X_i\right)\right)^2 = \sigma^2 + \mu^2,$$

and

$$E\left(\bar{X}^2\right) = \mathrm{Var}(\bar{X}) + (E(\bar{X}))^2 = \sigma^2/n + \mu^2,$$

from Theorems 4.2 and 4.3. So

$$\begin{aligned}
E\left(S^2\right) &= E\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right\} \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left(X_i^2\right) - nE\left(\bar{X}^2\right)\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - n\left(\sigma^2/n + \mu^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right] \\
&= \sigma^2 \equiv \mathrm{Var}(X).
\end{aligned}$$

□

4.3.3 Standard deviation and standard error

For an unbiased estimator $\tilde{\theta}$,

$$\mathrm{MSE}(\tilde{\theta}) = \mathrm{Var}(\tilde{\theta})$$

and therefore the sampling variance of the estimator is an important summary of its quality.

We usually prefer to focus on the standard deviation of the sampling distribution of $\tilde{\theta}$,

$$\text{s.d.}(\tilde{\theta}) = \sqrt{\mathrm{Var}(\tilde{\theta})}.$$

In practice we will not know s.d.$(\tilde{\theta})$, as it will typically depend on unknown features of the distribution of $X_1, \ldots, X_n$. However, we may be able to estimate s.d.$(\tilde{\theta})$ using the observed sample $x_1, \ldots, x_n$. We define the standard error, s.e.$(\tilde{\theta})$, of an estimator $\tilde{\theta}$ to be an estimate of the standard deviation of its sampling distribution, s.d.$(\tilde{\theta})$.

Standard error of an estimator is an estimate of the standard deviation of its sampling distribution

We proved that

$$\text{Var}[\bar{X}] = \frac{\sigma^2}{n} \Rightarrow \text{ s.d. } (\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

As σ is unknown, we cannot calculate this standard deviation. However, we know that $E\left(S^2\right) = \sigma^2$, i.e. that the sample variance is an unbiased estimator of the population variance. Hence S^2/n is an unbiased estimator for $\text{Var}(\bar{X})$. Therefore we obtain the standard error of the mean, s.e. $(\bar{X})$, by plugging in the estimate

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

of σ into s.d.$(\bar{X})$ to obtain

$$\text{s.e.}(\bar{X}) = \frac{s}{\sqrt{n}}.$$

Therefore, for the computer failure data, our estimate, $\bar{x} = 3.75$, for the population mean is associated with a standard error

$$\text{s.e.}(\bar{X}) = \frac{3.381}{\sqrt{104}} = 0.332.$$

Note that this is "a" standard error, so other standard errors may be available. Indeed, for parametric inference, where we make assumptions about $f(x)$, alternative standard errors are available. For example, if $X_1, \ldots, X_n$ are i.i.d. Poisson(λ) random variables, $E(X) = \lambda$, so $\bar{X}$ is an unbiased estimator of λ. $\text{Var}(X) = \lambda$, so another s.e.$(\bar{X}) = \sqrt{\hat{\lambda}/n} = \sqrt{\bar{x}/n}$. In the computer failure data example, this is $\sqrt{\frac{3.75}{104}} = 0.19$.

4.4 Confidence intervals

4.4.1 Introduction

An estimate $\tilde{\theta}$ of a parameter θ is sometimes referred to as a point estimate. The usefulness of a point estimate is enhanced if some kind of measure of its precision can also be provided. Usually, for an unbiased estimator, this will be a standard error, an estimate of the standard deviation of the associated estimator, as we have discussed previously. An alternative summary of the information provided by the observed data about the location of a parameter θ and the associated precision is a confidence interval.

Suppose that $x_1, \ldots, x_n$ are observations of random variables $X_1, \ldots, X_n$ whose joint pdf is specified apart from a single parameter θ. To construct a confidence interval for θ, we need to find a random variable $T(\mathbf{X}, \theta)$ whose distribution does not depend on θ and is therefore known. This random variable $T(\mathbf{X}, \theta)$ is called a *pivot* for θ. Hence we can find numbers h_1 and h_2 such that

$$P\left(h_1 \leq T(\mathbf{X}, \theta) \leq h_2\right) = 1 - \alpha \tag{4.1}$$

where $1 - \alpha$ is any specified probability. If (4.1) can be 'inverted',we can write it as

$$P\left(g_1(\mathbf{X}) \leq \theta \leq g_2(\mathbf{X})\right) = 1 - \alpha.$$

Hence with probability $1 - \alpha$, the parameter θ will lie between the random variables $g_1(\mathbf{X})$ and $g_2(\mathbf{X})$. Alternatively, the random interval $(g_1(\mathbf{X}), g_2(\mathbf{X}))$ includes θ with probability $1 - \alpha$. Now, when we observe $x_1, \ldots, x_n$, we observe a single observation of the random interval $(g_1(\mathbf{X}), g_2(\mathbf{X}))$, which can

be evaluated as $(g_1(\mathbf{x}), g_2(\mathbf{x}))$. We do not know if θ lies inside or outside this interval, but we do know that if we observed repeated samples, then $100(1-\alpha)\%$ of the resulting intervals would contain θ. Hence, if $1-\alpha$ is high, we can be reasonably confident that our observed interval contains θ. We call the observed interval $(g_1(\mathbf{x}), g_2(\mathbf{x}))$ a $100(1-\alpha)\%$ confidence interval for θ. It is common to present intervals with high confidence levels, usually $90\%, 95\%$ or 99%, so that $\alpha = 0.1, 0.05$ or 0.01 respectively.

4.4.2 Confidence interval for a normal mean

Let $X_1, \ldots, X_n$ be i.i.d. $N\left(\mu, \sigma^2\right)$ random variables. From the Central Limit Theorem (Section 3.8), we know that for large n,

$$\bar{X} \sim N\left(\mu, \sigma^2/n\right) \quad \Rightarrow \quad \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \sim N(0,1).$$

Suppose we know that $\sigma = 10$, so $\sqrt{n}(\bar{X}-\mu)/\sigma$ is a pivot for μ. Then we can use the distribution function of the standard normal distribution to find values h_1 and h_2 such that

$$P\left(h_1 \leq \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \leq h_2\right) = 1-\alpha$$

for a chosen value of $1-\alpha$ which is called the confidence level. So h_1 and h_2 are chosen so that the shaded area in Figure 4.1 is equal to the confidence level $1-\alpha$.

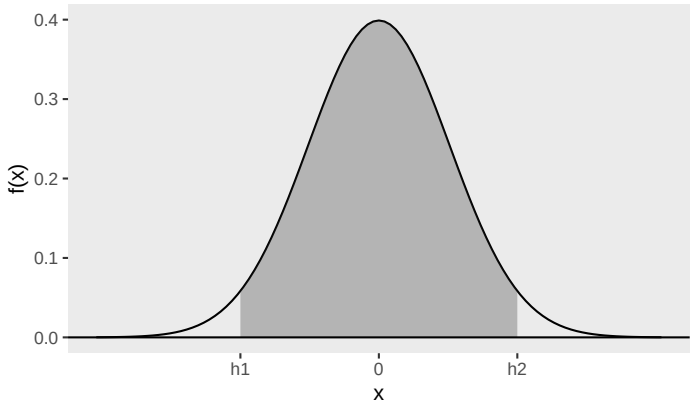


Figure 4.1: h_1 and h_2 are chosen to make the shaded area equal to the confidence level $1-\alpha$

It is common practice to make the interval symmetric, so that the two unshaded areas are equal (to $\alpha/2$), in which case

$$-h_1 = h_2 \equiv h \quad \text{and} \quad \Phi(h) = 1 - \frac{\alpha}{2}.$$

The most common choice of confidence level is $1-\alpha = 0.95$, in which case $h = 1.96 =$ `qnorm(0.975)`. You may also occasionally see 90% ($h = 1.645 =$ `qnorm(0.95)`) or 99% ($h = 2.58 =$ `qnorm(0.995)`) intervals.

Therefore we have

$$P\left(-1.96 \leq \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \leq 1.96\right) = 0.95$$
$$\Rightarrow P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Hence, $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$ and $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$ are the endpoints of a random interval which includes μ with probability 0.95. The observed value of this interval, $\left(\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}\right)$, is called a 95% confidence interval for μ.

Example 4.4 (Fast food waiting times)**.** For the fast food waiting time data, we have $n = 20$ data points combined from the morning and afternoon data sets. We have $\bar{x} = 67.85$ and $n = 20$. Hence, under the normal model assuming (just for the sake of illustration) $\sigma = 18$, a 95% confidence interval for μ is

$$67.85 - 1.96(18/\sqrt{20}) \leq \mu \leq 67.85 + 1.96(18/\sqrt{20})$$
$$\Rightarrow 59.96 \leq \mu \leq 75.74$$

The R command is

```
mean(fastfood) + c(-1, 1) * qnorm(0.975) * 18 / sqrt (20)
```

assuming `fastfood` is the vector containing 20 waiting times.

In reality, it is more likely that σ is unknown in this example, and we need to seek alternative methods for finding the confidence intervals.

4.4.3 Some remarks about confidence intervals

1. Notice that $\bar{x}$ is an unbiased estimate of $\mu, \sigma/\sqrt{n}$ is the standard error of the estimate and 1.96 (in general h in the above discussion) is a critical value from the associated known sampling distribution. The formula $(\bar{x} \pm 1.96\sigma/\sqrt{n})$ for the confidence interval is then generalised as:

   $$\text{Estimate} \ \pm \ \text{Critical value} \ \times \ \text{Standard error}$$

   where the estimate is $\bar{x}$, the critical value is 1.96 and the standard error is $\sigma/\sqrt{n}$. This is so much easier to remember. We will see that this formula holds in many of the following examples, but not all.

2. Confidence intervals are frequently used, but also frequently misinterpreted. A $100(1-\alpha)\%$ confidence interval for θ is a single observation of a random interval which, under repeated sampling, would include θ $100(1-\alpha)\%$ of the time.

3. A confidence interval is not a probability interval. You should avoid making statements like $P(1.3 < \theta < 2.2) = 0.95$. In the classical approach to statistics you can only make probability statements about random variables, and θ is assumed to be a constant.

4. If a confidence interval is interpreted as a probability interval, this may lead to problems. For example, suppose that X_1 and X_2 are i.i.d. $U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ random variables. Then $P\left(\min\left(X_1, X_2\right) < \theta < \max\left(X_1, X_2\right)\right) = \frac{1}{2}$ so $\left(\min\left(x_1, x_2\right), \max\left(x_1, x_2\right)\right)$ is a 50% confidence interval for θ, where x_1 and x_2 are the observed values of X_1 and X_2. Now suppose that $x_1 = 0.3$ and $x_2 = 0.9$. What is $P(0.3 < \theta < 0.9)$?

4.4.4 Confidence intervals using the CLT

4.4.4.1 Introduction

Confidence intervals are generally difficult to find. The difficulty lies in finding a pivot, i.e. a statistic $T(\mathbf{X}, \theta)$ such that

$$P\left(h_1 \leq T(\mathbf{X}, \theta) \leq h_2\right) = 1 - \alpha$$

for two suitable numbers h_1 and h_2, and also that the above can be inverted to put the unknown θ in the middle of the inequality inside the probability statement. One solution to this problem is to use

the powerful Central Limit Theorem (CLT) to claim normality, and then basically follow the above normal example for known variance.

4.4.4.2 Confidence intervals for μ using the CLT

The CLT allows us to assume the large sample approximation

$$\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \overset{\text{approx}}{\sim} N(0,1) \text{ as } n \to \infty.$$

Thus an (approximate) 95% confidence interval (CI) for μ is given by $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. But note that σ is unknown so this CI cannot be used unless we can estimate σ, i.e. replace the unknown s.d. of $\bar{X}$ by its estimated standard error. In this case, we get the CI in the familiar form:

$$\text{Estimate } \pm \text{ Critical value } \times \text{ Standard error}$$

Suppose that we do not assume any distribution for the sampled random variable X but assume only that $X_1, \ldots, X_n$ are i.i.d, following the distribution of X where $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. We know that the standard error of $\bar{X}$ is $s/\sqrt{n}$ where s is the sample standard deviation with divisor $n-1$. Then the following provides an (approximate) 95% CI for μ:

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}.$$

Example 4.5 (Computer failures)**.** For the computer failure data, $\bar{x} = 3.75, s = 3.381$ and $n = 104$. Under the model that the data are observations of i.i.d. random variables with population mean μ (but no other assumptions about the underlying distribution), we compute a 95% confidence interval for μ to be

$$\left(3.75 - 1.96\frac{3.381}{\sqrt{104}}, 3.75 + 1.96\frac{3.381}{\sqrt{104}}\right) = (3.10, 4.40).$$

If we can assume a distribution for X, i.e. a parametric model for X, then we can do slightly better in estimating the standard error of $\bar{X}$ and as a result we can improve upon the previously obtained 95% CI. Two examples follow.

Example 4.6 (Poisson)**.** If $X_1, \ldots, X_n$ are modelled as i.i.d. Poisson (λ) random variables, then $\mu = \lambda$ and $\sigma^2 = \lambda$. We know $\text{Var}(\bar{X}) = \sigma^2/n = \lambda/n$. Hence a standard error is $\sqrt{\hat{\lambda}/n} = \sqrt{\bar{x}/n}$ since $\hat{\lambda} = \bar{X}$ is an unbiased estimator of λ. Thus a 95% CI for $\mu = \lambda$ is given by

$$\bar{x} \pm 1.96\sqrt{\frac{\bar{x}}{n}}.$$

For the computer failure data, $\bar{x} = 3.75, s = 3.381$ and $n = 104$. Under the model that the data are observations of i.i.d. random variables following a Poisson distribution with population mean λ, we compute a 95% confidence interval for λ as

$$\bar{x} \pm 1.96\sqrt{\frac{\bar{x}}{n}} = 3.75 \pm 1.96\sqrt{3.75/104} = (3.38, 4.12).$$

We see that this interval is narrower ($0.74 = 4.12 - 3.38$) than the earlier interval $(3.10, 4.40)$, which has a length of 1.3. We prefer narrower confidence intervals as they facilitate more accurate inference regarding the unknown parameter.

Example 4.7 (Bernoulli)**.** If $X_1, \ldots, X_n$ are modelled as i.i.d. Bernoulli (p) random variables, then $\mu = p$ and $\sigma^2 = p(1-p)$. We know $\text{Var}(\bar{X}) = \sigma^2/n = p(1-p)/n$. Hence a standard error is $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{\bar{x}(1-\bar{x})/n}$, since $\hat{p} = \bar{X}$ is an unbiased estimator of p. Thus a 95% CI for $\mu = p$ is given by

$$\bar{x} \pm 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}.$$

For the example, suppose $\bar{x} = 0.2$ and $n = 10$. Then we obtain the 95% CI as

$$0.2 \pm 1.96\sqrt{(0.2 \times 0.8)/10} = (-0.048, 0.448)$$

Here n is too small for the large sample approximation to be accurate.

4.4.5 Exact confidence interval for the normal mean

For normal models we do not have to rely on large sample approximations, because it turns out that the distribution of

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S},$$

where S^2 is the sample variance with divisor $n-1$, is standard (easily calculated) and thus the statistic $T = T(\mathbf{X}, \mu)$ can be an exact pivot for any sample size $n > 1$.

The point about easy calculation is that for any given $1-\alpha$, e.g. $1-\alpha = 0.95$, we can calculate the critical value h such that $P(-h < T < h) = 1-\alpha$. Note also that the pivot T does not involve the other unknown parameter of the normal model, namely the variance σ^2. If indeed, we can find h for any given $1-\alpha$, then proceed as follows to find the exact CI for μ :

$$\begin{aligned} & P(-h \le T \le h) = 1-\alpha \\ \text{i.e. } & P\left(-h \le \sqrt{n}\frac{(\bar{X}-\mu)}{S} \le h\right) = 0.95 \\ \Rightarrow & P\left(\bar{X} - h\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + h\frac{S}{\sqrt{n}}\right) = 0.95 \end{aligned}$$

The observed value of this interval, $\left(\bar{x} \pm h\frac{s}{\sqrt{n}}\right)$, is the 95% confidence interval for μ. Remarkably, this also of the general form,

$$\text{Estimate } \pm \text{ Critical value } \times \text{ Standard error },$$

where the critical value is h and the standard error of the sample mean is $\frac{s}{\sqrt{n}}$. Now, how do we find the critical value h for a given $1-\alpha$? We need to introduce the t-distribution.

Let $X_1, \ldots, X_n$ be i.i.d $N\left(\mu, \sigma^2\right)$ random variables. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right).$$

Then, it can be shown (and will be in MATH2011) that

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1},$$

where t_{n-1} denotes the standard t distribution with $n-1$ degrees of freedom. The standard t distribution is a family of distributions which depend on one parameter called the degrees-of-freedom

(df) which is $n-1$ here. The concept of degrees of freedom is that it is usually the number of independent random samples, n here, minus the number of linear parameters estimated, 1 here for μ. Hence the df is $n-1$.

The probability density function of the t_k distribution is similar to a standard normal, in that it is symmetric around zero and 'bell-shaped', but the t-distribution is more heavy-tailed, giving greater probability to observations further away from zero. Figure 4.2 illustrates the t_k density function for various values of k.

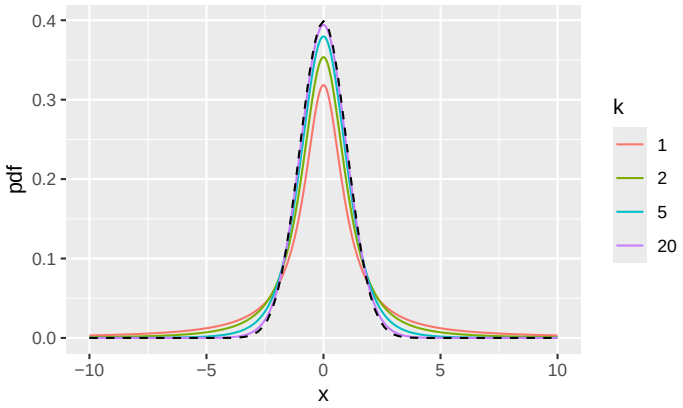


Figure 4.2: The pdf of the t_k distribution for various values of k. The dashed line is the standard normal pdf.

The values of h for a given $1-\alpha$ can be obtained using the R command `qt` (abbreviation for quantile of t). For example, we can find h for $1-\alpha = 0.95$ and $n = 20$

```
qt(0.975, df = 19)
```

```
## [1] 2.093024
```

Note that it should be 0.975 so that we are splitting 0.05 probability between the two tails equally and the df should be $n-1 = 19$. Indeed, modifying the above command, we obtain the following critical values for the 95% interval for different values of the sample size n.

n	2	5	10	15	20	30	50	100	1000
h	12.71	2.78	2.26	2.14	2.09	2.05	2.01	1.98	1.96

Note that the critical value approaches 1.96 (which is the critical value for the normal distribution) as $n \to \infty$, since the t-distribution itself approaches the normal distribution for large values of its df parameter.

If you can justify that the underlying distribution is normal then you can use the t-distribution-based confidence interval.

Example 4.8 (Fast food waiting times)**.** We would like to find a confidence interval for the true mean waiting time. If X denotes the waiting time in seconds, we have $n = 20, \bar{x} = 67.85$, $s = 18.36$. Hence, recalling that the critical value $h = 2.093$, from the command `qt(0.975, df = 19)`, a 95% confidence interval for μ is

$$67.85 - 2.093 \times 18.36/\sqrt{20} \leq \mu \leq 67.85 + 2.093 \times 18.36/\sqrt{20}$$
$$\Rightarrow 59.26 \leq \mu \leq 76.44.$$

In R, if the vector `fastfood` contains all the service times, we can find a 95% confidence interval with

```
mean(fastfood) + c(-1, 1) * qt(0.975, df = 19) * sd(fastfood) / sqrt(20)
```

```
## [1] 59.25467 76.44533
```

or a 90% confidence interval with

```
mean(fastfood) + c(-1, 1) * qt(0.95, df = 19) * sd(fastfood) / sqrt(20)
```

```
## [1] 60.74905 74.95095
```

or a 99% confidence interval with

```
mean(fastfood) + c(-1, 1) * qt(0.995, df = 19) * sd(fastfood) / sqrt(20)
```

```
## [1] 56.10113 79.59887
```

We can see clearly that the interval gets wider as the level of confidence is gets higher.

Example 4.9 (Weight gain)**.** Returning to the weight gain data from Example 1.3, we would like to find a confidence interval for the true average weight gain (final weight - initial weight). Here $n = 68, \bar{x} = 0.8672$ and $s = 0.9653$. Hence, a 95% confidence interval for μ is

$$0.8672 - 1.996 \times 0.9653/\sqrt{68} \leq \mu \leq 0.8672 + 1.996 \times 0.9653/\sqrt{68}$$
$$\Rightarrow 0.6335 \leq \mu \leq 1.1008$$

In R, we obtain the critical value 1.996 by

```
qt(0.975, df = 67)
```

```
## [1] 1.996008
```

Note that the interval here does not include the value 0, so it is very likely that the weight gain is significantly positive, which we will justify using what is called hypothesis testing.

4.5 Hypothesis testing

4.5.1 Hypothesis testing in general

4.5.1.1 Introduction

The manager of a new fast food chain claims that the average waiting time to be served in their restaurant is less than a minute. The marketing department of a mobile phone company claims that their phones never break down in the first three years of their lifetime. A professor of nutrition claims that students gain significant weight in the first year of their life in college away form home. How can we verify these claims? We will learn the procedures of hypothesis testing for such problems.

In statistical inference, we use observations $x_1, \ldots, x_n$ of univariate random variables $X_1, \ldots, X_n$ in order to draw inferences about the probability distribution $f(x)$ of the underlying random variable X. So far, we have mainly been concerned with estimating features (usually unknown parameters) of $f(x)$. It is often of interest to compare alternative specifications for $f(x)$. If we have a set of competing probability models which might have generated the observed data, we may want to determine which of the models is most appropriate. A proposed (hypothesised) model for $X_1, \ldots, X_n$ is then referred to as a hypothesis, and pairs of models are compared using hypothesis tests.

For example, we may have two competing alternatives, $f^{(0)}(x)$ (model H_0) and $f^{(1)}(x)$ (model H_1) for $f(x)$, both of which completely specify the joint distribution of the sample $X_1, \ldots, X_n$. Completely specified statistical models are called simple hypotheses. Usually, H_0 and H_1 both take the same parametric form $f(x, \theta)$, but with different values $\theta^{(0)}$ and $\theta^{(1)}$ of θ. Thus the joint distribution of the

sample given by $f(\mathbf{X})$ is completely specified apart from the values of the unknown parameter θ and $\theta^{(0)} \neq \theta^{(1)}$ are specified alternative values.

More generally, competing hypotheses often do not completely specify the joint distribution of $X_1, \ldots, X_n$. For example, a hypothesis may state that $X_1, \ldots, X_n$ is a random sample from the probability distribution $f(x;\theta)$ where $\theta < 0$. This is not a completely specified hypothesis, since it is not possible to calculate probabilities such as $P(X_1 < 2)$ when the hypothesis is true, as we do not know the exact value of θ. Such an hypothesis is called a composite hypothesis.

Examples of hypotheses:

- $X_1, \ldots, X_n \sim N\left(\mu, \sigma^2\right)$ with $\mu = 0, \sigma^2 = 2$.
- $X_1, \ldots, X_n \sim N\left(\mu, \sigma^2\right)$ with $\mu = 0, \sigma^2 \in \mathcal{R}_+$.
- $X_1, \ldots, X_n \sim N\left(\mu, \sigma^2\right)$ with $\mu \neq 0, \sigma^2 \in \mathcal{R}_+$.
- $X_1, \ldots, X_n \sim$ Bernoulli(p) with $p = \frac{1}{2}$.
- $X_1, \ldots, X_n \sim$ Bernoulli(p) with $p \neq \frac{1}{2}$.
- $X_1, \ldots, X_n \sim$ Bernoulli(p) with $p > \frac{1}{2}$.
- $X_1, \ldots, X_n \sim$ Poisson(λ) with $\lambda = 1$.
- $X_1, \ldots, X_n \sim$ Poisson(θ) with $\theta > 1$

4.5.1.2 Hypothesis testing procedure

A hypothesis test provides a mechanism for comparing two competing statistical models, H_0 and H_1. A hypothesis test does not treat the two hypotheses (models) symmetrically. One hypothesis, H_0, is given special status, and referred to as the null hypothesis. The null hypothesis is the reference model, and is assumed to be appropriate unless the observed data strongly indicate that H_0 is inappropriate, and that H_1 (the alternative hypothesis) should be preferred. Hence, the fact that a hypothesis test does not reject H_0 should not be taken as evidence that H_0 is true and H_1 is not, or that H_0 is better-supported by the data than H_1, merely that the data does not provide significant evidence to reject H_0 in favour of H_1.

A hypothesis test is defined by its critical region or rejection region, which we shall denote by C. C is a subset of $\mathcal{R}^n$ and is the set of possible observed values of $\mathbf{X}$ which, if observed, would lead to rejection of H_0 in favour of H_1, i.e.

$$\begin{array}{ll} \text{If } \mathbf{x} \in C & H_0 \text{ is rejected in favour of } H_1 \\ \text{If } \mathbf{x} \notin C & H_0 \text{ is not rejected} \end{array}$$

As $\mathbf{X}$ is a random variable, there remains the possibility that a hypothesis test will give an erroneous result. We define two types of error:

Type I error: H_0 is rejected when it is true
Type II error: H_0 is not rejected when it is false

The following table helps to understand further:

	H_0 true	H_0 false
Reject H_0	Type I error	Correct decision
Do not reject H_0	Correct decision	Type II error

When H_0 and H_1 are simple hypotheses, we can define

$$\begin{aligned} \alpha &= P(\text{Type I error}) = P(\mathbf{X} \in C) \quad \text{if } H_0 \text{ is true} \\ \beta &= P(\text{Type II error}) = P(\mathbf{X} \notin C) \quad \text{if } H_1 \text{ is true} \end{aligned}$$

Example 4.10 (Uniform)**.** Suppose that we have one observation from the uniform distribution on the range $(0, \theta)$. In this case, $f(x) = 1/\theta$ if $0 < x < \theta$ and $P(X \leq x) = \frac{x}{\theta}$ for$0 < x < \theta$. We want to

test $H_0 : \theta = 1$ against the alternative $H_1 : \theta = 2$. Suppose we decide arbitrarily that we will reject H_0 if $X > 0.75$. Then

$$\alpha = P(\text{Type I error}) = P(X > 0.75) \text{ if } H_0 \text{ is true}$$
$$\beta = P(\text{Type II error}) = P(X < 0.75) \text{ if } H_1 \text{ is true}$$

which will imply:

$$\alpha = P(X > 0.75 \mid \theta = 1) = 1 - 0.75 = 0.25,$$
$$\beta = P(X < 0.75 \mid \theta = 2) = 0.75/2 = 0.375.$$

Example 4.11 (Poisson)**.** The daily demand for a product has a Poisson distribution with mean λ, the demands on different days being statistically independent. It is desired to test the hypotheses $H_0 : \lambda = 0.7$ against the alternative $H_1 : \lambda = 0.3$. The daily demand will be measured for 20 days, and the null hypothesis will be rejected if there is no demand on at least 15 of these 20 days. Calculate the Type I and Type II error probabilities.

Let p denote the probability that the demand on a given day is zero. Then

$$p = e^{-\lambda} = \begin{cases} e^{-0.7} & \text{under } H_0 \\ e^{-0.3} & \text{under } H_1 \end{cases}$$

If X denotes the number of days out of 20 with zero demand, it follows that

$$X \sim \text{Binomial}\left(20, e^{-0.7}\right) \text{ under } H_0,$$
$$X \sim \text{Binomial}\left(20, e^{-0.3}\right) \text{ under } H_1.$$

Thus

$$\begin{aligned} \alpha &= P\left(\text{Reject } H_0 \mid H_0 \text{ true }\right) \\ &= P\left(X \geq 15 \mid X \sim \text{Binomial}\left(20, e^{-0.7}\right)\right) \\ &= 1 - P(X \leq 14 \mid X \sim \text{Binomial}(20, 0.4966)) \\ &= 1 - 0.98028 \\ &= 0.01923 \,(\texttt{1 - pbinom(14, size = 20, prob = 0.4966)}) \end{aligned}$$

Furthermore

$$\begin{aligned} \beta &= P\left(\text{Do not reject } H_0 \mid H_1 \text{ true }\right) \\ &= P\left(X \leq 14 \mid X \sim \text{Binomial}\left(20, e^{-0.3}\right)\right) \\ &= P(X \leq 14 \mid X \sim \text{Binomial}(20, 0.7408)) \\ &= 0.42023 \,(\texttt{1 - pbinom(14, size = 20, prob = 0.7408)}) \end{aligned}$$

Sometimes α is called the size (or significance level) of the test and $\omega \equiv 1 - \beta$ is called the power of the test. Ideally, we would like to avoid error so we would like to make both α and β as small as possible. In other words, a good test will have small size, but large power. However, it is not possible to make α and β both arbitrarily small. For example if $C = \emptyset$ then $\alpha = 0$, but $\beta = 1$. On the other hand if $C = \mathbf{S} = \mathcal{R}^n$ then $\beta = 0$, but $\alpha = 1$.

In this section we have seen how to compute the type I and II error rates given a hypothesis testing procedure. But how should we construct a hypothesis testing procedure in the first place? We typically do this by first finding a *test statistic*: a quantity which can be computed from the data, whose distribution is known under H_0. We can then check the actual value of the test statistic is a plausible sample from this null distribution (the distribution of the test statistic under H_0). This is all rather abstract. How does it work in a concrete example?

4.5.2 Testing a normal mean (t-test)

4.5.2.1 Introduction

Suppose that we observe data $x_1, \ldots, x_n$ which are modelled as observations of i.i.d. $N\left(\mu, \sigma^2\right)$ random variables $X_1, \ldots, X_n$, and we want to test the null hypothesis

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_0.$$

We recall that

$$\sqrt{n}\frac{(\bar{X} - \mu)}{S} \sim t_{n-1}$$

and therefore, when H_0 is true, often written as under H_0,

$$\sqrt{n}\frac{\left(\bar{X} - \mu_0\right)}{S} \sim t_{n-1}.$$

Here $\sqrt{n}\left(\bar{X} - \mu_0\right)/s$ is a test statistic: it can be calculated from the data, and its null distribution is known: a t distribution with $n-1$ degrees of freedom.

This test is called a t-test. We reject the null hypothesis H_0 in favour of the alternative H_1 if the observed test statistic seems unlikely to have been generated by the null distribution.

Example 4.12 (Weight gain)**.** For the weight gain data, if x denotes the differences in weight gain, we have $\bar{x} = 0.8672$, $s = 0.9653$ and $n = 68$. Hence our test statistic for the null hypothesis $H_0 : \mu = \mu_0 = 0$ is

$$\sqrt{n}\frac{(\bar{x} - \mu_0)}{s} = 7.41.$$

The observed value of 7.41 does not seem reasonable from Figure 4.3, which shows the density of the t-distribution with 67 degrees of freedom, and a vertical line at the observed value of 7.41. So there may be evidence here to reject $H_0 : \mu = 0$.

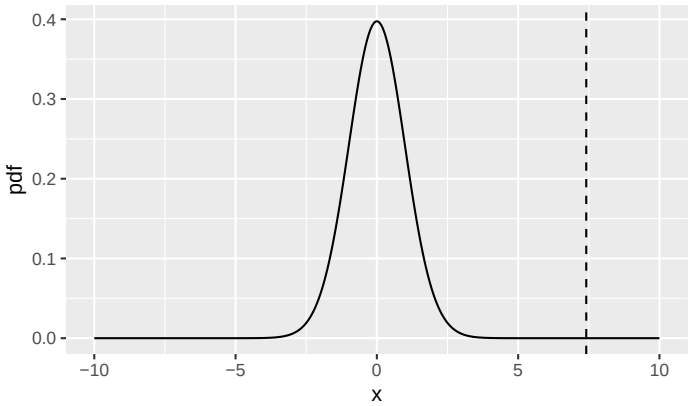


Figure 4.3: The density of the t_{67} distribution. The dashed line shows the observed value of the test statistic for the weight gain data.

Example 4.13 (Fast food waiting times)**.** Suppose the manager of the fast food outlet claims that the average waiting time is only 60 seconds. So, we want to test $H_0 : \mu = 60$. We have $n = 20, \bar{x} = 67.85, s = 18.36$. Hence our test statistic for the null hypothesis $H_0 : \mu = \mu_0 = 60$ is

$$\sqrt{n}\frac{(\bar{x} - \mu_0)}{s} = \sqrt{20}\frac{(67.85 - 60)}{18.36} = 1.91.$$

The observed value of 1.91 may or may not be reasonable from Figure 4.4, which shows the density of the t-distribution with 19 degrees of freedom, and a vertical line at the observed value of 1.91. This value is a bit out in the tail but we are not sure, unlike in the previous weight gain example. So how can we decide whether to reject the null hypothesis?

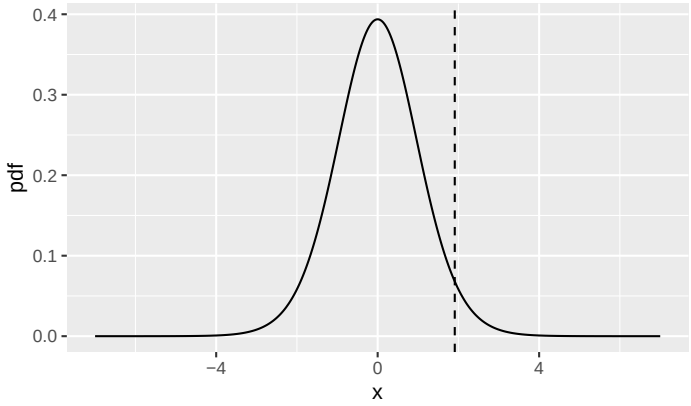


Figure 4.4: The density of the t_{19} distribution. The dashed line shows the observed value of the test statistic for the fast food data.

4.5.2.2 The significance level

In the weight gain example, it seems that there is enough evidence to reject H_0, but how extreme (far from the mean of the null distribution) should the test statistic be in order for H_0 to be rejected? The significance level of the test, α, is the probability that we will erroneously reject H_0 (called Type I error as discussed before). Clearly we would like α to be small, but making it too small risks failing to reject H_0 even when it provides a poor model for the observed data (Type II error). Conventionally, α is usually set to a value of 0.05, or 5%. Therefore we reject H_0 when the test statistic lies in a rejection region which has probability $\alpha = 0.05$ under the null distribution.

4.5.2.3 Rejection region

For the t-test, the null distribution is t_{n-1} where n is the sample size, so the rejection region for the test corresponds to a region of total probability $\alpha = 0.05$ comprising the 'most extreme' values in the direction of the alternative hypothesis. If the alternative hypothesis is two-sided, e.g. $H_1 : \mu \neq \mu_0$, then this is obtained as below, where the two shaded regions both have area (probability) $\alpha/2 = 0.025$.

The value of h depends on the sample size n and can be found in R with the `qt` command. Note that we need to put $n - 1$ in the `df` argument of `qt`. So for $n = 100$, we can find h with

```
qt(0.975, df = 99)
```

```
## [1] 1.984217
```

For some other values of n, we have

n	2	5	10	15	20	30	50	100	∞
h	12.71	2.78	2.26	2.14	2.09	2.05	2.01	1.98	1.96

where the last value for $n = \infty$ is obtained from the normal distribution.

However, if the alternative hypothesis is one-sided, e.g. $H_1 : \mu > \mu_0$, then the critical region will only be in the right tail: reject if the test statistic is greater than some value h. In this case, we choose h

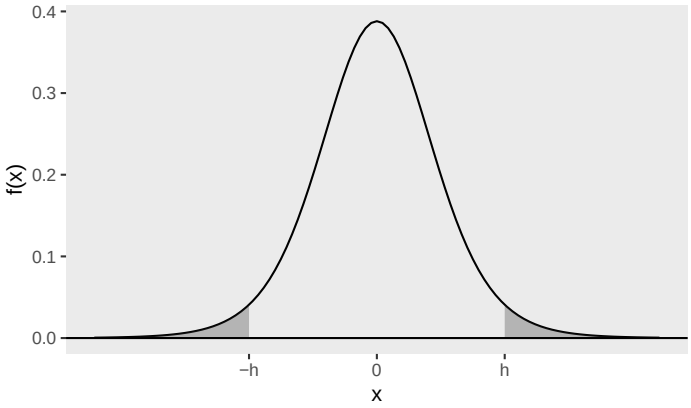


Figure 4.5: The shaded area shows the rejection region for a t-test. h is chosen to make the shaded area equal to the significance level $1 - \alpha$

such that the area to the right of h (under the t_{n-1} pdf) is α. So for $n = 100$ and $\alpha = 0.05$, we can find this h with

```
qt(0.95, df = 99)
```

```
## [1] 1.660391
```

For some other values of n, we have

n	2	5	10	15	20	30	50	100	∞
h	6.31	2.13	1.83	1.76	1.73	1.70	1.68	1.66	1.64

4.5.2.4 Summary of the t-test procedure

Suppose that we observe data $x_1, \ldots, x_n$ which are modelled as observations of i.i.d. $N\left(\mu, \sigma^2\right)$ random variables $X_1, \ldots, X_n$ and we want to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$:

1. Compute the test statistic
$$t = \sqrt{n}\frac{(\bar{x} - \mu_0)}{s}.$$

2. For chosen significance level α (usually 0.05) calculate the rejection region for t, which is of the form $|t| > h$ where $-h$ is the $\alpha/2$ percentile of the null distribution, t_{n-1}.

3. If your computed t lies in the rejection region, i.e. $|t| > h$, you report that H_0 is rejected in favour of H_1 at the chosen level of significance. If t does not lie in the rejection region, you report that H_0 is not rejected. (Never refer to 'accepting' a hypothesis.)

4.5.2.5 Examples

Example 4.14 (Fast food waiting times)**.** We would like to test $H_0 : \mu = 60$ against the alternative $H_1 : \mu > 60$, as this alternative will refute the claim of the store manager that customers only wait for a maximum of one minute. We calculated the observed value to be 1.91. This is a one-sided test and for a 5% level of significance, the critical value h will come from `qt(0.95, df = 19)` $= 1.73$. Thus the observed value is higher than the critical value so we will reject the null hypothesis, disputing the manager's claim regarding a minute wait.

Example 4.15 (Weight gain)**.** For the weight gain example $\bar{x} = 0.8671, s = 0.9653, n = 68$. Then, we would be interested in testing $H_0 : \mu = 0$ against the alternative hypothesis $H_1 : \mu \neq 0$ in the model that the data are observations of i.i.d. $N\left(\mu, \sigma^2\right)$ random variables.

- We obtain the test statistic

$$t = \sqrt{n}\frac{(\bar{x} - \mu_0)}{s} = \sqrt{68}\frac{(0.8671 - 0)}{0.9653} = 7.41.$$

- Under H_0 this is an observation from a t_{67} distribution. For significance level $\alpha = 0.05$ the rejection region is $|t| > 1.996$.
- Our computed test statistic lies in the rejection region, i.e. $|t| > 1.996$, so H_0 is rejected in favour of H_1 at the 5% level of significance.

In R we can perform the test as follows:

```
wtgain$diff <- wtgain$final - wtgain$initial
t.test(wtgain$diff)
```

```
##
##  One Sample t-test
##
## data:  wtgain$diff
## t = 7.4074, df = 67, p-value = 2.813e-10
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.6334959 1.1008265
## sample estimates:
## mean of x
## 0.8671612
```

This gives the results t = 7.4074, and df = 67.

4.5.2.6 p-values

The result of a test is most commonly summarised by rejection or non-rejection of H_0 at the stated level of significance. An alternative, which you may see in practice, is the computation of a p-value. This is the probability that the null distribution would have generated the actual observed value of the statistic or something more extreme. A small p-value is evidence against the null hypothesis, as it indicates that the observed data were unlikely to have been generated by the null distribution. In many examples a threshold of 0.05 is used, below which the null hypothesis is rejected as being insufficiently well-supported by the observed data.

For the t-test with a two-sided alternative, the p-value is given by:

$$p = P\left(|T| > |t_{\text{obs}}|\right) = 2P\left(T > |t_{\text{obs}}|\right),$$

where T has a t_{n-1} distribution and t_{obs} is the observed sample value.

However, if the alternative is one-sided and to the right then the p-value is given by:

$$p = P\left(T > t_{\text{obs}}\right),$$

where T has a t_{n-1} distribution and t_{obs} is the observed sample value.

A small p-value corresponds to an observation of T that is improbable (since it is far out in the low probability tail area) under H_0 and hence provides evidence against H_0. The p-value should not be misinterpreted as the probability that H_0 is true. H_0 is not a random event (under our models) and so

cannot be assigned a probability. The null hypothesis is rejected at significance level α if the p-value for the test is less than α.

$$\text{Reject } H_0 \text{ if p-value } < \alpha.$$

Example 4.16 (Fast food waiting times)**.** In the fast food example, a test of $H_0 : \mu = 60$ resulted in a test statistic $t = 1.91$. Then the p-value is given by

$$p = P(T > 1.91) = 0.036, \text{ when } T \sim t_{19}.$$

This is the area of the shaded region in Figure 4.6. In R it is

```
1 - pt(1.91, df = 19)
```

```
## [1] 0.03567359
```

The p-value 0.036 indicates some evidence against the manager's claim at the 5% level of significance but not the 1% level of significance.

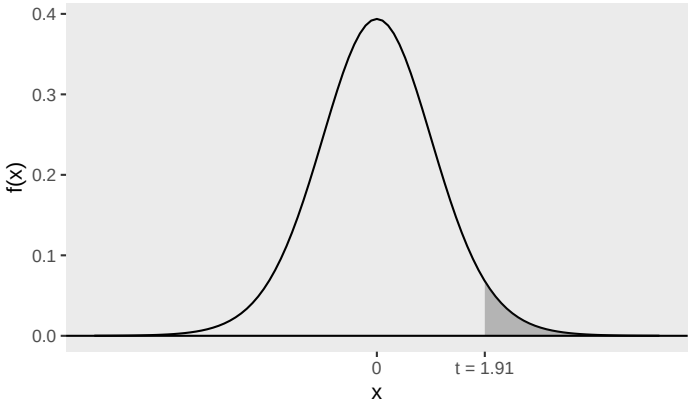


Figure 4.6: The area of the shaded region (under the t_{19} pdf) is the p-value for the one-sided hypothesis test in the fast food example

When the alternative hypothesis is two-sided the p-value has to be calculated from $P(|T| > t_{\text{obs}})$, where t_{obs} is the observed value and T follows the t-distribution with $n - 1$ df.

Example 4.17 (Weight gain)**.** Because the alternative is two-sided, the p-value is given by:

$$p = P(|T| > 7.41) = 2.78 \times 10^{-10} \approx 0.0, \text{ when } T \sim t_{67}.$$

This very small p-value indicates very strong evidence against the null hypothesis of no weight gain in the first year of college.

4.5.2.7 Equivalence of testing and interval estimation

Note that the 95% confidence interval for μ in the weight gain example has previously been calculated to be $(0.6335, 1.1008)$ in Example 4.9. This interval does not include the hypothesised value 0 of μ. Hence we can conclude that the hypothesis test at the 5% level of significance will reject the null hypothesis $H_0 : \mu = 0$.

This is because

$$\left| T_{\text{obs}} = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{s} \right| > h$$

implies and is implied by μ_0 being outside the interval

$$(\bar{x} - hs/\sqrt{n}, \bar{x} + hs/\sqrt{n}).$$

Notice that h is the same in both. For this reason we often just calculate the confidence interval and take the reject/do not reject decision merely by inspection.

4.5.3 Two sample t-tests

Suppose that we observe two samples of data, $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, and that we propose to model them as observations of

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N\left(\mu_X, \sigma_X^2\right)$$

and

$$Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} N\left(\mu_Y, \sigma_Y^2\right)$$

respectively, where it is also assumed that the X and Y variables are independent of each other. Suppose that we want to test the hypothesis that the distributions of X and Y are identical, that is

$$H_0 : \mu_X = \mu_Y, \quad \sigma_X = \sigma_Y = \sigma$$

against the alternative hypothesis

$$H_1 : \mu_X \neq \mu_Y.$$

In the Chapter 3 we proved that

$$\bar{X} \sim N\left(\mu_X, \sigma_X^2/n\right) \text{ and } \bar{Y} \sim N\left(\mu_Y, \sigma_Y^2/m\right)$$

and therefore

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

Hence, under H_0,

$$\bar{X} - \bar{Y} \sim N\left(0, \sigma^2\left[\frac{1}{n} + \frac{1}{m}\right]\right) \Rightarrow \sqrt{\frac{nm}{n+m}}\frac{(\bar{X} - \bar{Y})}{\sigma} \sim N(0,1).$$

The involvement of the (unknown) σ above means that this is not a pivotal test statistic. It will be proved in MATH2011 that if σ^2 is replaced by its unbiased estimator S^2, which here is the two-sample estimator of the common variance, given by

$$S^2 = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 + \sum_{i=1}^m \left(Y_i - \bar{Y}\right)^2}{n+m-2},$$

then

$$\sqrt{\frac{nm}{n+m}}\frac{(\bar{X} - \bar{Y})}{S} \sim t_{n+m-2}.$$

Hence

$$t = \sqrt{\frac{nm}{n+m}}\frac{(\bar{x} - \bar{y})}{s}$$

is a test statistic for this test. The rejection region is $|t| > h$ where $-h$ is the $\alpha/2$ (usually 0.025) percentile of t_{n+m-2}.

From the hypothesis testing, a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\bar{x} - \bar{y} \pm h\sqrt{\frac{n+m}{nm}}s,$$

where $-h$ is the $\alpha/2$ (usually 0.025) percentile of t_{n+m-2}.

Example 4.18 (Fast food waiting times)**.** In this example, we would like to know if there are significant differences between the AM and PM waiting times. Here $n = m = 10$, $\bar{x} = 68.9$, $\bar{y} = 66.8$, $s_x^2 = 538.22$ and $s_y^2 = 171.29$. From this we calculate,

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} = 354.8,$$

and

$$t_{\text{obs}} = \sqrt{\frac{nm}{n+m}}\frac{(\bar{x}-\bar{y})}{s} = 0.25.$$

This is not significant as the critical value $h =$ `qt(0.975, df = 18)` $= 2.10$ is larger in absolute value than 0.25. We can conduct the test easily in R:

```
fastfood_am <- c(38, 100, 64, 43, 63, 59, 107, 52, 86, 77)
fastfood_pm <- c(45, 62, 52, 72, 81, 88, 64, 75, 59, 70)
t.test(fastfood_am, fastfood_pm)
```

```
##
##  Welch Two Sample t-test
##
## data:  fastfood_am and fastfood_pm
## t = 0.24929, df = 14.201, p-value = 0.8067
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.9434  20.1434
## sample estimates:
## mean of x mean of y
##      68.9      66.8
```

It automatically calculates the test statistic as 0.249 and a p-value of 0.8067. It also obtains the 95% CI given by $(-15.94, 20.14)$.

4.5.4 Paired t-test

Sometimes the assumption that the X and Y variables are independent of each other is unlikely to be valid, due to the design of the study. The most common example of this is where $n = m$ and data are paired. For example, a measurement has been made on patients before treatment (X) and then again on the same set of patients after treatment (Y). Recall the weight gain example is exactly of this type. In such examples, we proceed by computing data on the differences

$$z_i = x_i - y_i, \quad i = 1, \ldots, n$$

and modelling these differences as observations of i.i.d. $N\left(\mu_z, \sigma_Z^2\right)$ variables $Z_1, \ldots, Z_n$. Then, a test of the hypothesis $\mu_X = \mu_Y$ is achieved by testing $\mu_Z = 0$, which is just a standard (one sample) t-test, as described previously.

Example 4.19. Water-quality researchers wish to measure the biomass to chlorophyll ratio for phytoplankton (in milligrams per litre of water). There are two possible tests, one less expensive than the other. To see whether the two tests give the same results, ten water samples were taken and each was measured both ways. The results are as follows:

Test 1 (x)	45.9	57.6	54.9	38.7	35.7	39.2	45.9	43.2	45.4	54.8
Test 2 (y)	48.2	64.2	56.8	47.2	43.7	45.7	53.0	52.0	45.1	57.5

To test the null-hypothesis

$$H_0 : \mu_Z = 0 \text{ against } H_1 : \mu_Z \neq 0$$

we use the test statistic $t = \sqrt{n}\frac{\bar{z}}{s_z}$, where $s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$.

From the hypothesis testing, a $100(1-\alpha)\%$ confidence interval for μ_Z is given by $\bar{z} \pm h\frac{s_z}{\sqrt{n}}$, where h is the critical value of the t distribution with $n-1$ degrees of freedom. In R we perform the test as follows:

```
x <- c(45.9, 57.6, 54.9, 38.7, 35.7, 39.2, 45.9, 43.2, 45.4, 54.8)
y <- c(48.2, 64.2, 56.8, 47.2, 43.7, 45.7, 53.0, 52.0, 45.1, 57.5)

t.test(x, y, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = -5.0778, df = 9, p-value = 0.0006649
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -7.531073 -2.888927
## sample estimates:
## mean difference
##           -5.21
```

This gives the test statistic $t_{\text{obs}} = -5.0778$ with a df of 9 and a p-value $= 0.0006649$. Thus we reject the null hypothesis. The associated 95% CI is $(-7.53, -2.89)$.

The values of the second test are significantly higher than the ones of the first test, and so the second test cannot be considered as a replacement for the first.

Chapter 5

Simple Linear Regression

5.1 What is regression?

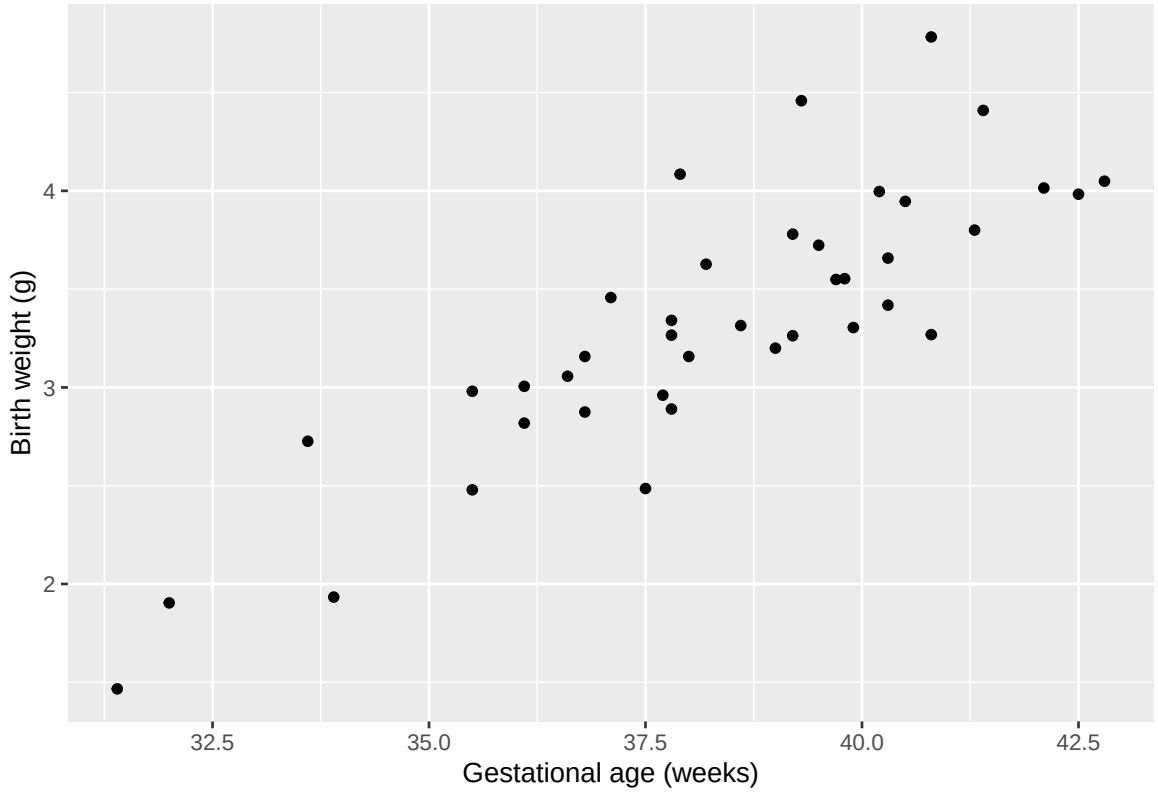
So far, we have looked at simple statistical models, with a single variable of interest. We have estimated the distribution of that variable based on random samples from that one variable. In many real-world problems, we instead want to understand the relationship between multiple variables.

Regression seeks to model how one "output" variable, called the **response** variable, Y, depends on one or more "input" variables x, which we will call **covariates**.

The response Y is sometimes called the dependent variable. The covariates x are sometimes called independent or explanatory variables.

In MATH1063, we will study only **simple** regression models, which have a single covariate x.

For instance, suppose we interested in how the birth weight of babies depends on gestational age (in weeks). The response, Y, is birth weight in g. The covariate, x is gestational age (in weeks). Suppose we have the following data (this data is fictional, but based on real data for male singleton births in Canada).



5.2 Simple linear regression

In simple linear regression, we model the dependence of the response on the covariate as a straight line plus an error term. That is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n,$$

where β_0 is an **intercept** parameter, β_1 is a **slope** parameter and ϵ_i is a random error term. We usually model $\epsilon_i \sim N(0, \sigma^2)$, with variance parameter σ^2. There are three unknown parameters, β_0, β_1 and σ^2, which we want to estimate from the data.

We have

$$E(Y_i) = \beta_0 + \beta_1 x_i.$$

Any choice of β_0 and β_1 give a different straight line. We call β_0 and β_1 the **regression parameters**.

Out of all possible choices of straight line (choices of the regression parameters β_0 and β_1), which one should we choose?

5.3 Estimating the regression parameters

To estimate the unknown regression parameters, we attempt to make the associated straight line as close as possible to the data. We measure the distance between the line and the data with the sum of squares criterion:

$$\mathrm{SS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

To estimate β_0 and β_1, we find the values which minimise this sum of squares criterion. These estimates are called the **least squares estimates**.

To find the least squares estimates, we take partial derivatives of $SS(\beta_0, \beta_1)$ with respect to β_0 and β_1, set them to zero, and solve for the parameters.

First, differentiate with respect to β_0, to give

$$\frac{\partial SS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i).$$

Setting to zero:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0,$$

so

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

which gives

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Next, differentiate with respect to β_1, to give

$$\frac{\partial SS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i).$$

Setting to zero gives

$$\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0,$$

or

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Substituting $\beta_0 = \bar{y} - \beta_1 \bar{x}$ gives

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0,$$

so

$$\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} = \beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right).$$

So

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Thus, the least squares estimates are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

These are unbiased estimators of the regression parameters. We do not derive this result here; the derivation will be covered in the second-year module, Statistical Modelling I).

5.4 Estimating the variance parameter

In addition to estimating the regression parameters β_0 and β_1, we also need to estimate the variance parameter σ^2 of the error terms.

We estimate σ^2 with

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates.

This estimator is unbiased for σ^2. We do not derive this result here; the derivation will be covered in the second-year module, Statistical Modelling I.

5.5 Model fitting in R

In practice, we do not need to apply these formulas by hand to find the least squares estimates. Instead, we can use the `lm` function in `R` to find the estimates. For instance, if the birth weight data is stored in a data frame called `bw`, with columns `weight` for birthweight (which is the response, y) and `ga` for gestational age (which is the covariate, x), then we can fit the simple linear regression model with:

```
mod <- lm(weight ~ ga, data = bw)
```

We can inspect the fitted model:

```
mod
```

```
##
## Call:
## lm(formula = weight ~ ga, data = bw)
##
## Coefficients:
## (Intercept)           ga
##     -5.1062       0.2203
```

We see $\hat{\beta}_0$ = -5.1062 and $\hat{\beta}_1$ = 0.2203.

We can get more information about the model fit, including standard errors for the estimates, by using the `summary` function in R:

```
summary(mod)
```

```
##
## Call:
## lm(formula = weight ~ ga, data = bw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67051 -0.26954 -0.08779  0.17978  0.90572
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.10619    0.86147  -5.927 7.16e-07 ***
## ga           0.22033    0.02245   9.814 5.74e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

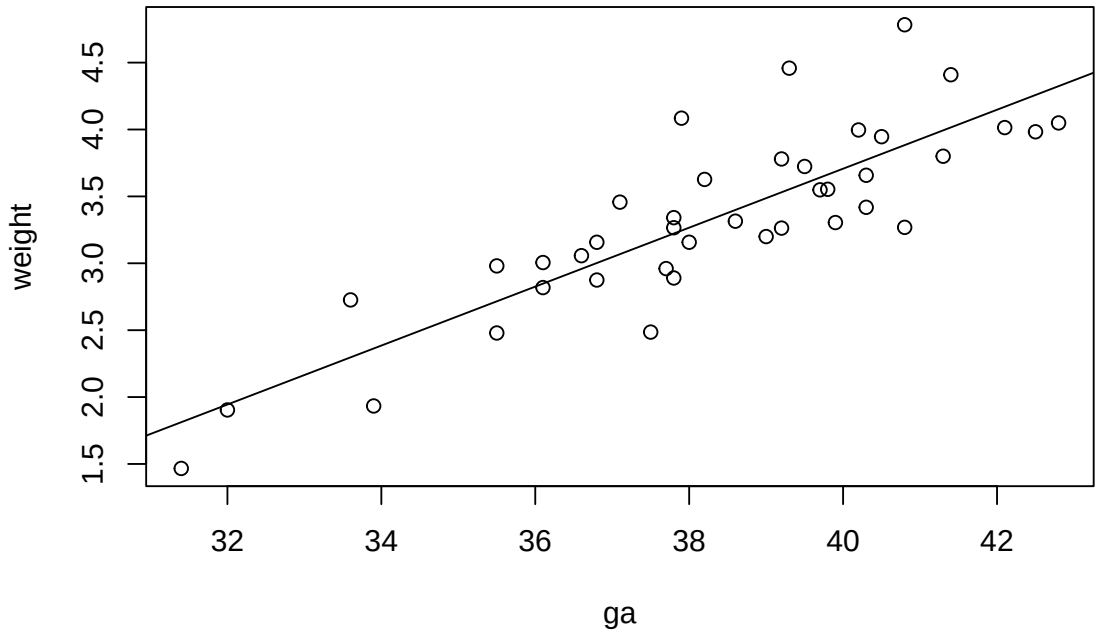
```
## Residual standard error: 0.3731 on 38 degrees of freedom
## Multiple R-squared:  0.7171, Adjusted R-squared:  0.7096
## F-statistic: 96.32 on 1 and 38 DF,  p-value: 5.735e-12
```

You can learn more about expressing uncertainty in simple linear regression the second-year module, Statistical Modelling I.

The "Residual standard error" (0.3731) is an estimate of σ. The variance estimate $\hat{\sigma}^2$ can be found by squaring this number (in this case, $\hat{\sigma}^2 = 0.1392$).

We can also plot the data with our fitted line overlaid:

```
plot(weight ~ ga, data = bw)
abline(mod)
```



5.6 Limitations of simple linear regression

While simple linear regression is a powerful and widely used tool, it has several important limitations:

- **Only one covariate:** Simple linear regression models the relationship between the response and a single covariate. In many real-world situations, multiple variables may influence the response.
- **Linearity assumption:** The model assumes that the relationship between the response and the covariate is linear. If the true relationship is non-linear, the model may not fit well.
- **Constant variance (homoscedasticity):** The errors are assumed to have constant variance across all values of the covariate. If the variance of the errors changes (heteroscedasticity), the model's estimates and inferences may be unreliable.
- **Normality of errors:** The errors are assumed to be normally distributed. If this assumption is violated, inference based on the model may not be valid.

These limitations, and methods to address them, will be covered in more detail in the second-year module, Statistical Modelling I.