



MAS1403

Quantitative Methods for Business Management

Semester 2

Dr. Dave Walshaw

School of Mathematics & Statistics

Chapter 1

Estimation

1.1 Introduction

Statistics is the science that studies the collection and interpretation of data to enable us to make **inferences** from the **sample data** to the **population** from which the sample has been drawn. As the sample we draw from the population gets bigger and bigger, it becomes increasingly difficult to “picture” the data just by looking at the raw observations; the data are easier to understand if we can find some way of *summarising* them.

Recall from semester 1 that data can be summarised in two ways: **graphically** and **numerically**. Various graphical techniques available for summarising data were presented, including stem and leaf plots, bar charts, histograms, relative frequency histograms and frequency polygons. You were also introduced to two methods of summarising data numerically, through measures of **location** and measures of **spread**.

A measure of location is a quantity which is ‘typical’ of the data; examples of such measures include

- (i) the **sample mean** (“add them up and divide by how many we have”);
- (ii) the **sample median** (“the one in the middle”), and
- (iii) the **sample mode** (“the value which occurs most often”).

A measure of spread is a value which quantifies the variability in the data (or how “spread out” the observations are); examples include

- (i) the **range** (“largest minus smallest”);
- (ii) the **variance** (“average squared distance of observations from the mean” – the standard deviation is the square root of this value), and
- (iii) the **inter-quartile range** (“upper quartile minus lower quartile” – this value represents the middle 50% of the data, as the lower quartile has one quarter of the data less than it, and the upper quartile has three-quarters of the data less than it).

1.2 Estimation

Last semester we concentrated on pure description of data, although we recognised that this might prompt us to ask pertinent questions about the population from which the sample was drawn. What exactly does the sample, often a tiny subset, tell us of the population? We can never observe the whole population, even if it is finite, except at enormous expense, and so the population mean and variance (or indeed any aspect of the population distribution) can never be known exactly. We call these unknown quantities **parameters** and use Greek letters to denote them: μ (“mu”) is the symbol commonly used for the population mean and σ (“sigma”) for the population standard deviation. Hopefully (and if we have a representative sample), the sample mean ($\bar{x}$) will be quite close to the true population mean μ ; likewise, the sample variance (S^2) will be a good estimator for σ . In this section, we concentrate on $\bar{X}$ as an estimator for μ . Note that we use upper case for the estimator and lower case for a particular estimate.

Before we can use our sample of n observations we must ask the question: Is $\bar{x}$ a “good” estimate of μ ? How do we **infer** (find something out about) the unknown μ using $\bar{x}$? So long as the sample size n is fairly large, we can hope that $\bar{x}$ is close to μ . But how close is it? To answer this question, we must make some plausible assumptions about the population.

Suppose $x_1, x_2, \dots, x_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution. Then their mean, $\bar{x}$, is an observation from a $N(\mu, \sigma^2/n)$ distribution. The standard deviation $\sigma/\sqrt{n}$ and is usually called the **standard error** to remind ourselves that $\bar{x}$ is used as an estimate of the unknown μ .

The Central Limit Theorem

Suppose now that $x_1, x_2, \dots, x_n$ are a random sample from *any* population, with mean μ and variance σ^2 . If n is large, then

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

no matter what the original probability distribution.

1.3 Interval estimation – Case 1: Known variance σ^2

The values we calculate for sample means and variances are **point estimates**; they are single values based on a limited sample of the whole population. Suppose that we wish to estimate the mean μ of a population. The natural estimate for μ is the sample mean $\bar{x}$. However, $\bar{x}$ is never exactly equal to μ ; all we really hope is that $\bar{x}$ will be close to μ . One way of improving our inference is to construct **interval estimates** or **confidence intervals**. We simply place an interval over the point estimate for μ which allows us to say (with a certain level of confidence) within what range the population mean lies. The calculation of these intervals depends on the size of our sample (n), the level of confidence we choose, and whether or not the population variance (σ^2) is known.

We know from the results above that, if our random sample is drawn from a normal distribution, or if n is large (i.e. $n \geq 30$), then

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

If we initially assume we know the population variance σ^2 , we can standardise $\bar{X}$ as we did last semester; i.e.

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}.$$

Recall that the standard normal distribution is $Z \sim N(0, 1)$, i.e. Z has zero mean and variance (and so standard deviation) 1; also recall that approximately 95% of the standard normal distribution lies between -1.96 and 1.96 , i.e.

$$\Pr(-1.96 < Z < 1.96) = 0.95.$$

If we think about this graphically, we can see this more clearly:

Since we have an expression for Z , we can rearrange this expression:

$$\begin{aligned} \Pr\left(-1.96 < \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} < 1.96\right) &= 0.95 \\ \Pr\left(\bar{X} - 1.96\sqrt{\sigma^2/n} < \mu < \bar{X} + 1.96\sqrt{\sigma^2/n}\right) &= 0.95. \end{aligned}$$

Thus, we can say that the two values $\bar{x} - 1.96\sqrt{\sigma^2/n}$ and $\bar{x} + 1.96\sqrt{\sigma^2/n}$ are the lower and upper bounds (respectively) of the (95%) confidence interval.

This can be written as

$$\bar{x} \pm 1.96\sqrt{\sigma^2/n}.$$

Consider the following example.

Example

A coffee machine fills cups with hot water; the variance of the filling process is known to be $\sigma^2 = 10 \text{ (ml)}^2$. A sample of 100 filled cups gives a sample mean and we have calculated a sample mean of $\bar{x} = 40\text{ml}$. What is the 95% confidence interval of the population mean μ ?

We already have a formula for the 95% confidence interval:

$$\bar{x} \pm 1.96\sqrt{\sigma^2/n}.$$

So, inputting our values, we get

$$\begin{aligned} 40 &\pm 1.96\sqrt{10/100}, & \text{i.e.} \\ 40 &\pm 0.61. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is (39.39, 40.61).

What would happen if the sample size increased to 200 and everything else remained the same? We'd get

$$\begin{aligned} 40 &\pm 1.96\sqrt{10/200}, & \text{i.e.} \\ 40 &\pm 0.44. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is (39.56, 40.44). This should be intuitive, since as the sample size increases we are becoming more sure of our estimate for the population value.

What would be the 99% confidence interval in this case? From tables for the standard normal distribution, we can find that

$$\Pr(-2.576 < Z < 2.576) = 0.99;$$

hence, the 99% confidence interval is given by

$$\bar{x} \pm 2.576\sqrt{\sigma^2/n},$$

in this case giving

$$\begin{aligned} 40 &\pm 2.576\sqrt{10/200}, & \text{i.e.} \\ 40 &\pm 0.58. \end{aligned}$$

Hence, the 99% confidence interval for the population mean μ is (39.42, 40.58). You should note that this gives a wider range than the 95% confidence interval. This is (again) intuitive; as you increase the percentage of certainty you want, you will naturally incorporate a larger range.

1.3.1 Case 1 (known σ^2): Summary

- (i) Calculate the sample mean $\bar{x}$ from the data;
- (ii) Calculate your interval! For example,
 - for a 90% confidence interval, use the formula

$$\bar{x} \pm 1.645 \times \sqrt{\sigma^2/n};$$

- for a 95% confidence interval, use the formula

$$\bar{x} \pm 1.96 \times \sqrt{\sigma^2/n};$$

- for a 99% confidence interval, use the formula

$$\bar{x} \pm 2.576 \times \sqrt{\sigma^2/n}.$$

1.3.2 Exercises

1. A company packs sacks of flour. The standard deviation of the filling process is $\sigma = 10$ g. A sample of 50 bags is taken and weighed and the resulting sample mean is 750g. Compute a 95% and 99% confidence interval for the mean weight of a bag of flour.
2. A company manufactures bolts with a process variance of 50. A sample of 100 bolts is taken and measured and their average length is calculated as 98mm. What is the 95% confidence interval for the mean length of bolts? If the bolts are designed to be 100mm long, is the process satisfactory?
3. AG Automotive claims that its latest hatchback will achieve a fuel efficiency figure of 70 miles per gallon (mpg) on average. A random sample of 8 vehicles was taken, and, after each vehicle was run for a month, the following sample of efficiency figures were obtained

59.1, 65.8, 76.1, 72.3, 69.3, 71.5, 67.9, 70.5

Assuming that the population standard deviation is known to be 4mpg, calculate a 95% confidence interval for the mean fuel efficiency figure of the AG Automotive hatchback. Do you think that the manufacturer's claim of 70mpg is justified?

1.4 Interval Estimation – Case 2: Unknown variance σ^2

If the population variance σ^2 is unknown, we can no longer use the normal distribution and instead have to use the t -distribution to calculate confidence intervals. We have seen that when our random sample follows a normal distribution, or indeed any distribution (if the sample size is large), then the sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$. From this, it follows that

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}},$$

where Z is the standard normal distribution, i.e. $Z \sim N(0, 1)$. However, if the population variance is unknown, then the quantity

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

does *not* have a $N(0, 1)$ distribution (note that the *population* variance σ^2 in Z has been replaced with the *sample* variance S^2 in T). Instead it has a Student's t -distribution. This distribution is similar to the $N(0, 1)$ distribution in that it is symmetrical and bell-shaped, but it is more heavily tailed to allow for greater uncertainty in $\bar{X}$ since the true variability is now unknown. Its exact shape is determined by one parameter called the “degrees of freedom”. Table 1.1 gives critical values of Student's t -distribution with various degrees of freedom. These numbers depend on two quantities: ν , the degrees of freedom, and p , a probability.

The expression for the confidence interval in this case is similar to the case where σ^2 is known:

$$\bar{x} \pm t_{p/2} \sqrt{s^2/n},$$

where σ^2 has been replaced with the sample variance s^2 and $t_{p/2}$ is the appropriate value from the t -distribution tables. Note that $t_{p/2}$ is the value such that

$$\Pr(-t_{p/2} < T < t_{p/2}) = 100(1 - p)\%.$$

But how do we find this value?

First, we need to find p . If we are looking for the 95% confidence interval, we are looking for the value of p which satisfies the equation

$$\begin{aligned} 100(1 - p)\% &= 95\%, & \text{i.e.} \\ p &= 0.05. \end{aligned}$$

We would look up the value in the t tables (table 1.1) in the p column, or in this case the 5% column.

We also need to know which row to look in. The rows are given as the degrees of freedom, ν , where $\nu = n - 1$. Hence, if our sample was of size $n = 10$ and we were looking for the 95% confidence interval, we would look in the $\nu = 9$ row and the $p = 5\%$ column to give us a value of 2.262 to use in our calculation.

Example

A sample of size 15 is taken from a larger population; the sample mean is calculated as 12 and the sample variance as 25. What is the 95% confidence interval for the population mean μ ?

We know that the confidence interval is given by

$$\bar{x} \pm t_{p/2} \sqrt{s^2/n},$$

where

$$\begin{aligned} n &= 15, \\ \nu &= n - 1 = 15 - 1 = 14, \\ p &= 5\%, \\ \bar{x} &= 12 \quad \text{and} \\ s^2 &= 25. \end{aligned}$$

We can find our t value by looking in the $p = 5\%$ column and the $\nu = 14$ row, giving a value of 2.145. Putting what we know into our expression, we get

$$\begin{aligned} 12 &\pm t_{2.5\%} \sqrt{\frac{25}{15}} \\ 12 &\pm 2.145 \sqrt{\frac{25}{15}} \quad \text{i.e.} \\ 12 &\pm 2.77. \end{aligned}$$

Hence, the confidence interval is (9.23, 14.77).

1.4.1 Application of Confidence Intervals

You might be asking: “why do we bother calculating confidence intervals?”. Firstly, by calculating a confidence interval for the population mean, it allows us to see how confident we are of the point estimate we have calculated. The wider the range, the less precise we can be about the population value.

Secondly, it allows us to start looking at differences between groups. If the confidence intervals for two samples do not overlap, this could suggest that they are from separate populations. Or if we have a known value for a population and this does not fall within the confidence interval of our sample, this could suggest that there is something different about this sample. For example, if the average daily taking for all shops in a chain was £12550, and we calculated the confidence interval for the population mean of one particular branch as (£11000, £11500), we can say that this branch is out of line with the company as a whole, and it appears that its daily takings are lower.

Example

A credit card company wants to determine the mean income of its card holders. It also wants to find out if there are any differences in mean income between males and females. A random sample of 225 male card holders and 190 female card holders was drawn, and the following results obtained:

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

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

95% confidence interval for male income

The true population variance, σ^2 , is unknown, so we can’t use the approach of section 1.3. Instead, we follow that of section 1.4 which uses the t -distribution, i.e.

$$\bar{x} \pm t_{p/2} \times \sqrt{s^2/n}.$$

Here,

$$\begin{aligned}\bar{x} &= 16450, \\ s^2 &= 3675^2 = 13505625 \quad \text{and} \\ n &= 225.\end{aligned}$$

The value $t_{p/2}$ must be found from table 1.1. Recall that the degrees of freedom, $\nu = n - 1$, and so here we have $\nu = 225 - 1 = 224$. Notice that table 1.1 only gives value of ν up to 29; for higher values, we use the ∞ row. Since we require a 95% confidence interval, we read down

the 5% column, giving a t value of 1.96 (recall that this is the same as the value used if σ^2 were *known* and we used the normal distribution – that’s because the t -distribution converges to the normal distribution as the sample size increases). Thus, the 95% confidence interval for μ is found as

$$\begin{aligned} 16450 \pm 1.96 \times \sqrt{13505625/225}, & \quad \text{i.e.} \\ 16450 \pm 480.2. & \end{aligned}$$

So, the 95% confidence interval is (£15969.80, £16930.20).

95% confidence interval for female income

Again, the true population variance, σ^2 , is unknown, so we can’t use the approach of section 1.3, and so again we use the t -distribution as in section 1.4:

$$\bar{x} \pm t_{p/2} \times \sqrt{s^2/n}.$$

Now,

$$\begin{aligned} \bar{x} &= 13220, \\ s^2 &= 3050^2 \\ &= 9302500, & \text{and} \\ n &= 190. \end{aligned}$$

Again, since the sample size is large, we use the ∞ row of table 1.1 to obtain the value of $t_{p/2}$, and so the 95% confidence interval for μ is found as

$$\begin{aligned} 13220 \pm 1.96 \times \sqrt{9302500/190}, & \quad \text{i.e.} \\ 13220 \pm 1.96 \times 221.27, & \quad \text{i.e.} \\ 13220 \pm 433.69. & \end{aligned}$$

So, the 95% confidence interval is (£12786.31, £13653.69).

Since the 95% confidence intervals for males and females *do not overlap*, there *is* evidence to suggest that males’ and females’ incomes, on average, are different. Further, it appears that male card holders earn more than women.

1.4.2 Case 2 (Unknown σ^2): Summary

- (i) Calculate the sample mean $\bar{x}$ and the sample variance s^2 from the data;
- (ii) For a $100(1 - p)\%$ confidence interval, look up the value of t under column p , row ν of table 1.1, remembering that $\nu = n - 1$. Note that, for a 90% confidence interval, $p = 10\%$, for a 95% confidence interval, $p = 5\%$ and for a 99% confidence interval, $p = 1\%$;
- (iii) Calculate your interval, using

$$\bar{x} \pm t_{p/2} \times \sqrt{s^2/n}.$$

1.4.3 Exercises

1. Suppose that you wish to construct a confidence interval for a population mean μ . A sample of size $n = 50$ is drawn and the sample mean $\bar{x} = 15$ computed. Given the following statements (a)-(e), decide which relate to case 1 (σ^2 is known) and which relate to case 2 (σ^2 is unknown).
 - (a) $s^2 = 91.39$.
 - (b) The sample variance is 91.39.
 - (c) The population standard deviation is 10.
 - (d) The sample standard deviation is 9.56.
 - (e) The process variance is known to be 100.
2. A class of students has sat an exam. A sample of 40 students is taken and their marks produced first. This sample has a mean of 55% and a sample variance of 100. Calculate the 95% confidence interval for the mean mark of the class as a whole.
3. A sample of 12 students is taken and their mean IQ calculated as 110 (with a sample variance of 220). What is the 95% and 99% confidence intervals for the population value based on this sample? What do you notice about the calculated interval as the confidence level increases? Do either of these two confidence intervals contain the known population mean IQ of 100?
4. The following are the number of cars caught speeding each day on one speed camera over a two week period.

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

What is the 95% confidence interval for this sample? How does this compare with the whole Northumbria police average of 8 per day per camera?

		p				
		50%	20%	10%	5%	1%
	1	1.00	3.078	6.314	12.706	63.657
	2	0.816	1.886	2.920	4.303	9.925
	3	0.765	1.638	2.353	3.182	5.841
	4	0.741	1.533	2.132	2.776	4.604
	5	0.727	1.476	2.015	2.571	4.032
	6	0.718	1.440	1.943	2.447	3.707
	7	0.711	1.415	1.895	2.365	3.449
ν	8	0.706	1.397	1.860	2.306	3.355
	9	0.703	1.383	1.833	2.262	3.250
	10	0.700	1.372	1.812	2.228	3.169
	11	0.697	1.363	1.796	2.201	3.106
	12	0.695	1.356	1.782	2.179	3.055
	13	0.694	1.350	1.771	2.160	3.012
	14	0.692	1.345	1.761	2.145	2.977
	15	0.691	1.341	1.753	2.131	2.947
	16	0.690	1.337	1.746	2.120	2.921
	17	0.689	1.333	1.740	2.110	2.898
	18	0.688	1.330	1.734	2.101	2.878
	19	0.688	1.328	1.729	2.093	2.861
	20	0.687	1.325	1.725	2.086	2.845
	21	0.686	1.323	1.721	2.080	2.831
	22	0.686	1.321	1.717	2.074	2.819
	23	0.685	1.319	1.714	2.069	2.807
	24	0.685	1.318	1.711	2.064	2.797
	25	0.684	1.316	1.708	2.060	2.787
	26	0.684	1.315	1.706	2.056	2.779
	27	0.684	1.314	1.703	2.052	2.771
	28	0.683	1.313	1.701	2.048	2.763
	29	0.683	1.311	1.699	2.045	2.756
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	∞	0.674	1.282	1.645	1.960	2.576

Table 1.1: Tabulated values of t for which $\Pr(|T| > t) = p$, where T has a t -distribution with ν degrees of freedom

Chapter 2

Hypothesis tests

2.1 Introduction

We have seen that confidence intervals can be used to make inferences about population parameters. Sometimes, you may be asked to assess whether or not a parameter takes a specific value. For example, whether the population mean $\mu = 5$. One way of re-expressing this question is to ask whether the parameter value is plausible in light of the data. A simple check to see whether the value is contained in a 95% confidence interval will provide an answer. An alternative method, called a **hypothesis test**, is available which does not require a confidence interval to be calculated. It is used extensively in reporting experimental results.

A hypothesis test is a rule for establishing whether or not a set of data is consistent with a hypothesis about a parameter of interest. The **null hypothesis** is a statement that a parameter has a certain value, and is usually written as H_0 . For example, $H_0 : \mu = 2.7$, or $H_0 : \sigma^2 = 12.4$. If the null hypothesis is not true, what alternatives are there? Usually, the **alternative hypothesis** is written as H_1 . Examples include $H_1 : \mu \neq 2.7$ and $H_1 : \sigma^2 \neq 12.4$.

Based on the information we have in our sample, we'd like to go with either the null hypothesis or the alternative hypothesis. We might use our sample as evidence to suggest that, for example, the population mean could well be equal to 2.7; alternatively, the sample might give evidence to the contrary and suggest that the population mean is not equal to 2.7, or for that matter the sample might suggest that the population mean is less than 2.7!

Here's a quick example. Suppose you are going on holiday to Kavos in March. Before you leave, a friend tells you that Kavos has an average of 10 hours sunshine a day during March. On the first three days of your holiday there are 7, 8 and 9 hours of sunshine respectively. You consider that this is evidence that your friend is wrong. Thus, the null hypothesis would state that the average sunshine hours per day is 10 (as suggested by your friend), and your alternative hypothesis might state that the average sunshine hours per day *is less than* 10. You might be tempted to go with the alternative hypothesis, since on your first three days you have observed less than 10 hours of sunshine. However, this sample of three days could be a fluke result – you might have chosen the most miserable period in March for years for your holiday. **Your test results are not conclusive; they only give you evidence *against* a particular belief and we**

should be aware of the *weight* of that evidence.

2.2 Methodology for hypothesis testing

All hypothesis tests follow the same basic methodology, although the actual calculations may vary depending on the data available.

1. State the null hypothesis (H_0)

We use a hypothesis test to throw light on whether or not this statement is true. For example, you might ask “is the population mean equal to 10?”, or “are the two population means equal?”; such hypotheses are expressed in the following way:

$$\begin{aligned} H_0 &: \mu = 10, && \text{and} \\ H_0 &: \mu_1 = \mu_2; && \text{or maybe} \\ H_0 &: \mu = c, \end{aligned}$$

Where c could be any constant.

2. State the alternative hypothesis (H_1)

This is the conclusion to be reached if the null hypothesis is found to be false. For example, “the population mean does not equal 10”, or even “the population mean is less than 10”; to test for two different populations, we might say “the two population means are different”. Remember, we can never reject the null hypothesis with certainty; the most we can say is that there is evidence against the null hypothesis, and so evidence in favour of the alternative. Such alternative hypotheses are expressed in the following way:

$$\begin{aligned} H_1 &: \mu \neq 10 && \text{or maybe} \\ H_1 &: \mu < 10. \end{aligned}$$

To test for two different populations, we might use

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

3. Calculate the test statistic

The value calculated from the sample which is used to perform the test is called the **test statistic**. It usually has a similar nature to the population value mentioned in the null hypothesis.

4. Find the p -value of the test

The probability that such an extreme test statistic occurs, *assuming that H_0 is true*, is called the **p -value**, and can be found by comparing the test statistic to values from statistical tables (the tables used will depend on the nature of the test) or using a computer package such as Minitab (see chapter 7).

5. Reach a conclusion

A small p -value suggests that our test statistic is unlikely to occur if H_0 is true, and so

we reject H_0 in favour of the alternative H_1 . But what constitutes a “small” p -value? One commonly used yardstick, or **significance level**, is 5%, or 0.05, though others can be used. The *smaller* the p -value, the *more* evidence there is to reject the null hypothesis; conversely, the *larger* the p -value, the *less* evidence we have to reject H_0 and so in this case we are more likely to retain the null hypothesis. The following table gives some guidelines on how to interpret your p -value:

p -value	Interpretation
p is bigger than 10%	no evidence against the null hypothesis: stick with H_0
p lies between 5% and 10%	<i>slight</i> evidence against H_0 , but not enough to reject it
p lies between 1% and 5%	moderate evidence against H_0 : reject it, and go with H_1
p is smaller than 1%	strong evidence against H_0 : reject it, and go with H_1

Table 2.1: Conventional interpretation of p -values

The above five steps are universal to all hypothesis testing.

2.3 Testing one mean

Here, from a **single** population, we draw a **single** sample, and we estimate the population mean μ with the sample mean $\bar{x}$. We’d then like to see how convincing a proposal (say c) for the population mean is, based on the information in our sample. As with the construction of confidence intervals, the choice of test statistic in step 3 above depends on whether or not the population variance (or standard deviation) is known.

2.3.1 Known population variance σ^2

If the population variance σ^2 is *known*, the steps involved in this hypothesis test are:

1. State the null hypothesis

$$H_0 : \mu = c;$$

that is, we believe the population mean μ to be equal to c .

2. State the alternative hypothesis

Here, we have three options: we can test whether or not the mean is *equal* to c ; *greater than* c ; or *less than* c . These would be written as

$$\begin{aligned} H_1 &: \mu \neq c; \\ H_1 &: \mu > c, \quad \text{or} \\ H_1 &: \mu < c \end{aligned}$$

(respectively). The first is alternative is known as a **two-tailed** alternative; the second and third are both called **one-tailed** alternatives.

3. Calculate the test statistic

Since, in this case, σ^2 is known, the test statistic is z , where

$$z = \frac{|\bar{x} - \mu|}{\sqrt{\sigma^2/n}},$$

which is just (the absolute value of) a realisation of $Z \sim N(0, 1)$.

4. Find the p -value of the test

We compare the test statistic found in step 3 to values from tables for the standard normal distribution (since Z follows a normal distribution). From table 2.2, we have the following values for a two-tailed test:

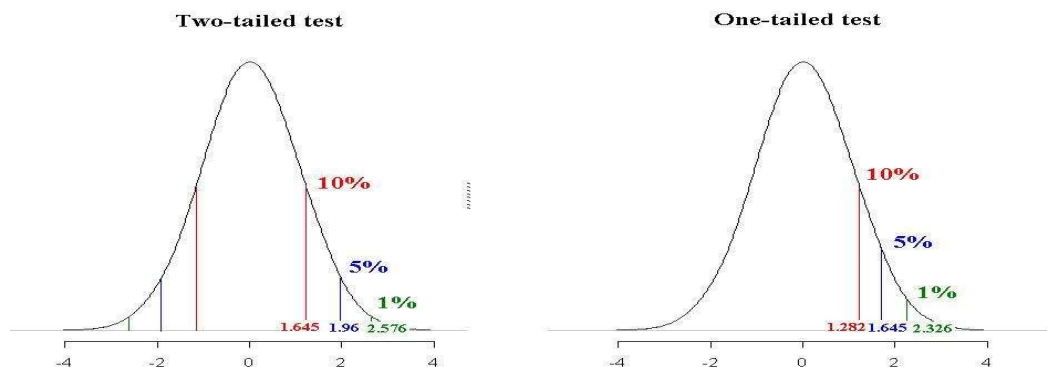
Significance level	10%	5%	1%
Critical value	1.645	1.96	2.576

Similarly, for a one-tailed test we have:

Significance level	10%	5%	1%
Critical value	1.282	1.645	2.326

By locating the position of our test statistic on one of the above tables (which one we use depends on whether we have constructed a one- or two-tailed alternative hypothesis), we can find the range within which our p -value lies. *[Note that using tables will not give an exact p -value, just a range for our p -value: using a computer package, such as Minitab, can give us an exact p -value]*

A picture might help:

**5. Conclusion**

Use table 2.1 to draw your conclusion, depending on the p -value obtained in step 4. Word your conclusions in the context of the original question.

2.3.2 Unknown population variance σ^2

As with the calculation of confidence intervals, things are a little more awkward when the population variance is unknown. Steps 1 and 2 remain unchanged, i.e.

1. State the null hypothesis, and
2. State the alternative hypothesis.

However, recall that the quantity

$$\frac{|\bar{x} - \mu|}{\sqrt{s^2/n}}$$

can no longer be regarded as (the absolute value of) a realisation from a standard normal distribution but instead is from a Student's t -distribution (on ν degrees of freedom). Thus, the remaining steps now are:

3. Calculate the test statistic

The test statistic now is

$$t = \frac{|\bar{x} - \mu|}{\sqrt{s^2/n}}.$$

Notice this is almost exactly the same as that used when we know the population variance; however, the population variance σ^2 has been replaced with the *sample* variance s^2 and the test statistic is called t instead of z to remind us that we are computing a realisation from a t -distribution rather than a normal distribution.

4. Find the p -value of the test

We compare the test statistic found in step 3 to t -distribution tables. We use table 2.3 to obtain critical values for certain levels of significance, and locate the position of our test statistic as before. However, since these critical values depend on the degrees of freedom, ν , we have to make sure we “read off” values from the correct row. For example, if our sample size was 12, then $\nu = 12 - 1 = 11$ and so, for a two-tailed test we’d get:

Significance level	10%	5%	1%
Critical value	1.796	2.201	3.106

Similarly, for a one-tailed test we’d have:

Significance level	10%	5%	1%
Critical value	1.363	1.796	2.718

As before, by locating the position of our test statistic on one of the above tables, we can find the range within which our p -value lies. Again, drawing a picture might help.

5. Conclusion

This bit's exactly the same as before – i.e. use table 2.1 to draw your conclusion, depending on the p -value obtained in the previous step. Don't forget to word your conclusions in the context of the original question!

2.4 Examples**Example 2.4.1**

A chain of shops believes that the average size of transactions is £130, and the variance is known to be £900. The takings of one branch were analysed and it was found that the mean transaction size was £123 over the 100 transactions in one day. Based on this sample, test the null hypothesis that the true mean is equal to £130.

Since σ^2 is known (we are given that $\sigma^2 = 900$), we use the approach of section 2.3.1.

Steps 1 and 2 (hypotheses)

Here, we state our null and alternative hypotheses. The null hypothesis is given in the question – i.e.

$$H_0 : \mu = £130.$$

We could test against a general alternative, i.e.

$$H_1 : \mu \neq £130.$$

Step 3 (calculating the test statistic)

Since σ^2 is known, the test statistic is

Step 4 (finding the p -value)

Since σ^2 is known, we use normal distribution tables (table 2.2) to obtain a range for our p -value. Our alternative hypothesis is two-tailed (i.e. $\neq$ rather than $<$ or $>$), and so our values are:

Significance level	10%	5%	1%
Critical value	1.645	1.96	2.576

Our test statistic $z = 2.33$ lies between the critical values of 1.96 and 2.576, and so our p -value lies between 1% and 5%.

Step 5 (conclusion)

Using table 2.1 to interpret our p -value, we see that there is moderate evidence against H_0 . Thus, we should reject H_0 in favour of the alternative hypothesis H_1 ; it appears that the population mean transaction size *is not equal to* £130.

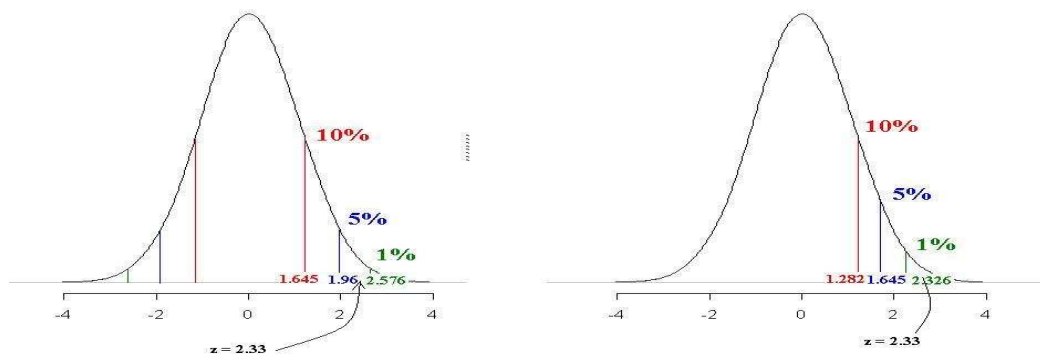
Alternatively, suppose we suspect that the proposed value of £130 is too high. We could have set up a one-tailed alternative hypothesis in step 2, i.e. we could have tested

$$\begin{aligned} H_0 &: \mu = £130 && \text{against} \\ H_1 &: \mu < £130. \end{aligned}$$

This is now a one-tailed test and the critical values from table 2.2 are

Significance level	10%	5%	1%
Critical value	1.282	1.645	2.326

The test statistic is (as before) 2.33, which now lies “to the right” of the last critical value in the table (2.326). Thus, our p -value is now smaller than 1%, and so, using table 2.1, we see that in this more specific test there is *strong* evidence against H_0 . Again, this can be seen more clearly with a diagram:



Notice that this one-tailed test is more specific than the two-tailed test previously carried out. If you're not sure whether you should perform a one-tailed test or a two-tailed test, (i.e. there might not be much difference between the proposed mean and the sample mean), it's usually safer to test against the more general two-tailed alternative.

Example 2.4.2

The batteries for a fire alarm system are required to last for 20000 hours before they need replacing. 16 batteries were tested; they were found to have an average life of 19500 hours and a standard deviation of 1200 hours. Perform a hypothesis test to see if the batteries do, on average, last for 20000 hours.

Steps 1 and 2 (hypotheses)

Suppose we suspect that the battery lifetimes are falling short of 20000 hours. Using a one-tailed test, our null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \mu = 20000 && \text{versus} \\ H_1 &: \mu < 20000. \end{aligned}$$

We use a one-tailed test because we are interested in whether the batteries are effective or not; they are only effective if they last at least 20000 hours.

Step 3 (calculating the test statistic)

Unlike the previous example, the population variance σ^2 is unknown (i.e. the question does not say “the population variance is ...”, or “the population standard deviation is ...”, for example). However, the *sample* standard deviation is given, based on a sample of size 16, and so the approach of section 2.3.2 can be used. Thus, the test statistic is given by

$$\begin{aligned} t &= \frac{|\bar{x} - \mu|}{\sqrt{s^2/n}} \\ &= \frac{|19500 - 20000|}{\sqrt{1200^2/16}} \\ &= \frac{500}{\sqrt{1440000/16}} \\ &= 1.667. \end{aligned}$$

Step 4 (finding the p -value)

Since σ^2 is unknown, we use t -distribution tables (table 2.3) to obtain a range for our p -value. The degrees of freedom, $\nu = n - 1 = 16 - 1 = 15$, and under a one-tailed test this gives the following critical values:

Significance level	10%	5%	1%
Critical value	1.341	1.753	2.602

Our test statistic of $t = 1.667$ lies between the critical values of 1.341 and 1.753, and so the corresponding p -value lies between 5% and 10%.

Step 5 (conclusion)

Using table 2.1 to interpret our p -value, we see that there is only *slight* evidence against the null hypothesis and certainly not enough grounds to reject it, so we retain H_0 . It appears that, on average, the batteries *do* last for 20000 hours.

2.5 Exercises

1. A machine for filling cans of Coke has a process variance of 400ml. A sample of 100 cans is taken and it is found that the average contents are 240ml. Is this consistent with the cans containing the stated weight of 250ml? Use a two-tailed test.
2. A company manufactures bolts. One production line produces bolts that are designed to be 100mm long. A sample of 50 bolts is taken and measured; they are found to have an average length of 97.5mm with variance 150mm. Are the bolts being made too short?
3. A company is in dispute with its work force. The workers claim that under a new flexitime system they are working longer than the standard 37.5 hour week. The time cards of 10 workers were selected at random and these showed the following hours worked:

35 40 45 41 36 37 39 38 42 32

Perform a hypothesis test to see if staff *are* working more than a standard week.

One-tailed test	10%	5%	2.5%	1%	0.5%
Two-tailed test	20%	10%	5%	2%	1%
Critical value	1.282	1.645	1.96	2.326	2.576

Table 2.2: Tabulated values of z for which $\Pr(Z > z) = p$ (1 tailed) and $\Pr(|Z| > z) = p$ (2 tailed) where Z has a standard normal distribution

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

Table 2.3: Tabulated values of t for which $\Pr(T > t) = p$ (1 tailed) and $\Pr(|T| > t) = p$ (2 tailed), where T has a t -distribution with ν degrees of freedom

Chapter 3

Hypothesis tests for two means

3.1 Introduction

Last week you were introduced to the concept of hypothesis testing in statistics, and we considered hypothesis tests for the mean if we have a single sample drawn from a single population. If we have *two* independent random samples from *two* populations, we can compare the two sample means in a test for two means (*c.f.* comparing one sample mean to a *proposed value* in the one-sample case). We use the same framework for hypothesis testing as for the one-sample tests:

1. State the **null hypothesis**, H_0 ;
2. State the **alternative hypothesis**, H_1 ;
3. Calculate a **test statistic**;
4. Find the **p -value**, and
5. Use table 2.1 to form your **conclusions**.

However, the calculations required for the test statistic in step 3 are slightly different.

3.2 Testing two means

Recall that, in the test for one mean, there were two cases: population variance (σ^2) known and population variance unknown. Similarly, when comparing two means, we can consider a test where both population variances are known and both are unknown.

3.2.1 Both population variances (σ_1^2 and σ_2^2) known

1. State the null hypothesis

This time, the null hypothesis is

$$H_0 : \mu_1 = \mu_2,$$

i.e. the two population means are equal.

2. State the alternative hypothesis

We usually test against the (two-tailed) alternative:

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

i.e. the population means *are not* equal. However, if we have reason to believe that one population mean is larger (or smaller!) than the other, we might want to use the (one-tailed) alternatives:

$$\begin{aligned} H_1 : \mu_1 &> \mu_2, & \text{or} \\ H_1 : \mu_1 &< \mu_2. \end{aligned}$$

3. Calculate the test statistic

The test statistic for a two-sample test (when both population variances are known) is

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

where $\bar{x}_1$, $\bar{x}_2$, n_1 and n_2 are the means and sample sizes of samples 1 and 2 (respectively), and σ_1^2 and σ_2^2 are the corresponding population variances.

4. Find the p -value of the test

As in the tests for one mean, we use statistical tables to obtain our p -value, or rather a *range* for our p -value. Since, in this case, both population variances are known, we refer to standard normal tables (i.e. table 2.2) as before, the critical values depending on whether we are testing against a one- or two-tailed alternative hypothesis.

5. Reach a conclusion

Use table 2.1 to form a conclusion (i.e. retain or reject the null hypothesis), remembering to also word your conclusions in plain English and in the context of the question posed.

Example

Before a training session for call centre employees a sample of 50 calls to the call centre had an average duration of 5 minutes, whereas after the training session a sample of 45 calls had an average duration of 4.5 minutes. The population variance is known to have been 1.5 minutes before the course and 2 minutes afterwards. Has the course been effective?

3.2.2 σ_1^2 and σ_2^2 unknown

In the more likely situation where the population variances are unknown, we use the test statistic

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

where s is a “pooled standard deviation”, and is found as

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

Also, as with the test for one mean when the population variance was unknown, we need to use t -tables (table 2.3) to obtain our p -value for the test. The degrees of freedom is now found as $\nu = n_1 + n_2 - 2$. Apart from these changes, the hypothesis test follows the same format as that for which both population variances are known.

Example

A company is interested in knowing if two branches have the same level of average transactions. The company sample a small number of transactions and calculates the following statistics:

$$\begin{array}{l|l} \text{Shop 1} & \bar{x}_1 = 130 \quad s_1^2 = 700 \quad n_1 = 12 \\ \text{Shop 2} & \bar{x}_2 = 120 \quad s_2^2 = 800 \quad n_2 = 15 \end{array}$$

Test whether or not the two branches have (on average) the same level of transactions.

Steps 1 and 2 (hypotheses)

Our null and alternative hypotheses are:

$$\begin{array}{ll} H_0 & : \mu_1 = \mu_2 \quad \text{versus} \\ H_1 & : \mu_1 \neq \mu_2. \end{array}$$

Step 3 (calculating the test statistic)

Since both population variances are unknown (only the *sample* values are given), the test statistic is

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}};$$

thus, we first need to obtain the pooled standard deviation s . This is given as

$$\begin{aligned}
 s &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \\
 &= \sqrt{\frac{(12 - 1) \times 700 + (15 - 1) \times 800}{12 + 15 - 2}} \\
 &= \sqrt{\frac{11 \times 700 + 14 \times 800}{25}} \\
 &= \sqrt{\frac{7700 + 11200}{25}} \\
 &= \sqrt{\frac{18900}{25}} \\
 &= \sqrt{756} \\
 &= 27.495.
 \end{aligned}$$

Thus,

$$\begin{aligned}
 t &= \frac{|130 - 120|}{27.495 \times \sqrt{\frac{1}{12} + \frac{1}{15}}} \\
 &= \frac{10}{27.495 \times \sqrt{0.15}} \\
 &= \frac{10}{10.649} \\
 &= 0.939.
 \end{aligned}$$

Step 4 (finding the p -value)

Since both population variances are unknown, we use t -tables to obtain our critical value. The degrees of freedom, $\nu = n_1 + n_2 - 2$, i.e. $\nu = 12 + 15 - 2 = 25$. Under a two-tailed test, and using table 2.3, we get the following critical values:

Significance level	10%	5%	1%
Critical value	1.708	2.060	2.787

Our test statistic $t = 0.939$ lies to the left of the first critical value, and so our p -value is bigger than 10%.

Step 5 (conclusion)

Using table 2.1, we see that, since our p -value is larger than 10%, we have no evidence to reject the null hypothesis. Thus, we retain H_0 and conclude that there is no significant difference between the average level of transactions at the two shops.

3.3 Exercises

1. A chain of record shops believes that its Northumberland Street store (Shop 1) is more successful than its Metro Centre branch (Shop 2). The management take a random sample of daily takings and obtains the following summary statistics:

$$\begin{array}{l|l} \text{Shop 1} & \bar{x}_1 = \text{£}15000 \quad s_1^2 = 400 \quad n_1 = 10 \\ \text{Shop 2} & \bar{x}_2 = \text{£}14250 \quad s_2^2 = 600 \quad n_2 = 20 \end{array}$$

Is the management's belief correct?

2. (a) An airline company, EasyAir, advertises 5 hour flights from Glasgow to Cairo. Another company flying the same route, RyanJet, suspects EasyAir of false advertising, and samples 20 of EasyAir's flights between Glasgow and Cairo. The mean flight time from this sample is 5 hours and 10 minutes with a standard deviation of 20 minutes. Are RyanJet's suspicions supported?
(b) RyanJet claim that their Glasgow to Cairo flights are, on average, faster than EasyAir's flights. A sample of 23 RyanJet flights between Glasgow and Cairo has a mean flight time of 5 hours and 4 minutes with a standard deviation of 22 minutes. Are RyanJet's flights on this route shorter than EasyAir's?
3. The mean age of students attending a college is 19.5 years. The ages of a sample of 7 students who have been offered college accommodation are given below. Test the hypothesis that younger students are favoured.

17.9 18.2 19.1 20.3 17.8 17.4 17.8

4. During the streaking phase of the 1970's, a psychological test designed to determine extroversion was applied to a group of 19 admitted male streakers and a control group of 19 male non-streakers. The results were

streakers	non-streakers
$\bar{x}_1 = 15.26$	$\bar{x}_2 = 13.90$
$s_1^2 = 2.62$	$s_2^2 = 4.11$

Are streakers more extrovert than non-streakers?

5. In the comparison of two kinds of paint, a consumer testing service finds that eight one-gallon cans of "Wilko's Best", and ten one-gallon cans of "Dulor", cover the following areas (in square feet):

"Wilko's Best"	542	546	550	548	540	537	548	553
"Dulor"	521	532	498	512	551	540	500	513 520 523

The population standard deviation coverage for "Wilko's Best" and "Dulor" is 31 square feet and 26 square feet respectively. Perform a test to find out whether or not there is any difference in mean coverage between the two brands.

One-tailed test	10%	5%	2.5%	1%	0.5%
Two-tailed test	20%	10%	5%	2%	1%
Critical value	1.282	1.645	1.96	2.326	2.576

Table 3.1: Tabulated values of z for which $\Pr(Z > z) = p$ (1 tailed) and $\Pr(|Z| > z) = p$ (2 tailed), where Z has a standard normal distribution

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

Table 3.2: Tabulated values of t for which $\Pr(T > t) = p$ (1 tailed) and $\Pr(|T| > t) = p$ (2 tailed), where T has a t -distribution with ν degrees of freedom

Chapter 4

Goodness-of-fit tests

4.1 Introduction

In this section, we address the question of whether our data follow any **pattern**, or fit a specified or assumed **probability distribution**. More precisely, we ask: Is there any evidence against the (null) hypothesis that the data are a random sample from a particular distribution, such as the Poisson distribution? The method first forms a frequency table of the data and then compares the observed frequencies with those that would be expected if the hypothesis was correct. If there is a large difference between observed and expected frequencies then this sheds doubt on the hypothesised distribution.

A simple example: traffic accidents

Consider the following data which is the number of traffic accidents involving children within two kilometres of schools, listed by days of the week.

Day	No. of accidents (Observed)
Monday	23
Tuesday	18
Wednesday	17
Thursday	19
Friday	23

We could use a hypothesis test from chapter 2 to find out, for example, if the mean number of accidents per day is equal to some specified value. However, in goodness-of-fit tests, we are more interested in the **distribution** of the data rather than just a single parameter. Looking at the data, there appears to be more accidents at the start and end of the week than during mid-week days; however, the above dataset is only a sample of all accidents that have taken place, and we need to decide whether or not this apparent effect is simply a result of sampling variation, or whether a pattern *really does exist* in the data.

Let's consider a hypothesis which states that no pattern exists in the data (i.e. there are no "peaks" in the data, and everything occurs **uniformly**). If this were true, we'd expect there to be (roughly) the same number of accidents on each day of the week. Thus, we would obtain the **expected frequencies**:

Day	No. of accidents (Expected)
Monday	20
Tuesday	20
Wednesday	20
Thursday	20
Friday	20

Although there *did* appear to be more accidents at the start of, and towards the end of, the week, comparing the expected frequencies in the above table with the real-life data (the **observed frequencies**), we see that they are fairly close. But how close do these observed and expected frequencies have to be before we conclude that traffic accidents occur uniformly throughout the week? To avoid subjectivity, we should compare these two sets of frequencies through a formal hypothesis test, known as a **chi-squared (χ^2) goodness-of-fit test**. The hypothesis tests we have looked at so far (tests for one mean and tests for two means) have compared a calculated test statistic to the standard normal distribution or the t -distribution; goodness-of-fit tests use the chi-squared (χ^2) distribution. This distribution, like the t -distribution, has one parameter called the degrees of freedom (ν); its critical values can be found in table 4.1.

4.2 Goodness-of-fit tests: basic framework

The procedure for carrying out a goodness-of-fit test is as follows:

1. State the null hypothesis (H_0)

If we think our frequencies occur uniformly, then, in the context of the last example (traffic accidents), this would be

H_0 : There are the same number of accidents each day of the week.

If we think our data might follow a particular distribution, then the null hypothesis might be, for example,

H_0 : The number of accidents follows a Poisson distribution.

You have already done some work on probability distributions for continuous and discrete data; we will review some of these distributions shortly.

2. State the alternative hypothesis (H_1)

This (as is often the case) is just the opposite statement to the null hypothesis; so in the context of goodness-of-fit tests, this might be

H_1 : There are *not* the same number of accidents each day of the week, or

H_1 : The number of accidents does *not* follow a Poisson distribution.

3. Calculate the test statistic

The test statistic is

$$X^2 = \sum \frac{(O - E)^2}{E},$$

where O and E represent the **observed** and **expected** frequencies respectively. This is most easily calculated by drawing up a table:

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
$\vdots$	$\vdots$	$\vdots$

and then summing the final column. **For the goodness-of-fit test to work, all expected frequencies must be ≥ 5 ; to achieve this, adjacent categories can be “pooled”.**

4. Find the p -value of the test

As before, a range for our p -value can be found by comparing our test statistic to statistical tables. The test statistic X^2 follows a χ^2 distribution for which critical values are given in table 4.1. As with the t -distribution, the row we use depends on the degrees of freedom, which can be calculated as

$$\nu = (\text{number of categories after pooling}) - (\text{number of parameters estimated}) - 1.$$

Once we have the 10%, 5% and 1% critical values, we can find out where our test statistic lies and so obtain a corresponding range for our p -value.

5. Reach a conclusion

This bit's the same as before – use table 2.1 to interpret your p -value. Remember, generally, to reject H_0 , we need a p -value of less than 5%. Don't forget to write a sentence in the context of the question, i.e. “the data appear to follow a Poisson distribution”.

Back to the traffic accidents example

Let us now test the hypothesis that the number of traffic accidents occurs uniformly throughout the week; i.e. on average, there are an equal number of traffic accidents each day.

Steps 1 and 2 (hypotheses)

Our hypotheses are

H_0 : There are the same number of accidents each day of the week versus

H_1 : There are *not* the same number of accidents each day of the week.

Step 3 (calculating the test statistic)

This is the hard bit! Remember, the test statistic is

$$X^2 = \sum \frac{(O - E)^2}{E}.$$

Drawing up a table usually helps! Thus,

Day	Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
Monday	23	20	0.45
Tuesday	18	20	0.2
Wednesday	17	20	0.45
Thursday	19	20	0.05
Friday	23	20	0.45

So,

$$\begin{aligned} X^2 &= 0.45 + 0.2 + 0.45 + 0.05 + 0.45 \\ &= 1.6. \end{aligned}$$

Notice we didn't have to pool any adjacent categories since all expected frequencies were ≥ 5 .

Step 4 (finding the p -value)

Using table 4.1, with degrees of freedom

$$\begin{aligned} \nu &= (\text{number of categories after pooling}) - (\text{number of parameters estimated}) - 1 \\ &= 5 - 0 - 1 \\ &= 4, \end{aligned}$$

we obtain the following critical values:

Significance level	10%	5%	1%
Critical value	7.78	9.49	13.28

In goodness-of-fit tests, there are no two-tailed tests, and so you don't have to worry about your choice!

Since our test statistic $X^2 = 1.6$ lies to the left of the first critical value, our p -value is bigger than 10%.

Step 5 (conclusion)

Using table 2.1 to interpret our p -value, we see that there is no evidence to reject the null hypothesis. Thus, we retain H_0 , and conclude that there is insufficient evidence to suggest that accidents do not occur uniformly, i.e. we conclude that the number of accidents is, on the whole, the same for each day of the week (this is because the observed and expected frequencies were so close).

If we are testing the null hypothesis that the data occur uniformly (as in this example), the expected frequencies are easy to calculate. However, things get a little more tricky if we want to test the hypothesis that the data adhere to a more complicated distribution. Before we look at this, the next section will quickly review some probability models introduced in semester 1.

4.3 Probability distributions

A **probability distribution** of a discrete random variable X is the list of all possible values X can take and the probabilities associated with them. For example, if the random variable X is the outcome of a roll of a (fair, six-sided) dice, then the probability distribution for X is

x	1	2	3	4	5	6
$\Pr(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

This is an example of a discrete probability distribution, since the random variable X can only take integer values. There are a number of ‘standard’ probability distributions which data often adopt; the two *discrete* probability distributions you studied in semester 1 are the **binomial distribution** and the **Poisson distribution**.

4.3.1 The binomial distribution

The binomial distribution is used to model the number of successes in a series of n independent trials. If each trial results in either a “success” or a “failure” and the probability that any particular trial is a “success” is p , then the random variable X = number of successes out of n trials has a **binomial distribution** with probabilities of the form

$$\Pr(X = r) = {}^nC_r \times p^r \times (1 - p)^{n-r},$$

where ${}^nC_r = \frac{n!}{r!(n-r)!}$ (or the nCr button on your calculator), and r takes any of the values $0, 1, \dots, n$ (remember that $r!$ is the factorial function, and $r! = r \times (r - 1) \times (r - 2) \times \dots \times 3 \times 2 \times 1$). The mean of a binomial distribution is given by $n \times p$, and the variance is given by $n \times p \times (1 - p)$.

The following are examples of an experiment containing a binomial random variable:

- (i) 100 students of equal ability sit an exam. The total number passing is recorded;

- (ii) A fair dice is thrown 20 times and the number of sixes obtained is recored;
- (iii) ten barley seeds are sown in a petri dish. The number of germinating seeds is recorded.

Why is the random variable $X = \text{number of rainy days in a week}$ unlikely to have a binomial distribution?

In the context of goodness-of-fit tests, we can use the the formula for calculating probabilities from a binomial distribution to calculate expected frequencies based on this distribution; **the expected frequency is just the sample size multiplied by the associated probability.**

4.3.2 The Poisson distribution

This distribution is used to model data which are counts of (random) events in a certain area or time interval, without a fixed upper limit. The probability distribution has one parameter – λ – and its probabilities take the form

$$\Pr(X = r) = \frac{\lambda^r e^{-\lambda}}{r!},$$

where r can take values $0, 1, 2, \dots$. The mean and variance of a Poisson distribution are both equal to λ .

The following are practical examples in which a Poisson distribution might be appropriate:

- (i) Number of radioactive emissions per unit of time;
- (ii) seedlings per unit area;
- (iii) knots per cubic foot of wood.

Again, in the context of goodness-of-fit tests, we can use the formula for calculating probabilities from a Poisson distribution to calculate expected frequencies based on this distribution.

4.4 A More Complex Example

Consider the following data:

Number of claims	Observed frequency (O)
0	144
1	91
2	32
3	11
4	2
5 +	0
	280

The data represents the number of small factories in northern England in which industrial injuries resulted in claims for compensation between April 2003 and March 2004 (Whitaker, L., *Biometrika*, 2005, **10**, 55). Clearly, here we are not expecting the same number of claims in each class – we would expect the number of factories to decrease as the number of claims increases. We could, however, use a probability distribution to represent the number of accidents we could expect. But which probability distribution might we use?

Well, the data are discrete (since we are looking at counts), and for discrete data we have looked at two probability distributions: the binomial distribution and the Poisson distribution. Recall the set-up for a binomial distribution to be appropriate: we need n trials, each with two outcomes (“success” and “failure”), and the probability of “success”, p , should be constant across all trials. In this example, it is difficult to see how we would use a binomial distribution; however, the Poisson distribution might be more appropriate. Remember that for a Poisson distribution to be valid, all we really need are data which are counts over a certain time interval, without a fixed upper limit. Thus, the Poisson distribution might be used here – we have counted the number of factories with 0 claims, 1 claim, 2 claims, etc., and there’s no natural fixed upper limit to the number of claims (the last class is just 5+).

Recall that the mean of a Poisson random variable is equal to the rate parameter λ . Thus, we can estimate λ by setting it equal to the sample mean, i.e.

$$\begin{aligned}
 \lambda &= \frac{0 \times 144 + 1 \times 91 + 2 \times 32 + 3 \times 11 + 4 \times 2}{280} \\
 &= \frac{196}{280} \\
 &= 0.7.
 \end{aligned}$$

Once we have this we can proceed as before; the expected probabilities based on the Poisson distribution will be calculated using the Poisson formula on the previous page, and the expected frequencies found by multiplying these probabilities by the sample size.

Steps 1 and 2 (*hypotheses*)

Since we think the Poisson distribution might be an appropriate model for our data, we test

H_0 : Claims follow a Poisson distribution against

H_1 : Claims do *not* follow a Poisson distribution.

Step 3 (*calculating the test statistic*)

Recall that, for goodness-of-fit tests, the test statistic is

$$X^2 = \sum \frac{(O - E)^2}{E}.$$

We already have the O 's – these are just the observed frequencies. What we need to calculate are the E 's (the expected frequencies). Unlike the traffic accidents example, we are now testing the hypothesis that the data follow a particular distribution – the Poisson distribution. Thus, we first of all use the formula for the Poisson distribution to obtain expected *probabilities*, and then multiply these by the the total number of accidents to obtain the expected frequencies. From earlier, we know that Poisson probabilities are found using

$$\Pr(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}.$$

We have estimated λ as 0.7; thus, we just need to substitute this into the formula to calculate our probabilities for different values of r . For example, the expected probability of no deaths is

$$\begin{aligned} \Pr(X = 0) &= \frac{e^{-0.7} \times 0.7^0}{0!} \\ &= 0.4966. \end{aligned}$$

Similarly,

$$\begin{aligned} \Pr(X = 1) &= \frac{e^{-0.7} \times 0.7^1}{1!} \\ &= 0.3476. \end{aligned}$$

We can do this for all our categories and then convert these to frequencies by multiplying by the total number of accidents we have observed (280). This gives

Number of claims	Expected probability	Expected frequency (E)
0	0.4966	139.048
1	0.3476	97.328
2	0.1217	34.076
3	0.0284	7.952
4	0.0050	1.4
5 +	0.0007	0.196
		280

For the χ^2 test to be valid, the expected frequencies must be at least 5; hence, we have to adjust our categories to allow this to happen. We can do this by “pooling” the last three categories, to give:

Number of claims	Observed frequency (O)	Expected frequency (E)
0	144	139.048
1	91	97.328
2	32	34.076
3+	13	9.548

Notice that the observed frequencies must also be “pooled” across the last three categories. The test statistic can then be found by calculating $\frac{(O-E)^2}{E}$ for each category, and then summing over all categories, i.e.

Number of claims	O	E	$\frac{(O-E)^2}{E}$
0	144	139.048	0.176
1	91	97.328	0.411
2	32	34.076	0.126
3+	13	9.548	1.248

Thus,

$$\begin{aligned}
 X^2 &= \sum \frac{(O - E)^2}{E} \\
 &= 0.176 + 0.411 + 0.126 + 1.248 \\
 &= 1.961.
 \end{aligned}$$

Step 4 (finding the p -value)

Recall that for goodness-of-fit tests, we use the χ^2 distribution to obtain our p -value. Thus, using table 4.1 with degrees of freedom

$$\begin{aligned}
 \nu &= (\text{number of categories after pooling}) - (\text{number of parameters estimated}) - 1 \\
 &= 4 - 1 - 1 \\
 &= 2,
 \end{aligned}$$

we obtain the following values:

Significance level	10%	5%	1%
Critical value	4.61	5.99	9.21

Our test statistic $X^2 = 1.961$ lies to the left of the first critical value, and so our p -value is bigger than 10%.

Step 5 (conclusion)

Using table 2.1 to interpret our p -value, we find that there is no evidence against the null hypothesis, and so we should retain H_0 . We can say that the observed values match those we would expect to see from a Poisson distribution with a value of $\lambda = 0.7$ quite well, and there is insufficient evidence to suggest that our data do not follow the proposed distribution.

4.5 Exercises

1. The following table contains data on the number of customers a shop has on each day of the week. Does the number of customers occur uniformly throughout the week?

Day	Frequency
Monday	95
Tuesday	81
Wednesday	84
Thursday	120
Friday	117
Saturday	110
Sunday	93
Total	700

2. The following table contains data on the number of complaints received per day at a major retail bank's branches.

Number of Complaints	Frequency
0	270
1	140
2	65
3	14
4 +	5
Total	494

Propose an appropriate distribution for these data and test to see if it is consistent with the data.

- 3*. The Elves Toy Co. makes toy trains. For quality control purposes, toy trains coming off the production line are regularly inspected for defects; defective trains are always thrown away. Once every week, five trains are randomly selected from the production line. The probability that a train has a defect and is thrown away is 0.25. The table below shows the number of weeks in which 0, 1, ..., 5 trains were defective and thus thrown away in 2007.

No. of trains thrown away	No. of weeks
0	10
1	21
2	14
3	6
4	1
5	0

Propose an appropriate probability distribution for the number of defective trains, and test to see if it is consistent with the data observed.

* Prize question

		p				
		50%	10%	5%	1%	0.1%
	1	0.45	2.17	3.84	6.63	10.83
	2	1.39	4.61	5.99	9.21	13.82
	3	2.37	6.25	7.82	11.34	16.27
	4	3.36	7.78	9.49	13.28	18.47
	5	4.34	9.24	11.07	15.09	20.52
	6	5.35	10.64	12.59	16.81	22.46
	7	6.35	12.02	14.07	18.48	24.32
ν	8	7.34	13.36	15.51	20.09	26.13
	9	8.34	14.68	16.92	21.67	27.88
	10	9.34	15.99	18.31	23.21	29.59
	12	11.34	18.55	21.03	26.22	32.91
	15	14.34	22.31	25.00	30.58	37.70
	20	19.34	28.41	31.41	37.57	45.32
	25	24.34	34.38	37.65	44.31	52.62
	30	29.34	40.26	43.77	50.89	59.70

Table 4.1: This table contains values of x for which $\Pr(X^2 > x) = p$, where X^2 has a χ^2 -distribution with ν degrees of freedom

Chapter 5

Tests of independence using the χ^2 distribution

5.1 Introduction

In the previous chapter, we used a χ^2 test to test hypotheses about the probability distribution of a random variable. We are now going to see that the test is much more versatile because it can be used to compare the probability distributions of more than one population, and that this is equivalent to a test of the **association** between two **categorical variables**.

Suppose we are interested in the probabilities that a graduate has a permanent job, temporary job, or is unemployed six months after graduation, and we want to compare these probabilities for female graduates with those for male graduates. As we now have *two* categorical variables (employment status and gender), it will be worth looking at different ways of presenting and storing categorical data.

5.2 Presenting categorical data

The following table shows the first few rows of a computer file of data for a sample of graduates in **raw form**, that is, with a row for each graduate. The first column merely identifies the graduate, the second is their employment status six months after graduation (0 = permanent job, 1 = temporary job, 2 = unemployed) and the third is their gender (0 = male, 1 = female).

Case number	Employment status	Gender
1	2	0
2	0	0
3	1	1
4	0	1
5	1	1
6	0	0
7	0	1
$\vdots$	$\vdots$	$\vdots$

Most software will store data in this form, although sometimes the categories don't have to be coded as numbers and you can enter the category name (e.g. male/female or permanent/temporary/unemployed) directly.

As employment status and gender are categorical with three and two categories respectively, only 3×2 possible pairs of numbers can appear in the second and third columns. The data can therefore be consolidated into **frequency form** using a column for each categorical variable and another for counts or frequencies, as shown:

Employment status	Gender	Frequency
0	0	100
0	1	90
1	0	33
1	1	40
2	0	25
2	1	22

The most succinct way of representing categorical data is to arrange the frequencies in a **contingency table**. When there are only two categories, this is easy; each column represents a category of one categorical variable, and each row represents a category of the other categorical variable. The frequencies are written in the appropriate cell of the table. Notice that each cell of the contingency table corresponds to one of the frequencies of the table above.

	Permanent	Temporary	Unemployed
Male	100	33	25
Female	90	40	22

It is often useful to include the column and row totals, so we repeat the table with these below:

	Permanent	Temporary	Unemployed	Total
Male	100	33	25	158
Female	90	40	22	152
Total	190	73	47	310

5.3 The χ^2 test for independence

In a χ^2 test for independence, we ask: “Is there any evidence in the data of an **association** between two categorical variables”? As before, the **observed frequencies** are compared with **expected frequencies** calculated under a null hypothesis. In this case, the null hypothesis is that of independence between the variables.

5.3.1 Tests of independence: basic framework

As always, we now outline the basic framework for our hypothesis test. The steps to consider in a χ^2 test of independence are exactly the same as for the other hypothesis tests considered in this course so far, i.e.

1. State the null hypothesis (H_0)

The null hypothesis is that there is no association between the two categorical variables, or that the two categorical variables are independent; i.e.

$$\begin{aligned} H_0 &: \text{There is no association between the two categorical variables,} && \text{or} \\ H_0 &: \text{The two categorical variables are independent.} \end{aligned}$$

2. State the alternative hypothesis (H_1)

This is just the opposite statement to the null hypothesis:

$$\begin{aligned} H_1 &: \text{There is an association between the two categorical variables,} && \text{or} \\ H_1 &: \text{The two categorical variables are not independent.} \end{aligned}$$

3. Calculate the test statistic

The test statistic is

$$X^2 = \sum \frac{(O - E)^2}{E},$$

where O and E represent the **observed** and **expected** frequencies (respectively). This is most easily calculated by drawing up a table:

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
$\vdots$	$\vdots$	$\vdots$

and then summing the final column. Remember, the observed frequencies are the frequencies given in the table; *you* have to calculate the expected frequencies yourself! In the example concerning graduate employment status and gender, if (as the null hypothesis specifies) the two variables *are* independent, then, for example,

$$\text{Pr(Female and Temporary)} = \text{Pr(Female)} \times \text{Pr(Temporary)}.$$

Thus, the expected *frequency* for Female and Temporary is given by

$$\begin{aligned} & \text{Pr(Female and Temporary)} \times \text{overall sample size} \\ &= \frac{\text{row total for Female}}{\text{overall sample size}} \times \frac{\text{column total for Temporary}}{\text{overall sample size}} \times \text{overall sample size} \\ &= \frac{\text{row total for Female} \times \text{column total for Temporary}}{\text{overall sample size}} \\ &= \frac{152 \times 73}{310} \\ &= 35.794. \end{aligned}$$

Generally, for an $r \times c$ contingency table with r rows and c columns, expected frequencies are calculated using

$$E = \frac{\text{row total} \times \text{column total}}{\text{overall sample size}}.$$

As with the χ^2 goodness-of-fit test, the χ^2 test for independence is only valid if all $E > 5$. If necessary, adjacent row or columns can be pooled to ensure this.

4. Find the p -value for the test

A range for the p -value for a test of independence can be found in tabulated values from the χ^2 distribution (table 4.1). The degrees of freedom is found as

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1).$$

5. Reach a conclusion

This bit's the same as ever – use table 2.1 to interpret your p -value, and either retain (stick with) or reject the null hypothesis. Remember to always write a sentence in the context of the question, i.e. “the two categorical variables do not appear to be independent”.

5.3.2 Example: Employment status and gender

We now perform a χ^2 test of independence to assess if there is any evidence of a relationship between employment status of graduates six months after graduation and gender.

Steps 1 and 2 (*hypotheses*)

Since we are testing for a relationship between employment status and gender, the null hypothesis for this test is

$$H_0 : \text{Employment status and gender are independent,}$$

i.e. there is no association between employment status and gender (i.e. it doesn't matter whether you're male or female, you have an equal chance of being in a permanent job, a temporary job, or unemployed six months after graduation!).

The alternative is simply

$$H_1 : \text{Employment status and gender are *not* independent,}$$

i.e. there *is* an association between employment status and gender (i.e. males and females do *not* have an equal chance of being in a permanent job, a temporary job, or unemployed six months after graduating!).

Step 3 (*calculating the test statistic*)

Recall that, for any χ^2 test, the test statistic is

$$X^2 = \sum \frac{(O - E)^2}{E},$$

where O and E are the observed and expected frequencies (respectively). We already have the observed frequencies:

	Permanent	Temporary	Unemployed	Total
Male	100	33	25	158
Female	90	40	22	152
Total	190	73	47	310

To calculate the test statistic, we also need the expected frequencies! Remember that, if the null hypothesis is true, then

$$E = \frac{\text{row total} \times \text{column total}}{\text{overall sample size}}.$$

Thus, for “cell 1” (Male and Permanent), we have

$$\begin{aligned} E_1 &= \frac{158 \times 190}{310} \\ &= \frac{30020}{310} \\ &= 96.839. \end{aligned}$$

For “cell 2” (Male and Temporary), we have

$$\begin{aligned} E_2 &= \frac{158 \times 73}{310} \\ &= \frac{11534}{310} \\ &= 37.206. \end{aligned}$$

Similarly, for “cell 3” (Male and Unemployed), we have

$$\begin{aligned} E_3 &= \frac{158 \times 47}{310} \\ &= \frac{7426}{310} \\ &= 23.955. \end{aligned}$$

The corresponding frequencies for the other cells are shown in the table below:

	Permanent	Temporary	Unemployed	Total
Male	96.839	37.206	23.955	158
Female	93.161	35.794	23.045	152
Total	190	73	47	310

Notice that none of the expected frequencies are ≤ 5 , and so we can proceed straight to the calculation of the test statistic without having to pool any categories. Thus, we have

O	E	$\frac{(O-E)^2}{E}$
100	96.839	0.103
33	37.206	0.475
25	23.955	0.046
90	93.161	0.107
40	35.794	0.494
22	23.045	0.047
		1.272

and so our test statistic is $X^2 = 1.272$.

Step 4 (*finding the p -value for the test*)

As with goodness-of-fit tests, we refer to χ^2 tables (table 4.1) to obtain a range for our p -value. However, in tests for independence, the degrees of freedom is given by

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1).$$

In our example we have two rows and three columns, and so

$$\begin{aligned}\nu &= (2 - 1) \times (3 - 1) \\ &= 1 \times 2 \\ &= 2.\end{aligned}$$

Notice that the degrees of freedom are *not* one less than the total number of cells in the table! Referring to table 4.1 (in chapter 4), we obtain the following critical values for 2 degrees of freedom:

Significance level	10%	5%	1%
Critical value	4.61	5.99	9.21

Our test statistic $X^2 = 1.272$ lies to the left of the first critical value; thus our p -value is bigger than 10%.

Step 5 (*form your conclusion*)

Since our p -value is larger than 10%, there is insufficient evidence to reject the null hypothesis. It appears that the two categorical variables (employment status and gender) *are* independent; i.e., we can conclude that employment status and gender are not associated, and so whether you are male or female, you have the same chance of being in a permanent job, a temporary job, or unemployed six months after graduation! (Of course, this conclusion is drawn from the information in this particular sample.)

5.3.3 Another example (Past exam question)

In a market research survey 200 people are shown a proposed design for the new Mini Cooper car, and they are asked if they like the design. The responses, broken down by age groups, are shown in the table below.

	Under 21	21–35	Over 35	Total
Liked design	24	38	70	132
Disliked design	35	17	16	68
Total	59	55	86	200

Is there any evidence to suggest that age is associated with attitude to the proposed design?

Solution

Steps 1 and 2 (*hypotheses*)

Step 3 (*test statistic*)

Recall that the test statistic is:

$$X^2 = \sum \frac{(O - E)^2}{E}.$$

We have the O 's – these are just the **O**bserved frequencies. We need the “ E ”'s as well! Remember, these are given by:

$$E = \frac{\text{row total} \times \text{column total}}{\text{overall sample size}}.$$

For “cell 1” (Liked design and Under 21), we get:

Step 4 (*p-value*)

Our degrees of freedom to use Table 4.1 is given by

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

$$=$$
$$=$$

Referring to Table 4.1, we thus get:

<i>p</i> -value	10%	5%	1%
Critical value			

Step 5 (*Conclusions*)

5.4 Exercises

- Two groups of students were given I.Q. tests: group 1 consisted of 30 students classified by their peers as “drinkers”, and group 2 consisted of 28 students classified by their peers as “non-drinkers”. The number of students with a “below norm”, “norm” and “above norm” I.Q. were counted, and the results summarised in the contingency table:

	Below norm	Norm	Above norm	Total
Drinkers	12	10	8	30
Non-drinkers	9	10	9	28
Total	21	20	17	58

Is there evidence of an association (or lack of independence) between the two factors?

- A large retail chain is been taken to an industrial tribunal by one of its female staff for sexual discrimination in its employment strategy. Does the company have a discriminatory employment strategy?

	Male	Female	Total
Shelf Stackers	34	50	84
Checkout Staff	39	40	79
Buyers	12	8	20
Section Managers	10	1	11
Total	95	99	194

Chapter 6

Using Minitab

6.1 Introduction

Minitab is an easy-to-use statistical package which can carry out a wide variety of statistical tasks. You may use Minitab in many different courses during your time as a student, for many different purposes, so it is worth putting a little effort into familiarising yourself with the basics at an early stage. That way, you will be able to adapt to more sophisticated uses of the package later on in your course. There are three basic kinds of object Minitab works with:

- **Data column**

This is the most frequently used object type in Minitab. Columns are denoted C1, C2, C3, etc. They can be used to store a collection of observations relating to a particular variable.

- **Constants**

These are denoted K1, K2, K3, etc, and each stores a single number of interest.

- **Data arrays**

These are matrices of numbers, and are denoted by M1, M2, M3, etc.

All of the data columns, constants and data arrays relating to a particular problem are stored in a working environment called a **worksheet**. Worksheets can be saved and loaded from disk, for later use. Several worksheets may be opened simultaneously, and these can be saved together as a **project**. Minitab has both a command language and a menu driven interface. This module will concentrate on the latter, which is more intuitive, especially for the beginner.

To load Minitab, login to a PC in any University computing cluster. Select Start – Programs – Statistical Software – Minitab 16 for Windows – Minitab 16.

You should now have a spreadsheet (“data window”) ready to input data. In Minitab, there are two main windows; the Session window and the Worksheet window. The Worksheet allows you to view and edit the data columns of the current worksheet. It is normally empty on startup, so the first step is to load the data in. Always enter data in the white boxes – the grey boxes are for column titles. Use the arrow keys to move around the worksheet.

6.2 Hypothesis tests for the mean

6.2.1 Tests for one mean

Recall the hypothesis tests for one mean (section 2.3). Here, from a single population, we draw a single sample, and we estimate the population mean μ with $\bar{x}$. We'd then like to see how convincing a proposal for the population mean is, based on the information in our sample. Our null hypothesis is

$$H_0 : \mu = c.$$

The alternative could be

$$\begin{aligned} H_1 &: \mu \neq c, \\ H_1 &: \mu > c \quad \text{or} \\ H_1 &: \mu < c. \end{aligned}$$

The test statistic we calculate, and the statistical tables we consult to obtain our p -value, depend on whether or not the population variance is known. If the population variance (σ^2) is *known*, we use the test statistic

$$z = \frac{|\bar{x} - \mu|}{\sqrt{\sigma^2/n}}.$$

If the population variance is unknown, we use

$$t = \frac{|\bar{x} - \mu|}{\sqrt{s^2/n}},$$

where s^2 is the *sample* variance.

If the population variance is known, the test statistic is called z ; another name for this hypothesis test is therefore the **one-sample z test**. If the population variance is unknown, the test statistic is called t , and we consult tables of probabilities for the t -distribution to obtain our critical value; another name for this hypothesis test is the **one-sample t test**.

Test for one mean in Minitab

This example is the same as that in question 3 of the exercises in chapter 2. A company is in dispute with its workforce. The workers claim that under a new flexitime system they are working longer than the standard 37.5 hours per week. The time cards of 10 workers were selected at random and these showed the following hours worked:

35 40 45 41 36 37 39 38 42 32

Are the staff working more than a standard week?

We can do this in Minitab! Simply enter the data in column 1 of the worksheet, and then select Stat – Basic Statistics – 1-Sample t (we are performing a one sample t test because we are testing for one mean and the population variance is unknown). A box will appear like that in figure 6.1. Entering C1 in the box labelled Variables, and 37.5 in the Test mean box will test the null hypothesis against the general two-tailed alternative; since we want to know if the workers are working *longer* than the standard week, we need to test the null hypothesis

$$H_0 : \mu = 37.5 \quad \text{against}$$

$$H_1 : \mu > 37.5.$$

To do this, select Options, and change the Alternative to greater than. Clicking OK twice gives the following output.

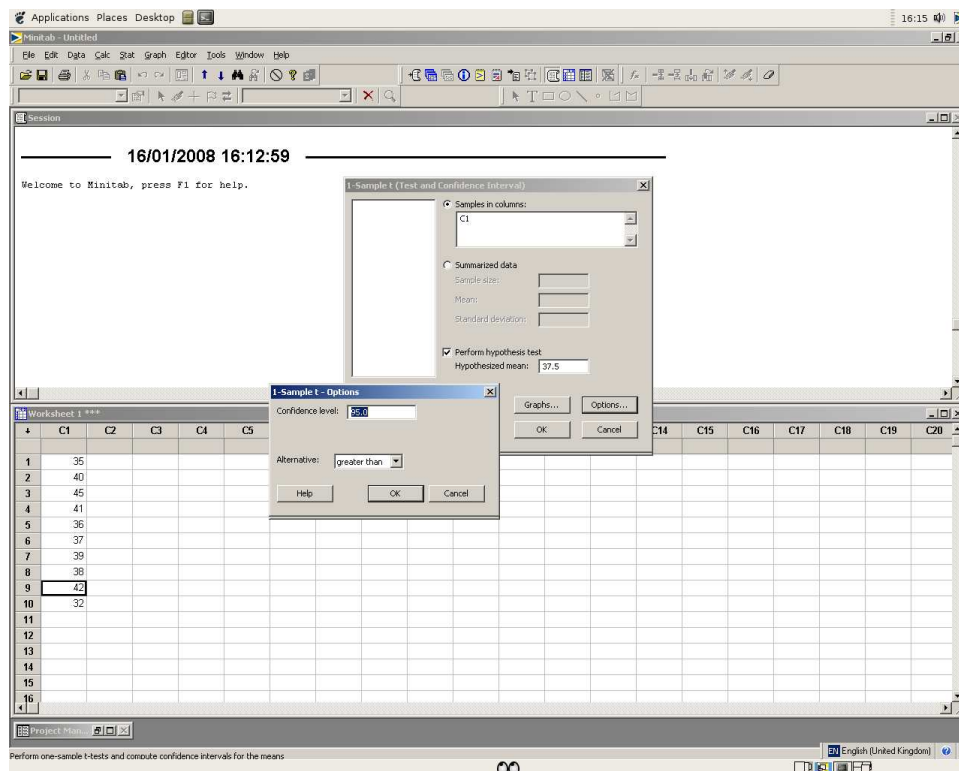


Figure 6.1: Screenshot showing the one-sample t test in Minitab

One-Sample T: Hours worked

Test of mu = 37.5 vs mu > 37.5

Variable	N	Mean	StDev	SE Mean	95% Lower Bound	T	P
Hours worked	10	38.5	3.75	1.19	36.33	0.84	0.210

Notice the benefit of using a computer package such as Minitab over performing the test by hand: instead of just obtaining a *range* for our p -value, we get the *exact* value for p . When the question above was done by hand a few weeks ago, we knew only that the p -value was bigger than 10% – Minitab gives us the exact p -value of 21%. [Oh, and another benefit is that we don't have to use a calculator and a pen and paper!!]

6.2.2 Tests for two means

If we have **two** independent random samples from **two** populations, we can compare the two sample means. The null hypothesis for such a test is

$$H_0 : \mu_1 = \mu_2.$$

The alternative hypothesis could be one of

$$\begin{aligned} H_1 &: \mu_1 \neq \mu_2, \\ H_1 &: \mu_1 > \mu_2 \quad \text{or} \\ H_1 &: \mu_1 < \mu_2, \end{aligned}$$

depending on the question. The formula we use for the test statistic will depend on whether or not both population variances are known. In the case where both *are* known, we use

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

If both are unknown, we use

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

where s is the “pooled standard deviation” (see section 3.2.2). Thus, similar to the one-sample case, we have a **two-sample z test** if both population variances are known, and a **two-sample t test** if both are unknown.

Test for two means in Minitab

The test for two means in Minitab is similar to that for one mean; however, this time we have two columns of data! For example, if we compare the average flight times (in minutes) of two flight companies between Newcastle and Palma, we might observe the following data:

EasyJet	150	147	141	158	155	133	140
Britannia	162	158	163	152	156	149	150

To test for a general difference in the flight times between the two companies, we test the null hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{against}$$

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

We enter these data in columns 1 and 2 of a fresh Minitab worksheet, and select Stat – Basic Statistics – 2-Sample t. We select Samples in different columns and enter C1 in Column 1 and C2 in Column 2. We tick Assume equal variances (always tick this), and then click OK. Figure 6.2 shows the options in Minitab.

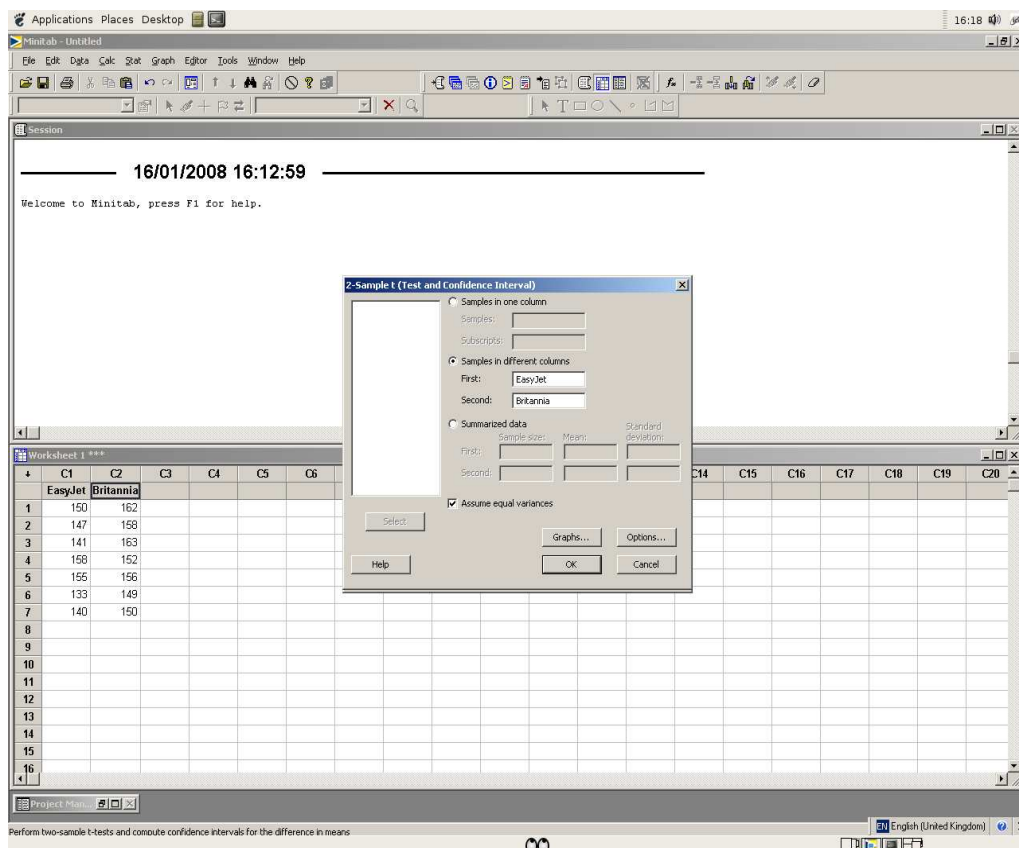


Figure 6.2: Screenshot showing the two-sample t test in Minitab

Two-Sample T-Test and CI: EasyJet, Britannia

Two-sample T for EasyJet vs Britannia

	N	Mean	StDev	SE Mean
EasyJet	7	146.29	8.86	3.4
Britannia	7	155.71	5.62	2.1

Difference = mu(EasyJet) - mu(Britannia)

Estimate for difference: -9.3

95% CI for difference: (-18.07, -0.79)

T-test of difference = 0 (vs not = 0): T-value = -2.38 P-value = 0.035

DF = 12

Both use Pooled StDev = 7.42

Since the p -value is 0.035, or 3.5%, it lies between 1% and 5% and so we have moderate evidence to reject the null hypothesis in favour of the alternative. It appears the two flight companies *do* have different flight times on this route (and Britannia's look longer).

6.3 Tests of independence using the χ^2 distribution

We can also perform test of independence in Minitab (see chapter 5 of these notes). Going back to the employment status and gender example, we have the following contingency table of observed frequencies:

	Permanent	Temporary	Unemployed	Total
Male	100	33	25	158
Female	90	40	22	152
Total	190	73	47	310

The aim here was to test for an association between the two categorical variables. The null and alternative hypotheses are

$$H_0 : \text{There is no association between employment status and gender} \quad \text{versus}$$

$$H_1 : \text{There is an association between employment status and gender.}$$

We simply enter the table the way it is into empty columns of a Minitab worksheet (excluding column totals), as shown in figure 6.3. Selecting Stat – Tables – Chi-Square Test will calculate all the expected frequencies and perform the test for us!

From the Minitab output, we see that our p -value is 0.529, or 52.9%. Thus, we retain the null hypothesis; it appears that there is no association between employment status and gender (again, exactly the same as when we performed this test by hand; compare Minitab's calculations to those on pages 45–47).

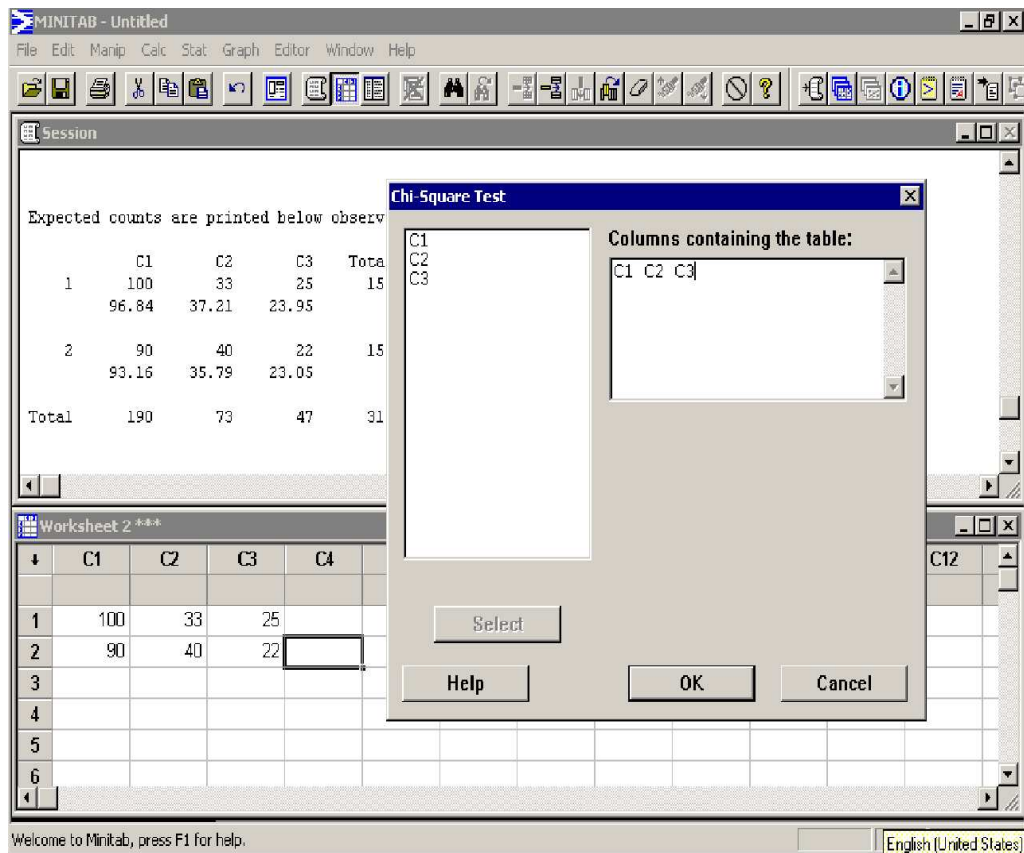


Figure 6.3: Screenshot showing the chi-squared test in Minitab

Chi-Square Test: C1,C2,C3

Expected counts are printed below observed counts

	C1	C2	C3	Total
1	100	30	25	158
	96.84	37.21	23.95	
2	90	40	22	152
	93.16	35.74	23.05	
Total	190	73	47	310

$$\text{Chi-Sq} = 0.103 + 0.476 + 0.046 + 0.107 + 0.494 + 0.047 = 1.273$$

$$\text{DF} = 2, \text{ P-Value} = 0.529$$

Chapter 7

Correlation and linear regression

7.1 Introduction

This chapter concerns itself with relationships between *continuous variables*, such as height and weight, market value and number of transactions, or temperature and sales of ice cream. How would you expect the three pairs of variables listed here to relate to one another? You have already seen (in semester 1) how to present paired data through a **scatter diagram**; this descriptive analysis is now supplemented with more formal techniques. The data consist of n pairs of observations on two variables X and Y :

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

These data could have arisen from a random sample of n individuals from a population, or from an experiment in which one variable, usually the X variable, is held fixed or controlled at certain chosen levels and independent measurements of the **response** variable, conventionally Y , are taken at each of these levels. The first step is *always* to plot the data on a scatter diagram.

Example: ice cream sales

Consider the following data (shown overleaf) for ice cream sales at Luigi Minchella's ice cream parlour. Is there any relationship between average temperature and ice cream sales? How would you *describe* this relationship?

Figure 7.1 shows a scatter plot of these data, drawn in Minitab. Looking at this diagram, we can immediately see that the two variables are related; in fact, it appears that, as the average temperature increases, sales also increase. We say such a relationship is a **positive** relationship. If there was a “downwards” or “downhill” slope to the scatter diagram, we would say the relationship was **negative**. It looks like we could also draw a straight line through the middle of the data without the points straying too much from this line. Thus, we can also say that the two variables have a **linear** relationship. The more the points stray from this line, the weaker the (linear) relationship between the two variables. So average temperatures and ice cream sales have a *strong, positive, linear* relationship.

Month	Average Temp (°C)	Sales (£ 000's)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80

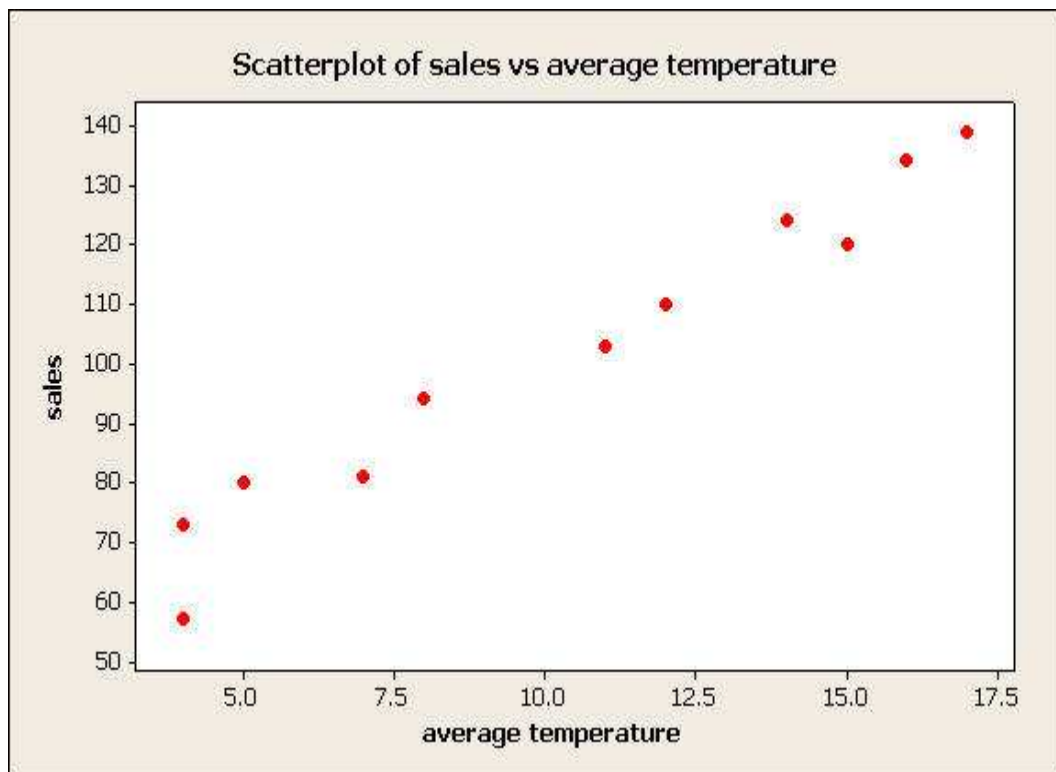


Figure 7.1: Scatter plot showing the relationship between average temperature and ice cream sales

7.2 Correlation

The only measure of association we will be covering here is the **Pearson product moment correlation coefficient**, r . Sometimes r is more briefly referred to as the **sample linear correlation coefficient**. The formula for r is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

$$\begin{aligned} S_{XY} &= \left(\sum xy \right) - n\bar{x}\bar{y}, \\ S_{XX} &= \left(\sum x^2 \right) - n\bar{x}^2, \\ S_{YY} &= \left(\sum y^2 \right) - n\bar{y}^2. \end{aligned}$$

The sample linear correlation coefficient r always lies between -1 and $+1$. Also, if r is near $+1$, there is a strong *positive* linear relationship between the two variables; if r is near -1 there is a strong *negative* relationship. If r is near zero, there is *no* linear relationship between the variables. Note that this does not imply no relationship at all, simply no *linear* relationship.

Example: ice cream sales

The easiest way to calculate the correlation coefficient between two variables (other than using a computer!) is to draw up a table:

x	y	x^2	y^2	xy
4	73	16	5329	292
4	57	16	3249	228
7	81	49	6561	567
8	94	64	8836	752
12	110	144	12100	1320
15	124	225	15376	1860
16	134	256	17956	2144
17	139	289	19321	2363
14	124	196	15376	1736
11	103	121	10609	1133
7	81	49	6561	567
5	80	25	6400	400
120	1200	1450	127674	13362

We have a sample size of 12 (not 24!), and so $n = 12$. Thus, using the sums of the first two columns we can find the sample means of temperature (X) and ice cream sales (Y):

$$\begin{aligned} \bar{x} &= \frac{120}{12} \\ &= 10 \quad \text{and} \end{aligned}$$

$$\begin{aligned} \bar{y} &= \frac{1200}{12} \\ &= 100. \end{aligned}$$

Similarly,

$$\begin{aligned}
 S_{XY} &= \left(\sum xy \right) - n\bar{x}\bar{y} \\
 &= 13362 - 12 \times 10 \times 100 \\
 &= 13362 - 12000 \\
 &= 1362,
 \end{aligned}$$

$$\begin{aligned}
 S_{XX} &= \left(\sum x^2 \right) - n\bar{x}^2 \\
 &= 1450 - 12 \times 10 \times 10 \\
 &= 1450 - 1200 \\
 &= 250 \quad \text{and}
 \end{aligned}$$

$$\begin{aligned}
 S_{YY} &= \left(\sum y^2 \right) - n\bar{y}^2 \\
 &= 127674 - 12 \times 100 \times 100 \\
 &= 127674 - 120000 \\
 &= 7674.
 \end{aligned}$$

Thus,

$$\begin{aligned}
 r &= \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} \\
 &= \frac{1362}{\sqrt{250 \times 7674}} \\
 &= \frac{1362}{1385.099274} \\
 &= 0.983 \text{ (to 3 decimal places).}
 \end{aligned}$$

We have a correlation coefficient of 0.983. Remember, this implies that there is a strong, positive (linear) relationship between average temperature and ice cream sales since the correlation coefficient is very close to +1. This is what we might expect, given the pattern shown in figure 7.1.

Remember, if your calculated correlation coefficient does not lie between -1 and $+1$ then you've done something wrong! You should also check to see if your calculated coefficient agrees with what you can see in the scatter diagram – an “uphill” slope ties in with a positive correlation coefficient, and a “downhill” slope with a negative correlation coefficient. If there appears to be just a random scatter of points, you might expect to get a correlation coefficient which is close to zero. The closer to a straight line the points will lie, then the closer to either $+1$ or -1 the correlation coefficient should be.

Data sets which show non-linear association can be analysed by the technique described above after transforming the data to linearity, or by using **rank correlation** methods which are outside the scope of this course.

7.3 Simple linear regression

A correlation analysis may establish a linear relationship but does not allow us to *use* it to say, predict the value of one variable given the value of another. **Regression analysis** allows us to do this and more. It is also applicable when one of the variables (X) is controlled.

We will assume that the scatter plot of Y versus X shows roughly a straight line and, in addition, that the **spread** in the Y -direction is roughly constant with X .

Look at the scatter plot of ice cream sales against average temperatures. A “line of best fit” can be drawn through the data, and from this line we can make predictions of ice cream sales based on temperature for temperatures which we have no data for. The problem is, everyone’s line of best fit is bound to be slightly different! And so everyone’s predictions will be slightly different! The aim of regression analysis is to find the very best line which goes through the data in a less subjective way. We do this through the **regression equation**. Before you can understand a regression equation, you need to know what a straight line equation actually is.

7.3.1 Equation of a straight line

Consider the equation

$$y = 2x.$$

We can draw the line which has this equation quite easily by drawing up a table:

x	0	1	2	3	4	5
$y = 2x$	0	2	4	6	8	10

In the space below, plot the line which has the equation $y = 2x$.

You probably noticed that we only needed two points to draw the line; this is true for *all* straight line equations. In the space below, plot the line which has the equation $y = 3 + 4x$.

There might be many equations which produce lines which pass through the ice cream data; however, in regression analysis, we want to find the equation which gives the “best” line, i.e. the equation which gives a straight line lying closer to the data than any other equation.

7.3.2 The regression equation

We assume a simple linear regression model for the data (and for the population from which the data have been drawn) in which

$$Y = \alpha + \beta x + \epsilon$$

where Y is the **response** variable, X is the **explanatory** variable, and ϵ (“epsilon”) is a random error with zero mean and constant variance. The unknown parameters α (“alpha”) and β (“beta”) represent the intercept and slope of the population regression line $\alpha + \beta x$. Obviously, we need to find α and β ; the best values will minimise the gaps between the regression line and the data. These “gaps” are known as the **residuals**. The values of α and β which give rise to the “best” regression line, i.e. the line which minimises the residuals, are

$$\begin{aligned}\hat{\beta} &= \frac{S_{XY}}{S_{XX}} \quad \text{and} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x},\end{aligned}$$

where S_{XY} and S_{XX} are as before. The “hats” on α and β are there to remind ourselves that we have *estimated* α and β using our sample data; these estimates will change from sample to sample. Since the error term (ϵ) is assumed to have zero mean, in practice we don’t estimate this and just ignore it in any further analysis.

Example: ice cream sales

We now use simple linear regression to fit a regression line through the ice cream sales data. The equation of the regression line is

$$y = \hat{\alpha} + \hat{\beta}x,$$

where we can find $\hat{\alpha}$ and $\hat{\beta}$ using

$$\begin{aligned}\hat{\beta} &= \frac{S_{XY}}{S_{XX}} \quad \text{and} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}.\end{aligned}$$

Notice that $\hat{\beta}$ must be found first, since we need $\hat{\beta}$ in order to calculate $\hat{\alpha}$. Thus,

$$\begin{aligned}\hat{\beta} &= \frac{1362}{250} \\ &= 5.448 \quad \text{and} \\ \hat{\alpha} &= 100 - 5.448 \times 10 \\ &= 100 - 54.48 \\ &= 45.52.\end{aligned}$$

Thus, the regression equation is

$$y = 45.52 + 5.448x$$

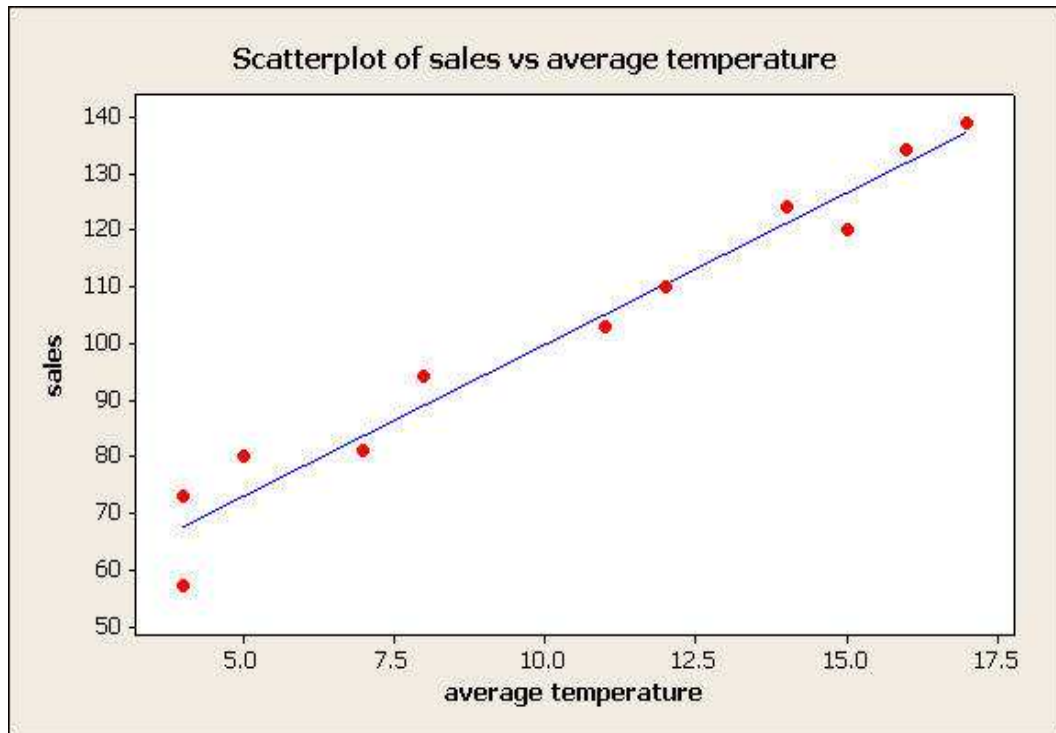


Figure 7.2: Scatter plot showing the relationship between average temperature and ice cream sales, with superimposed regression line

The scatter plot in figure 7.2 has this regression line superimposed. We can use this regression equation to predict ice cream sales for a given temperature. For example, if we were interested in knowing what the sales would be if the monthly average temperature was 10°C , we can either

- (1) take a reading from the graph, or
- (2) substitute 10 into our regression equation and solve for y .

The second approach is probably better, since the accuracy of a reading from your graph depends on how accurately you have drawn your graph in the first place. Thus, using our regression equation, the number of sales we can expect if the average temperature is 10°C is

$$\begin{aligned}
 y &= 45.52 + 5.448 \times 10 \\
 &= 45.52 + 54.48 \\
 &= 100,
 \end{aligned}$$

i.e. £100,000.

You should only use your regression line to make predictions within the range of the observed data; we cannot be certain that an association between the two variables will continue in the future, and even if it does, it might not be linear.

7.4 Extensions

Regression is a very flexible tool and there are a range of extensions that can be investigated; for example, non-linear regression for situations where a linear relationship between the two variables might not be justified. Multiple regression can be used when there is more than one predictor variable. There are also techniques for using regression on ordinal data, which are particularly useful when considering survey results. The calculations for these more complex regressions are beyond the scope of this course, and are invariably performed by a computer.

7.5 Exercises

1. Consider the following data for a company's monthly advertising expenditure and their sales.
 - (a) Produce a scatter plot for these data, and comment on the relationship between advertising and sales.
 - (b) Calculate the sample correlation coefficient. Does this agree with what you can see in your plot in part (a)?
 - (c) Perform a linear regression analysis on these data, and obtain the linear regression equation.
 - (d) Plot the regression line on your scatter diagram in part (a).
 - (e) If the company were to spend £112,000 on advertising in a month, what could we expect their sales to be?

Month	Advertising (£000's)	Sales (£ Millions)
January	100	11.2
February	90	12.1
March	110	13.2
April	120	15.1
May	115	14.2
June	95	10.2
July	105	12.5
August	130	16.6
September	118	14.8
October	100	10.8
November	115	11.2
December	128	15.9

- 2*. (a) In this chapter we have looked at the **Pearson Product Moment Correlation Coefficient**; however, this is just one of many correlation coefficients we could use to quantify the strength of linear dependence between two variables. Write down the name of one other correlation coefficient, give its formula, and explain how it differs to the Pearson Product Moment Correlation Coefficient.
- (b) What regression models could be used for data which do not display a linear association?

* Prize question

Chapter 8

Time Series and Forecasting

8.1 Introduction

A **time series** is a collection of observations made sequentially in time. When observations are made continuously, the time series is said to be **continuous**; when observations are taken only at specific time points, the time series is said to be **discrete**. In this course we consider only discrete time series, where the observations are taken at equal intervals. The observations $x_1, x_2, \dots, x_n$ should be regarded as the realisation of random quantities $X_1, X_2, \dots, X_n$.

The first step in the analysis of time series is usually to plot the data against time, in a **time series plot**.

8.1.1 Time series plots in Minitab

Once the data has been entered in the Minitab worksheet (or loaded from a data file), you can plot the data, or analyse it, using one of Minitab's built-in time series analysis functions. Selecting `Graph - Time Series Plot - Simple`, and then entering the column which contains your data in `Series` will produce a simple plot of your data against time. However, if the `Time/Scale` option is chosen, you can select the period over which your data has been collected (e.g. hourly, quarterly, monthly); you can also enter the start year (e.g. 1970) and the start month (e.g. 7 for July) to produce better labels for your time axis. Using the `Labels` option will also allow you to include a relevant title for your time series plot. Figure 8.1 (left-hand-side) shows a time series plot of monthly average air temperatures in England and Wales between 1970 and 1978 (inclusive). The right-hand-side plot in the same figure shows the same time series plot, but with the time axis labelled correctly (using the `Time/Scale` option in Minitab). The plot on the right-hand-side also has a title added.

Figures 8.2, 8.3 and 8.4 show time series plots of some other datasets; figure 8.2 shows a time series plot of the monthly means of daily relative sunspot numbers, figure 8.3 shows a time series plot of the (four-monthly) sales of a department store (between 1994 and 1997), and figure 8.4 shows the total precipitation in New York City each year from 1870 to 1968.

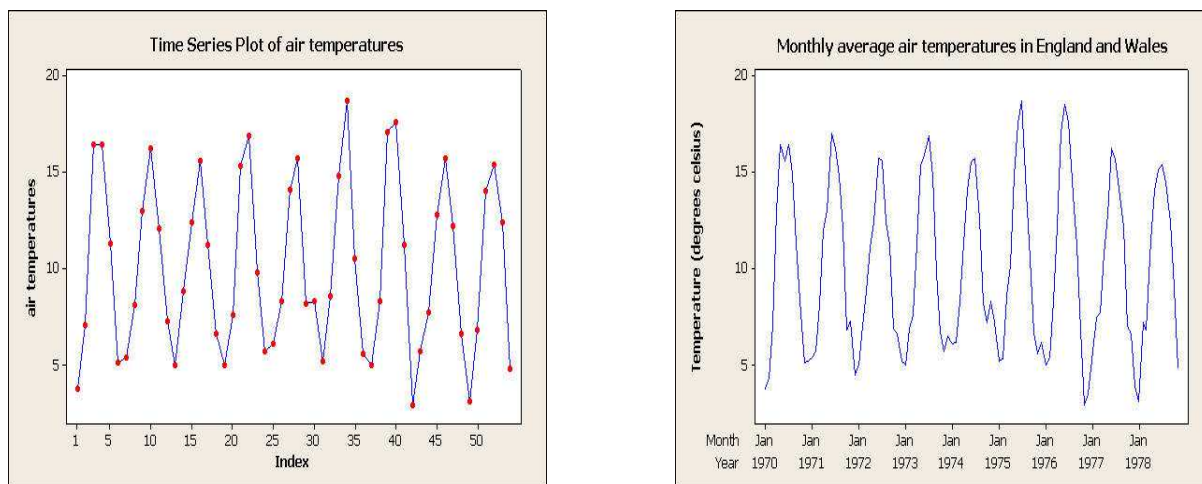


Figure 8.1: Left: Time series plots of monthly average air temperatures in England and Wales (right-hand-side plot shows edited time axis and inserted title)

8.1.2 Example: sales figures for a department store

Suppose we have the following four-monthly sales figures (in thousands of pounds) for a department store:

	Jan–Apr	May–Aug	Sep–Dec
1994	8	13	10
1995	10	14	11
1996	10	15	11
1997	11	16	13

Figure 8.3 shows a time series plot of these data; from this plot, we can see that the sales figures clearly exhibit seasonal patterns, with sales peaking in the second “season” (May–August) and dipping either side of this. We can also see that, overall, sales have increased through time, with 1997 showing the highest sales figures for all seasons. Thus, there is also a clear increasing trend in the series. The next two sections will demonstrate how both the trend and seasonal patterns can be estimated.

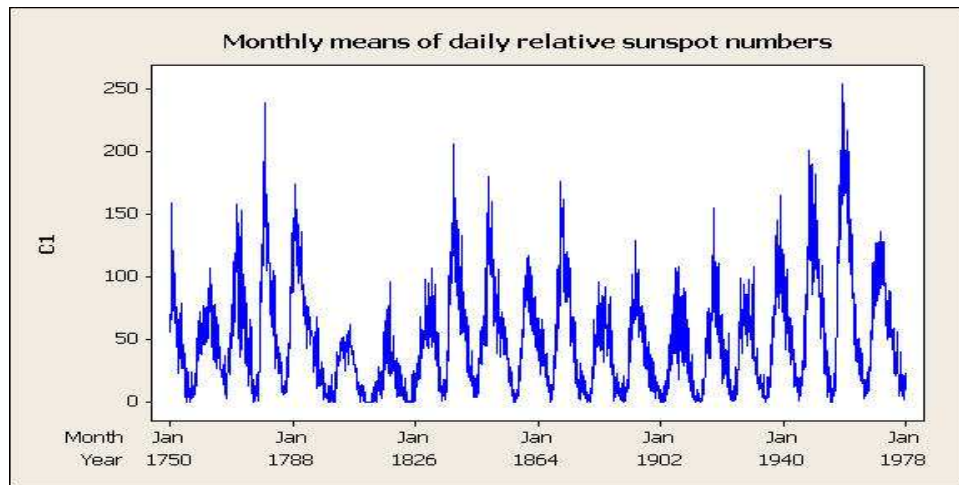


Figure 8.2: Time series plot of the monthly means of daily relative sunspot numbers



Figure 8.3: Time series plot of four-monthly sales for a department store

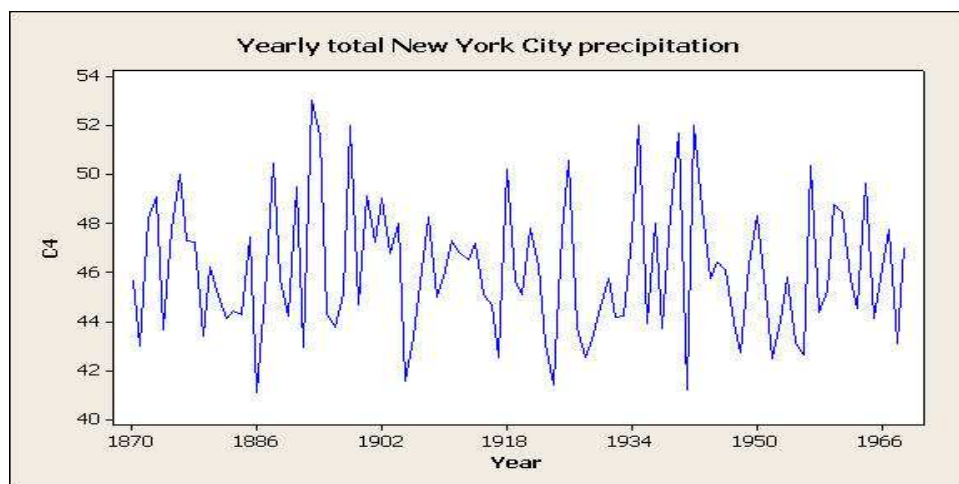


Figure 8.4: Time series plot of yearly total precipitation in New York

8.2 Isolating the trend

There are several methods we could use for isolating the trend. The method we will study is based on the notion of **moving averages**. To calculate a moving average, we simply average over the cycle around an observation. For example, for the department store sales data, we have three “seasons” (Jan–Apr, May–Aug and Sep–Dec) and so a full cycle consists of three observations. Thus, to calculate the first moving average we would take the first three values of the time series and calculate their mean, i.e.

$$\frac{8 + 13 + 10}{3} = 10.33.$$

Similarly, the second moving average is

$$\frac{13 + 10 + 10}{3} = 11.$$

The rest of the moving averages have been calculated in this way, and are shown in the table below.

	Moving averages		
	Jan–Apr	May–Aug	Sep–Dec
1994	*	10.33	11.00
1995	11.33	11.67	11.67
1996	12.00	12.00	12.33
1997	12.67	13.33	*

Obviously, there’s no moving average associated with the first and last data points, as there’s no observation before the first, or after the last, in order to calculate the moving average at these points! The length of the cycle over which to average is often obvious; for example, much data is presented quarterly or monthly, and that can provide a natural cycle around which to base the process. In our example, we have three clearly defined “seasons”, and so a cycle of length 3 would seem like the obvious choice. Figure 8.5 shows a plot of the original data with the moving averages overlaid. This clearly shows the upward rise in sales over the four years.

These moving averages represent the trend in our time series. We might be interested only in this trend, and wish to project this trend into the future – in our example, we might want to forecast sales for Jan–Apr 1998, or maybe even further ahead. Note the linearity in the plot of our moving averages in figure 8.5; in chapter 7 you were introduced to a technique known as **linear regression**, which calculates the equation of the line of best fit for a scatter plot of paired data. If we consider our units of time to be the explanatory variable, and if a linear relationship exists between our time variable and the data (which it seems to in our example), we could use linear regression techniques to estimate a straight line trend in our data. This straight line trend could then be extended into the future to make forecasts of sales at future time points.

Recall from chapter 6 that the simple linear regression model is given by

$$Y = \alpha + \beta X + \epsilon,$$

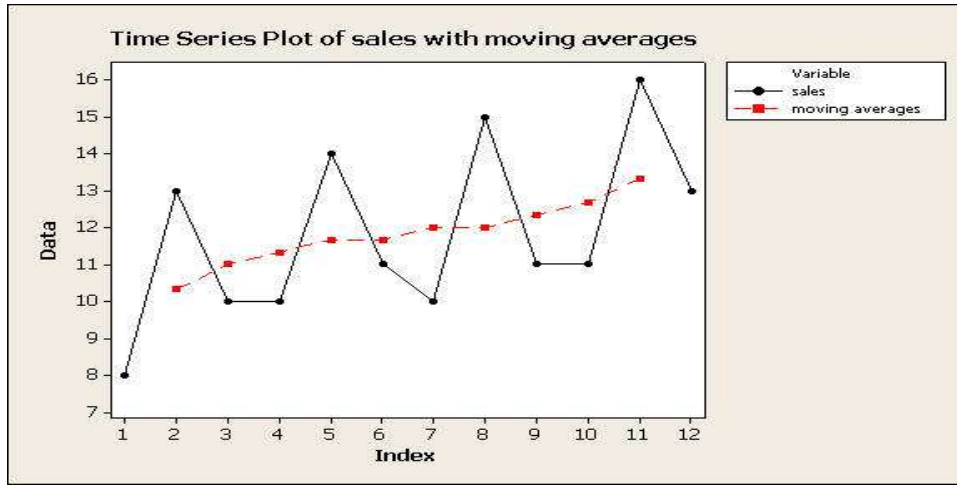


Figure 8.5: Time series plot of sales, with overlaid moving averages

where Y is the **response** variable, X is the **explanatory** variable, and ϵ is a random error. We now reformulate this model so that

$$Y = \alpha + \beta T + \epsilon,$$

where T is the “time point”. The first observation in the series has a time point of 1 (i.e. $T = 1$), the second has a time point of $T = 2$, and so on. So if we wanted to predict sales in the period Jan–Apr 1998, we would substitute $T = 13$ into the regression equation, since the last observed time point was for Sep–Dec 1997 and was the 12th time point. But, before we do this, we need to estimate α and β !

The formulae for α and β are exactly the same as in chapter 7, but now X is replaced with T , i.e.

$$\begin{aligned}\hat{\beta} &= \frac{S_{TY}}{S_{TT}} \quad \text{and} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{t},\end{aligned}$$

where

$$\begin{aligned}S_{TY} &= \left(\sum ty \right) - n\bar{t}\bar{y} \quad \text{and} \\ S_{TT} &= \left(\sum t^2 \right) - n\bar{t}^2.\end{aligned}$$

Remember, the easiest way to calculate these quantities is to draw up a table! note that for y we use the moving average values, *not* the actual observations! Notice also that we don’t have any observations at time points 1 and 12, since we were unable to calculate moving averages here.

t	y (moving averages)	ty	t^2
2	10.33	20.66	4
3	11.00	33.00	9
4	11.33	45.32	16
5	11.67	58.35	25
6	11.67	70.02	36
7	12.00	84.00	49
8	12.00	96.00	64
9	12.33	110.97	81
10	12.67	126.7	100
11	13.33	146.63	121
65	118.33	791.65	505

Thus,

$$\begin{aligned}\bar{t} &= \\ &= \end{aligned}$$

$$\begin{aligned}\bar{y} &= \\ &= \end{aligned}$$

Similarly,

$$\begin{aligned}S_{TY} &= \left(\sum ty \right) - n\bar{t}\bar{y} \\ &= \\ &= \end{aligned}$$

$$\begin{aligned}S_{TT} &= \left(\sum t^2 \right) - n\bar{t}^2 \\ &= \\ &= \end{aligned}$$

Thus,

$$\hat{\beta} = \frac{S_{TY}}{S_{TT}}$$

$$=$$

$$=$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t}$$

$$=$$

$$=$$

So, the regression equation for our trend is

$$y =$$

8.3 Isolating the seasonal effects

We have seen how we can use simple averaging over the cycle around an observation to isolate the trend in our series, and how to estimate this trend using the linear regression techniques discussed in chapter 7. We can use the linear trend equation to make forecasts of future sales, but these forecasted values will assume sales increase linearly only (since the regression equation is an equation of a straight line). Going back to the time series plot of our data (figures 8.3 and 8.5), we can see that the underlying level of the sales data does increase in a linear fashion, but there are clear cycles around this increase. In this section we will discuss how to estimate these so-called **seasonal effects**.

First of all, we calculate the **seasonal deviations** by subtracting the moving average for each observation from the original observation. Again, we cannot calculate values for the first and last observations. For example, the seasonal deviation for May–Aug 1994 is found as

$$13 - 10.33 = 2.67.$$

Similarly, for Sep–Dec 1994, we have

$$10.00 - 11.00 = -1.$$

The other seasonal deviations, along with the **seasonal means**, are shown in the table below:

	seasonal deviations		
	Jan–Apr	May–Aug	Sep–Dec
1994	*	2.67	–1
1995	–1.33	2.33	–0.67
1996	–2	3	–1.33
1997	–1.67	2.67	*
means	–1.67	2.6675	–1

We can now calculate the **seasonal effects**. The seasonal effect for each season is the seasonal mean for that season minus the overall mean. The overall mean from the table above is found as

$$\begin{aligned} \frac{2.67 - 1 - 1.33 + \dots + 2.67}{10} &= \frac{2.67}{10} \\ &= 0.267. \end{aligned}$$

Thus, the seasonal effects for each season are

$$\begin{aligned} \hat{s}_1 &= -1.67 - 0.267 \\ &= -1.937 \end{aligned}$$

$$\begin{aligned} \hat{s}_2 &= 2.6675 - 0.267 \\ &= 2.4005 \end{aligned}$$

$$\begin{aligned} \hat{s}_3 &= -1 - 0.267 \\ &= -1.267. \end{aligned}$$

Note that the seasonal effects should add up to give zero. Ours don't – we have

$$\begin{aligned}\hat{s}_1 + \hat{s}_2 + \hat{s}_3 &= -1.937 + 2.4005 - 1.267 \\ &= -0.8035.\end{aligned}$$

Thus, we have to make an adjustment so they *do* add up to give zero. To do this, we find the mean of our seasonal effects, and then subtract this from each of the seasonal effects. In this example, the mean of our seasonal effects is

$$\begin{aligned}\frac{-1.937 + 2.4005 - 1.267}{3} &= \frac{-0.8035}{3} \\ &= -0.26783.\end{aligned}$$

Thus, if we subtract this from each of the seasonal effects, the *adjusted* seasonal effects will then add up to give zero. Thus, the *adjusted* seasonal effects are

$$\begin{aligned}\hat{s}_1 &= -1.937 - (-0.26783) \\ &= -1.66917\end{aligned}$$

$$\begin{aligned}\hat{s}_2 &= 2.4005 - (-0.26783) \\ &= 2.66833\end{aligned}$$

$$\begin{aligned}\hat{s}_3 &= -1.267 - (-0.26783) \\ &= -0.99917.\end{aligned}$$

Just be careful with double negatives! Now the seasonal effects *do* sum to give zero!

8.4 Forecasting

There are many ways in which we can forecast future observations. One way is to use the linear regression equation for the trend in our series. For the department store sales data, recall that this was

$$y = 10.043 + 0.275t.$$

To predict average sales in Jan–Apr 1998, we would substitute $t = 13$ into the above equation, since this would be our 13th observation. Doing so, gives

$$\begin{aligned}y &= 10.043 + 0.275 \times 13 \\ &= 10.043 + 3.575 \\ &= 13.618.\end{aligned}$$

However, we're not quite done yet! This assumes that our data follow a straight line; looking at the time series plot in figure 8.3, clear cycles around an increasing trend can be seen. Thus, we now need to **add** in our seasonal effect. The seasonal effect for Jan–Apr is $\hat{s}_1 = -1.66917$. Thus, our “full” forecast for sales in Jan–Apr 1998 is

$$13.618 + (-1.66917) = 11.949,$$

i.e. £11,949, or just under £12,000.

8.5 Exercises

Consider the following data for sales (in thousands of pounds) at a newly opened sandwich shop on a large industrial estate. The sandwich shop only opens five days a week as many of the factories shut at weekends.

	Mon	Tues	Wed	Thurs	Fri
Week beginning 21/02/05	28	16	23	44	54
Week beginning 28/02/05	31	22	27	51	60
Week beginning 7/03/05	35	25	30	51	65

- Produce a time series plot of the sales data, either by hand or using *Minitab*. Remember to label your axes and give your plot a suitable title.
- Comment on your time series plot in part (a).
- Calculate the moving averages for the sales data, based on a five-observation cycle (Hint: Your first moving average will be for Wednesday 23/02/05, and is $(28 + 16 + 23 + 44 + 54)/5 = 33$).
- Using the moving averages obtained in part (c), find the linear trend

$$y = \hat{\alpha} + \hat{\beta}t.$$

- Calculate the seasonal deviations by subtracting the moving average for each observation from the original observation, and calculate the seasonal means.
- Calculate the overall mean, and hence calculate the seasonal effects for Monday, ..., Friday (Remember, if your seasonal effects don't sum to give zero, you need to make a suitable adjustment).
- Use the regression equation obtained in part (d), and the seasonal effects in part (f), to forecast sales at the sandwich shop on Monday 14th March 2005 and Tuesday 15th March 2005.

Chapter 9

Linear programming (I)

9.1 Introduction

Decision making is a process that has to be carried out in many areas of life. Usually there is a particular aim in making one decision rather than another. Two particular aims that are often considered in commerce are maximising profit and minimising loss.

During, and after, the Second World War a group of American mathematicians developed some mathematical methods to help with decision making. They sought to produce mathematical models of situations in which all the requirements, constraints and objectives were expressed as algebraic equations. They then developed methods for obtaining the **optimal solution** – the maximum or minimum value of a required function.

In this chapter we will study problems for which all the algebraic expressions are linear, that is, of the form

$$(\text{a number})x + (\text{a number})y;$$

For example,

$$\text{Profit} = 4x + 3y$$

is a linear equation in x and y , and

$$16x + 18y \leq 25$$

is a linear *inequation* or *inequality* in x and y . This area of mathematics is called **linear programming**.

Linear programming methods are some of the most widely used method employed to solve management and economic problems. They have been applied in a variety of contexts, some of which will be discussed in this chapter, with enormous savings in money and resources.

9.2 Formulating linear programming problems

The first step in formulating a linear programming problem is to determine which quantities you need to know to solve the problem. These are called the **decision variables**.

The second step is to decide what the **constraints** are in the problem. For example, there may be a limit on resources or a maximum or minimum value a decision variable may take, or there could be a relationship between two decision variables.

The third step is to determine the objective to be achieved. This is the quantity to be maximised or minimised, that is *optimised*. The function of the decision variables that is to be *optimised* is called the **objective function**.

The examples which follow illustrate the varied nature of problems that can be modelled by a linear programming model. We will not, at this stage, attempt to solve these problems but instead concentrate on producing the objective function and the constraints, writing these in terms of the decision variables. As an aid to this it is often useful to summarise all the given information in the form of a table (as illustrated in example 9.2.1).

9.2.1 Example (A chair manufacturer)

A manufacturer makes two kinds of chairs, A and B, each of which has to be processed in two departments, I and II. Chair A has to be processed in department I for 3 hours and in department II for 2 hours. Chair B has to be processed in department I for 3 hours and in department II for 4 hours.

The time available in department I in any given month is 120 hours, and the time available in department II, in the same month, is 150 hours.

Chair A has a selling price of £10 and chair B has a selling price of £12.

The manufacturer wishes to maximise his income. How many of each chair should be made in order to achieve this objective? You may assume that all chairs made can be sold.

In this section, **we will not attempt to solve this problem**; we will simply *formulate* the situation as a linear programming problem. You'll notice that there's a lot of information given in the question – this is typical of a linear programming problem. Sometimes it's easier to summarise the information given in a table:

Type of chair	Time in dept. I (hours)	Time in dept. II (hours)	Selling price (£)
A	3	2	10
B	3	4	12
Total time available	120	150	

To formulate this linear programming problem, we consider the following three steps:

1. What are the **decision variables**? (i.e. which quantities do you need to know in order to solve the problem?)
2. What are the **constraints**?
3. What is the **objective**?

Step 1: Decision variables

To find out what the decision variables are, read through the question and identify the things you'd like to know in order to solve the problem. You can usually do this by going straight to the last sentence of the question. The last sentence here is

“How many of each chair should be made...”

Thus, we'd like to know the number of type A chairs to make, and the number of type B chairs to make. These are our decision variables, and are usually denoted with lower case letters (x and y if we have two decision variables, x , y and z if we have three, for example). Thus, our decision variables are

$$\begin{aligned} x &= \text{number of type A chairs made} & \text{and} \\ y &= \text{number of type B chairs made.} \end{aligned}$$

Step 2: Constraints

Identifying the constraints is probably the hardest bit. To understand this bit, consider what could happen in each department. For example, if we focus on what could happen in department I:

Since: the production of 1 type A chair uses 3 hours,
 then: the production of x type A chairs takes $3 \times x = 3x$ hours.
 Similarly: the production of 1 type B chair uses 3 hours,
 so: the production of y type B chairs takes $3 \times y = 3y$ hours.

The total time used is therefore

$$(3x + 3y) \text{ hours.}$$

Since only 120 hours are available in department I, one constraint is

$$\begin{aligned} (3x + 3y) \text{ hours} &\leq 120 \text{ hours,} & \text{or just} \\ (3x + 3y) &\leq 120. \end{aligned}$$

Considering department II in a similar way, we get:

Since: the production of 1 type A chair uses 2 hours,
 then: the production of x type A chairs takes $2 \times x = 2x$ hours.
 Similarly: the production of 1 type B chair uses 4 hours,
 so: the production of y type B chairs takes $4 \times y = 4y$ hours.

The total time used is therefore

$$(2x + 4y) \text{ hours.}$$

Since only 150 hours are available in department II, a second constraint is

$$\begin{aligned} (2x + 4y) \text{ hours} &\leq 150 \text{ hours,} && \text{or just} \\ (2x + 4y) &\leq 150. \end{aligned}$$

In addition to the two constraints we have identified above, we also require that x and y are non-negative (since we can't make a negative number of chairs!), and so we also have the following two constraints:

$$\begin{aligned} x &\geq 0 && \text{and} \\ y &\geq 0. \end{aligned}$$

These are called the **non-negativity constraints**.

Step 3: Objective function

Our objective here is to maximise income. If we make x type A chairs, then we get $\pounds 10 \times x = \pounds 10x$, since each type A chair sells for $\pounds 10$.

Similarly, if we make y type B chairs, then we get $\pounds 12 \times y = \pounds 12y$, since each type B chair sells for $\pounds 12$.

The total income is then

$$\pounds Z = \pounds(10x + 12y).$$

The aim is to maximise income, so we'd like to maximise

$$Z = 10x + 12y,$$

where z is the objective function.

Thus, to summarise, we have the following linear programming problem:

Maximise $Z = 10x + 12y$ subject to the constraints

$$\begin{aligned} 3x + 3y &\leq 120, \\ 2x + 4y &\leq 150, \\ x &\geq 0 && \text{and} \\ y &\geq 0. \end{aligned}$$

9.2.2 Example (A book publisher)

A book publisher is planning to produce a book in two different bindings: paperback and library. Each book goes through a sewing process and a gluing process. The table below gives the time required, in minutes, for each process and for each of the bindings:

	Sewing (mins)	Gluing (mins)
Paperback	2	4
Library	3	10

The sewing process is available for 7 hours per day and the gluing process for 15 hours per day. The profits are 25p on a paperback edition and 60p on a library edition. How many books in each binding should be manufactured to maximise profits? (assume that the publisher sells as many of each type of book as is produced.)

You'll notice that in this question, the table has been started off for you. It might be a good idea to extend this table to include all the information given by adding the restrictions on time and profits, i.e.

	Sewing (mins)	Gluing	Profit (P)
Paperback	2	4	25
Library	3	10	60
Total time	420 (in minutes!)	900 (in minutes!)	

Notice that the time available has been converted to minutes to be consistent with the other times given.

Step 1: Decision variables

The decision variables are the number of books to be made in each binding. Let

$$\begin{aligned} x &= \text{number in paperback binding} & \text{and} \\ y &= \text{number in library binding.} \end{aligned}$$

Step 2: Constraints

The constraints are:

$$\begin{aligned} \text{sewing : } 2x + 3y &\leq 420 & \text{and} \\ \text{gluing : } 4x + 10y &\leq 900, \end{aligned}$$

together with the non-negativity conditions

$$\begin{aligned} x &\geq 0 & \text{and} \\ y &\geq 0. \end{aligned}$$

Step 3: Objective function

For each paperback edition, the publisher makes 25p profit. Since we make x number of paperback bindings, the publisher will make $25x$ pence profit.

Similarly, for each library edition, the publisher makes 60p profit. Since we make y number of library bindings, the publisher will make $60y$ pence profit.

The objective is to maximise the profit P pence. The total profit is $25x + 60y$ – thus our aim is to maximise

$$P = 25x + 60y.$$

Thus, to summarise, we have the following linear programming problem:

Maximise $P = 25x + 60y$ subject to the constraints

$$\begin{aligned} 2x + 3y &\leq 420, \\ 4x + 10y &\leq 900, \\ x &\geq 0 \quad \text{and} \\ y &\geq 0. \end{aligned}$$

9.2.3 Solving simultaneous equations

We are often interested in the range of values of x and y that satisfy the constraints. One way to find the 'largest' values of x and y that simultaneously satisfy the constraint inequalities is to solve the simultaneous equations given by the two main constraint equations, with the inequalities replaced by an '=' sign. For the book publishers example, we can solve

$$\begin{aligned} 2x + 3y &= 420, \\ 4x + 10y &= 900 \end{aligned}$$

as follows:

9.3 Exercises

1. A factory is to install two types of machine – A and B. Type A requires one operator and occupies 3m^2 of floor space. Type B requires 2 operators and occupies 4m^2 of floor space.

The maximum number of operators available is 40 and the total floor space available is 100m^2 .

Given that the weekly profits on type A and type B machines are £75 and £120 (respectively), formulate a linear programming problem to find the number of each machine that should be bought in order to maximise the profit. *Do not attempt to solve this problem.*

Hint: to identify the inequalities, try completing the following table:

	Number of operators	Floor space (m^2)
Type A		
Type B		
Limits		

Find the values of each machine type that require 40 operators and 100m^2 of floor space.

2. A small toy manufacturer makes 2 kinds of toys, A and B. Each type of toy spends time on a machine, and time with a craftsman who adds detail by hand.

Each of toy A takes up 3 hours of machine time and 2 hours of the craftsman's time. Each of toy B takes up 3 hours of machine time and 4 hours of the craftsman's time.

The total machine time available each week is 40 hours. The manufacturer employs two part-time craftsmen who both work 25 hours each week – the manufacturer operates a family-friendly policy of no overtime.

The manufacturer will make £10 profit on each of toy A sold, and £12 profit on each of toy B. Formulate this scenario as a linear programming problem which will determine how many of each type of toy should be made in order to maximise the manufacturer's profit. *Do not attempt to solve this problem.*

Find the values of each toy that require 40 hours of machine time and 25 hours of craftsmen's time.

Chapter 10

Linear Programming II – NOT EXAMINED

10.1 Graphical solutions for two–variable problems

In last week’s lecture we discussed how to *formulate* a linear programming problem; this week, we consider how to *solve* such problems. Linear programming problems that involve only two decision variables, x and y , may be solved graphically, that is by drawing lines associated with the constraints, identifying a region of possible solutions and then locating a point in this region that has a particular property (i.e. a point which maximises profit or minimises cost). To solve linear programming problems using graphical methods you need to be able to use the following basic mathematical ideas.

10.1.1 Sets of points defined by a linear inequality

Any equation of the form $ax + by = c$, where a , b and c are numbers, is called a **linear equation**. For example, $3x + 4y = 12$ is a linear equation. In the x, y plane this is the equation of a straight line, and this line may be drawn by identifying any two points on it (since it only takes two points to draw a straight line).

For example, for the equation $3x + 4y = 12$, we can identify two points which lie on the straight line by considering what happens when $x = 0$ and what happens when $y = 0$.

- When $x = 0$, we have

$$3 \times 0 + 4y = 12 \quad \text{i.e.}$$

$$0 + 4y = 12 \quad \text{i.e.}$$

$$4y = 12 \quad \text{i.e.}$$

$$y = 3.$$

So one point on the line $3x + 4y = 12$ is at $x = 0, y = 3$, or $(0, 3)$ (remember, we go “along the corridor and up the stairs” when plotting points given by coordinates, so for this point we’d go “0 along, and 3 up”).

- When $y = 0$, we have

$$3x + 4 \times 0 = 12 \quad \text{i.e.}$$

$$3x + 0 = 12 \quad \text{i.e.}$$

$$3x = 12 \quad \text{i.e.}$$

$$x = 4.$$

So another point on the line $3x + 4y = 12$ is at $x = 4, y = 0$, or $(4, 0)$.

Since we have the coordinates of two points which lie on the line, and the line is a straight line, we can plot the line with this equation! (figure 10.1)

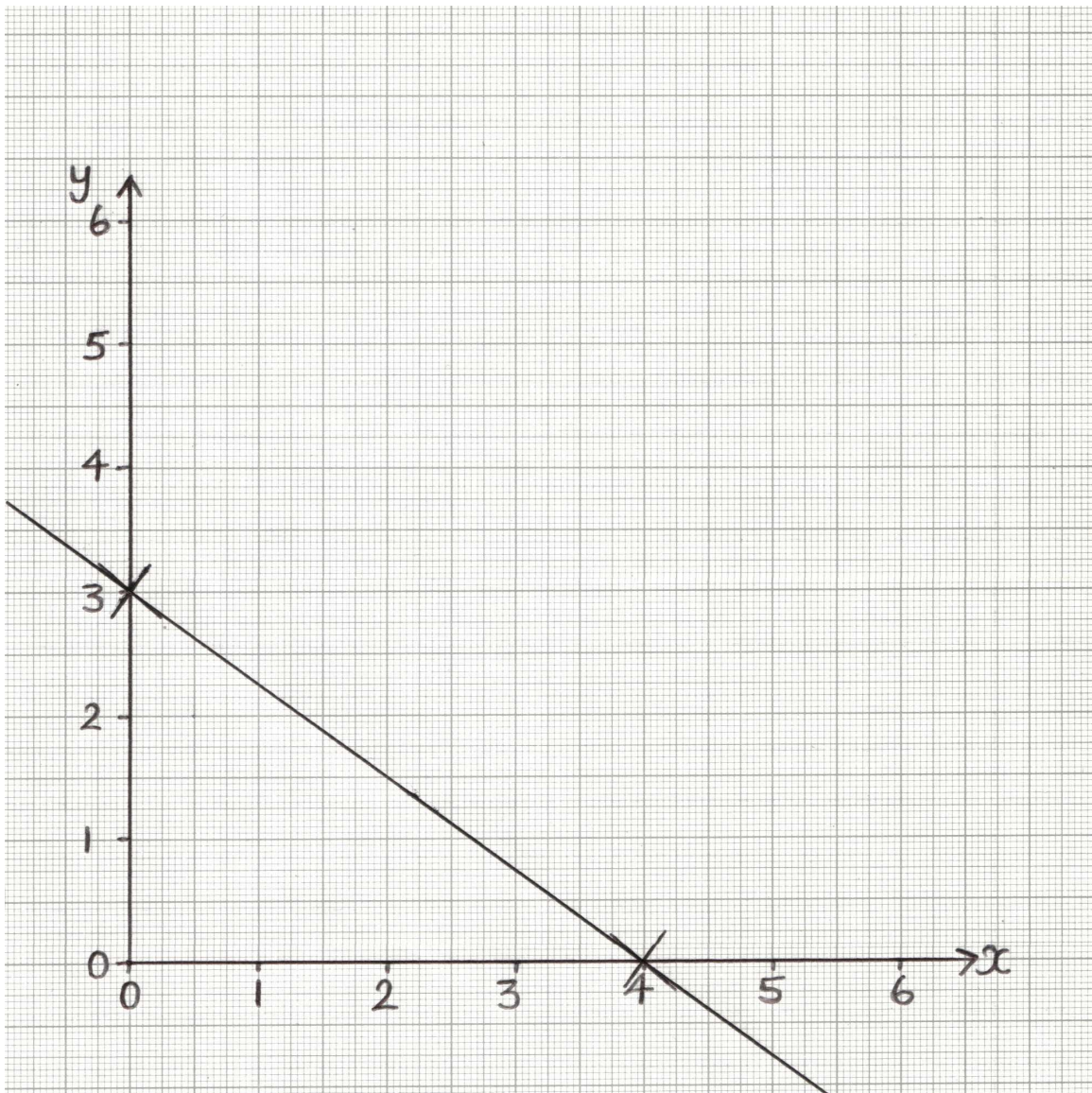


Figure 10.1: Plot of the line $3x + 4y = 12$

Any straight line divides the xy -plane into two half-planes. If the equation of the line is $ax + by = c$ then on one side of the line you have $ax + by < c$ and on the other side of the line you have $ax + by > c$.

Example

Show on a diagram the region for which $3x + 4y \leq 12$.

The first step is to draw the line $3x + 4y = 12$ (as in figure 10.1). We are interested in the region which is *less than or equal to 12*; we can find out where abouts this is by seeing if the origin lies in the required region or not. Now at the origin, $x = 0$ and $y = 0$, so we have $3x + 4y = 3 \times 0 + 4 \times 0 = 0$, which is less than 12. Thus, the origin *does* lie in the required region, and the half-plane required contains the origin. So all points on the line or below it satisfy the inequality. We call this the **admissible set** defined by the inequality. We show this on the graph by shading out points which do not lie in the required region, or shading out the **inadmissible set** of points (see figure 10.2).

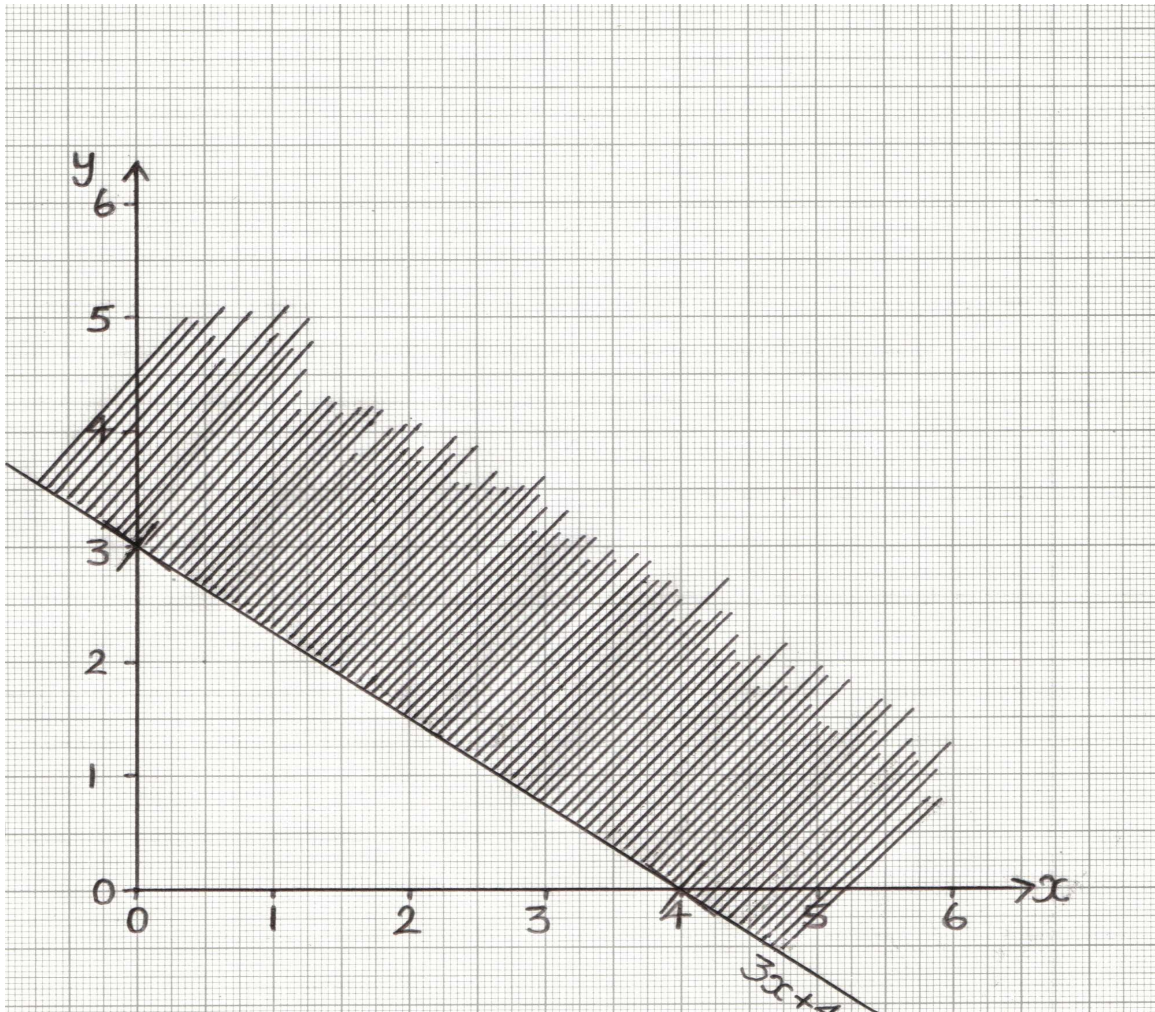


Figure 10.2: Indicating the region for which $3x + 4y \leq 12$

10.1.2 Sets of points defined by a collection of inequalities

Each inequality in a linear programming problem will produce an admissible set. To find a solution to the problem you need to find the set of points which satisfies all the inequalities *simultaneously*. This is obtained graphically by drawing a diagram like the one in figure 10.2 but showing *all* the inequalities. **The required region is then the one which does not contain any shading marks.**

Example

Indicate on a diagram the region for which

$$\begin{aligned} 3x + 4y &\leq 12, \\ 3x + 2y &\leq 9, \\ x &\geq 0 \quad \text{and} \\ y &\geq 0. \end{aligned}$$

We've already looked at how to show the admissible set for the first inequality $3x + 4y \leq 12$; the same procedure can be used for the inequality $3x + 2y \leq 9$. Firstly, we need to plot the line $3x + 2y = 9$. Again, it's easiest to consider what happens when $x = 0$ and $y = 0$. For example,

- When $x = 0$, we have

$$\begin{aligned} 3 \times 0 + 2y &= 9 && \text{i.e.} \\ 0 + 2y &= 9 && \text{i.e.} \\ 2y &= 9 && \text{so} \\ y &= 4.5. \end{aligned}$$

- Similarly, when $y = 0$, we have

$$\begin{aligned} 3x + 2 \times 0 &= 9 && \text{i.e.} \\ 3x + 0 &= 9 && \text{i.e.} \\ 3x &= 9 && \text{so} \\ x &= 3. \end{aligned}$$

So the points $x = 0, y = 4.5$ and $x = 3, y = 0$, i.e. $(0, 4.5)$ and $(3, 0)$ lie on the line with equation $3x + 2y = 9$. Figure 10.3 shows a plot of the line with this equation. The plot also shows the set of admissible points for our inequality; remember, we want the set of points such that $3x + 2y \leq 9$, so the region we require lies on, or underneath, the line drawn. The inadmissible set has been shaded out.

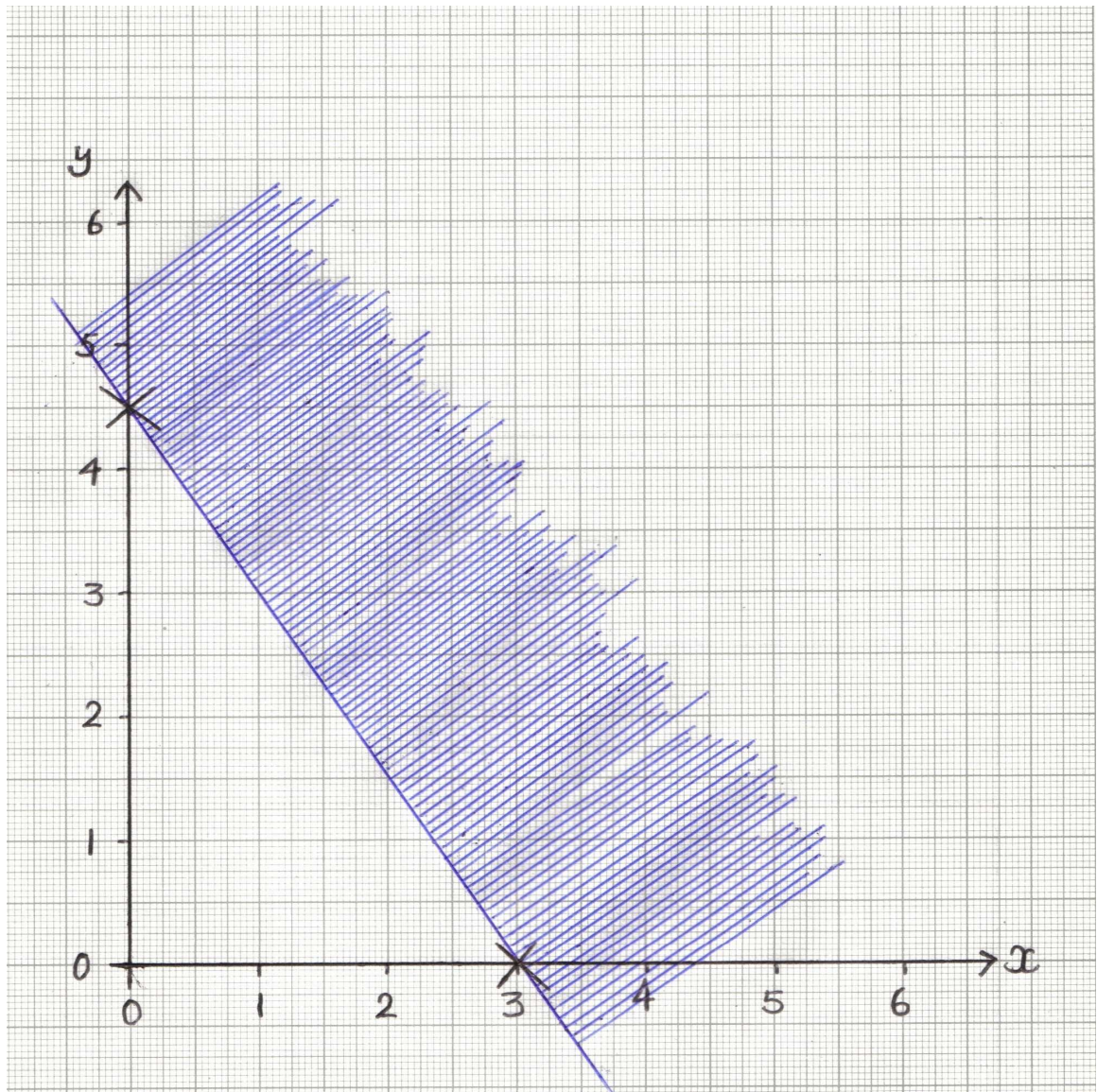


Figure 10.3: Plot of the line $3x + 2y = 9$; also shown is the admissible region for the inequality $3x + 2y \leq 9$

For the inequality $x \geq 0$ and the inequality $y \geq 0$, we shade out the areas below the x axis and to the left of the y axis.

Combining these with the inequality $3x + 4y \leq 12$ considered previously, we get

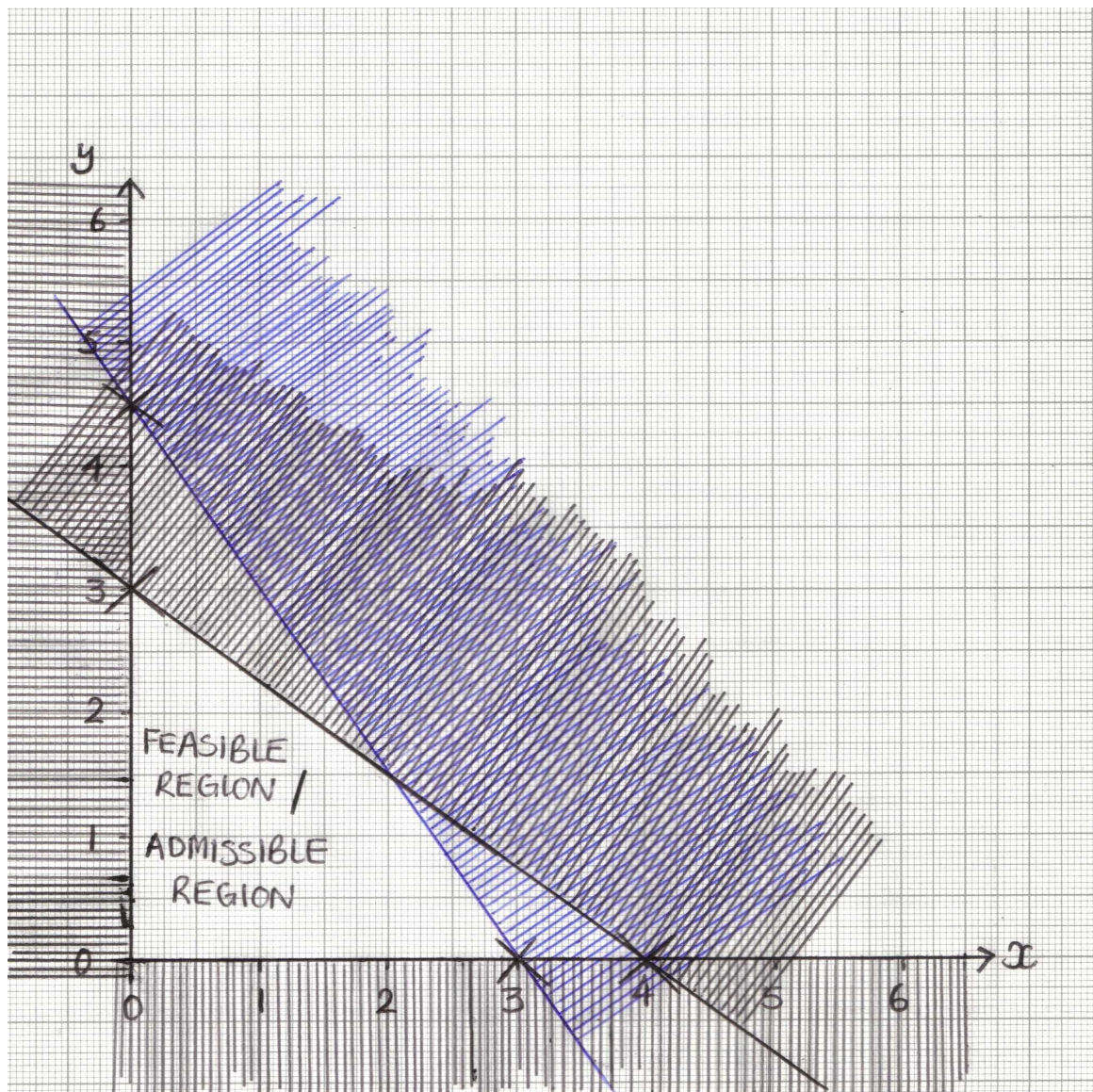


Figure 10.4: Diagram showing the admissible region satisfying the inequalities $3x + 4y \leq 12$, $3x + 2y \leq 9$, $x \geq 0$ and $y \geq 0$

We now use these techniques to obtain feasible solutions to the linear programming problems discussed last week, and look at how to find the *optimal* solution for each of these problems.

Recall the scenarios:

1. a chair manufacturer,
2. a book publisher.

10.2 Solving linear programming problems

10.2.1 Example (A chair manufacturer)

Recall the problem posed in example 9.2.1:

A manufacturer makes two kinds of chairs, A and B, each of which has to be processed in two departments, I and II. Chair A has to be processed in department I for 3 hours and in department II for 2 hours. Chair B has to be processed in department I for 3 hours and in department II for 4 hours.

The time available in department I in any given month is 120 hours, and the time available in department II, in the same month, is 150 hours.

Chair A has a selling price of £10 and chair B has a selling price of £12.

The manufacturer wishes to maximise his income. How many of each chair should be made in order to achieve this objective? You may assume that all chairs made can be sold.

Last week we considered how to *formulate* this situation as a linear programming problem; using the graphical techniques outlined in section 10.1, we will now consider how to *solve* this problem.

Recall that things are made much easier if we summarise the information given in a table; we can then read down the columns to obtain the inequalities for the linear programming problem. The table below summarises the information given in this particular problem.

Type of chair	Time in dept. I (hours)	Time in dept. II (hours)	Selling price (£)
A	3	2	10
B	3	4	12
Total time available	120	150	

Recall the three steps involved in formulating a linear programming problem:

1. Identify the **decision variables**
2. Identify the **constraints**
3. State the **objective function**

In this example, the decision variables were identified as

$$\begin{aligned}
 x &= \text{number of type A chairs made} & \text{and} \\
 y &= \text{number of type B chairs made.}
 \end{aligned}$$

The constraints were

$$\begin{aligned} 3x + 3y &\leq 120, \\ 2x + 4y &\leq 150, \\ x &\geq 0 \quad \text{and} \\ y &\geq 0. \end{aligned}$$

We then identified that the objective was to maximise the income, which we called Z , where

$$Z = 10x + 12y.$$

So, in summary, the linear programming problem is:

Maximise $Z = 10x + 12y$ subject to the following constraints:

$$\begin{aligned} 3x + 3y &\leq 120, \\ 2x + 4y &\leq 150, \\ x &\geq 0 \quad \text{and} \\ y &\geq 0. \end{aligned}$$

To find the **feasible region** for this problem (i.e. the region which satisfies all of our inequalities), we proceed by indicating, on a diagram, the region for which all of the inequalities hold (as we did in the example in section 10.1.2).

The first inequality is $3x + 3y \leq 120$. To show this on a diagram, we first need to plot the line $3x + 3y = 120$.

- When $x = 0$, we have

$$\begin{aligned} 3 \times 0 + 3y &= 120 && \text{i.e.} \\ 3y &= 120 && \text{i.e.} \\ y &= 40. \end{aligned}$$

- When $y = 0$, we have

$$\begin{aligned} 3x + 3 \times 0 &= 120 && \text{i.e.} \\ 3x &= 120 && \text{i.e.} \\ x &= 40. \end{aligned}$$

These points are plotted on figure 10.5, and the line with equation $3x + 3y = 120$ drawn. Since we want $3x + 3y \leq 120$, our region of interest lies on or below the line, and so we shade out the space above the line.

Now consider the second inequality $2x + 4y \leq 150$. Again, to show this on a diagram, we first need to plot the line $2x + 4y = 150$.

- When $x = 0$, we have

$$\begin{aligned}2 \times 0 + 4y &= 150 && \text{i.e.} \\4y &= 150 && \text{i.e.} \\y &= 37.5.\end{aligned}$$

- When $y = 0$, we have

$$\begin{aligned}2x + 4 \times 0 &= 150 && \text{i.e.} \\2x &= 150 && \text{i.e.} \\x &= 75.\end{aligned}$$

Again, these points are plotted on figure 10.5, and the line with equation $2x + 4y = 150$ drawn. Since we want $2x + 4y \leq 150$, our region of interest lies on or below the line, and so we shade out the space above the line.

On figure 10.5 we also shade out the inadmissible regions for the two non-negativity constraints.

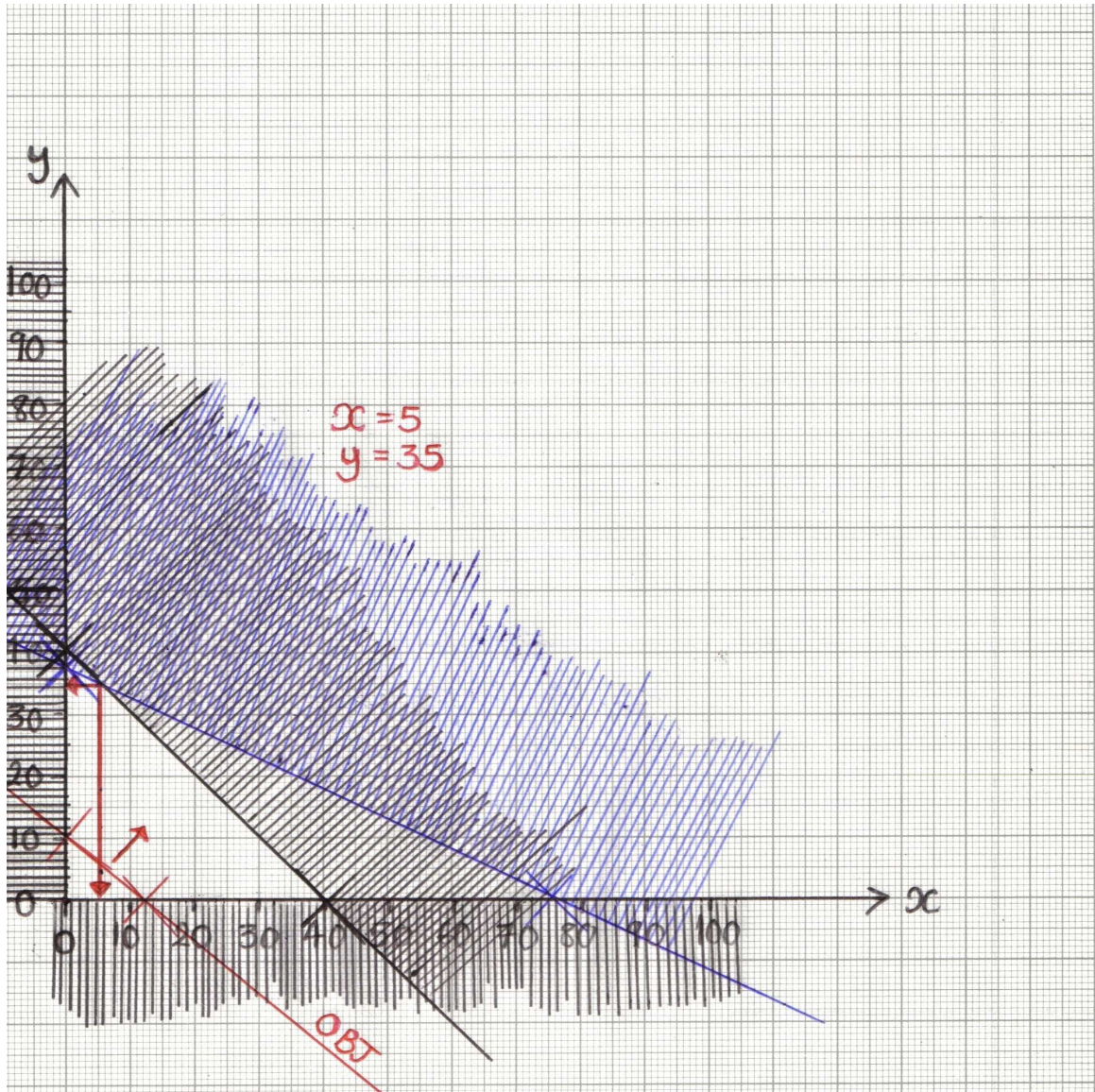


Figure 10.5: Feasible region and objective function for the chair manufacturing problem

The unshaded region in figure 10.5 shows the feasible region associated with our set of inequalities. What we must do now is find the point in that region which meets our objective – i.e. the point in that region which maximises income. One way of doing this is to also plot the objective function. Now our objective function is

$$Z = 10x + 12y,$$

where Z is our income. When Z takes different values we get a family of parallel straight lines. We need to choose a starting value for Z in order to be able to plot the objective function. It's often a good idea to try a value which is a multiple of both the coefficients of x and y . The coefficient of x is 10 and the coefficient of y is 12, so we could try a starting value of $Z = 120$. Thus, the objective function is now

$$10x + 12y = 120.$$

We can plot this line in the same way as before – i.e. consider what happens when x and y are zero.

- When $x = 0$, we have

$$\begin{aligned} 10 \times 0 + 12y &= 120 && \text{i.e.} \\ 12y &= 120 && \text{i.e.} \\ y &= 10. \end{aligned}$$

- When $y = 0$, we have

$$\begin{aligned} 10x + 12 \times 0 &= 120 && \text{i.e.} \\ 10x &= 120 && \text{i.e.} \\ x &= 12. \end{aligned}$$

This line is also plotted on figure 10.5. Notice that this line does not give the optimal income; the origin represents zero income, and we want to move as far away from this as possible. The largest value of Z (income) will occur at the point in the feasible region that is *furthest* from the origin, but still parallel to the objective line. You can find this point by sliding a ruler over the feasible region so that it is always parallel to the objective function drawn. This will enable you to identify the point that is furthest from the origin. In fact, you should notice that the point in the feasible region furthest from the origin (and parallel to the objective function) is the intersection of the two lines with equations $3x + 3y = 120$ and $2x + 4y = 150$.

All points in the feasible region satisfy our inequalities, but only one point maximises income. Once this point has been identified, we can simply “read off” the x and y values. Doing so give $x = 5$ and $y = 35$, and so, in order to maximise income, we should make 5 type A chairs and 35 type B chairs. This will give an income of

$$\begin{aligned} Z &= 10x + 12y && \text{i.e.} \\ Z &= 10 \times 5 + 12 \times 35 && \text{i.e.} \\ Z &= 50 + 420 \\ &= 470, \end{aligned}$$

i.e. £470.

10.2.2 Example (A book publisher)

Recall the problem considered in example 9.2.2:

A book publisher is planning to produce a book in two different bindings: paperback and library. Each book goes through a sewing process and a gluing process. The sewing process is available for 7 hours per day and the gluing process for 15 hours per day. The profits are 25p on a paperback edition and 60p on a library edition. How many books in each binding should be manufactured to maximise profits? (assume that the publisher sells as many of each type of

book as is produced.)

The table below summarises the information given in the above paragraph:

	Sewing (mins)	Gluings	Profit (P)
Paperback	2	4	25
Library	3	10	60
Total time	420 (in minutes!)	900 (in minutes!)	

The linear programming problem is summarised below:

Maximise $P = 25x + 60y$ subject to the constraints

$$\begin{aligned} 2x + 3y &\leq 420, \\ 4x + 10y &\leq 900, \\ x &\geq 0 \quad \text{and} \\ y &\geq 0. \end{aligned}$$

As before, we solve this problem by first of all drawing the feasible region satisfying all of our inequalities. For the inequality $2x + 3y \leq 420$, we draw the line $2x + 3y = 420$ and then shade out the area *above* the line (since the required region is less than, or equal to, 420). So:

- When $x = 0$, we have

$$\begin{aligned} 2 \times 0 + 3y &= 420 && \text{i.e.} \\ 3y &= 420 && \text{i.e.} \\ y &= 140. \end{aligned}$$

- Similarly, when $y = 0$, we have

$$\begin{aligned} 2x + 3 \times 0 &= 420 && \text{i.e.} \\ 2x &= 420 && \text{i.e.} \\ x &= 210. \end{aligned}$$

We consider the inequality $4x + 10y \leq 900$ in a similar way; we draw the line with equation $4x + 10y = 900$ and then shade out the region above this line.

- When $x = 0$, we have

$$\begin{aligned} 4 \times 0 + 10y &= 900 && \text{i.e.} \\ 10y &= 900 && \text{i.e.} \\ y &= 90. \end{aligned}$$

- Similarly, when $y = 0$, we have

$$\begin{aligned} 4x + 10 \times 0 &= 900 && \text{i.e.} \\ 4x &= 900 && \text{i.e.} \\ x &= 225. \end{aligned}$$

Both these lines are plotted on figure 10.6, and the inadmissible regions shaded out. Also shaded out are the inadmissible regions for the two non-negativity constraints.

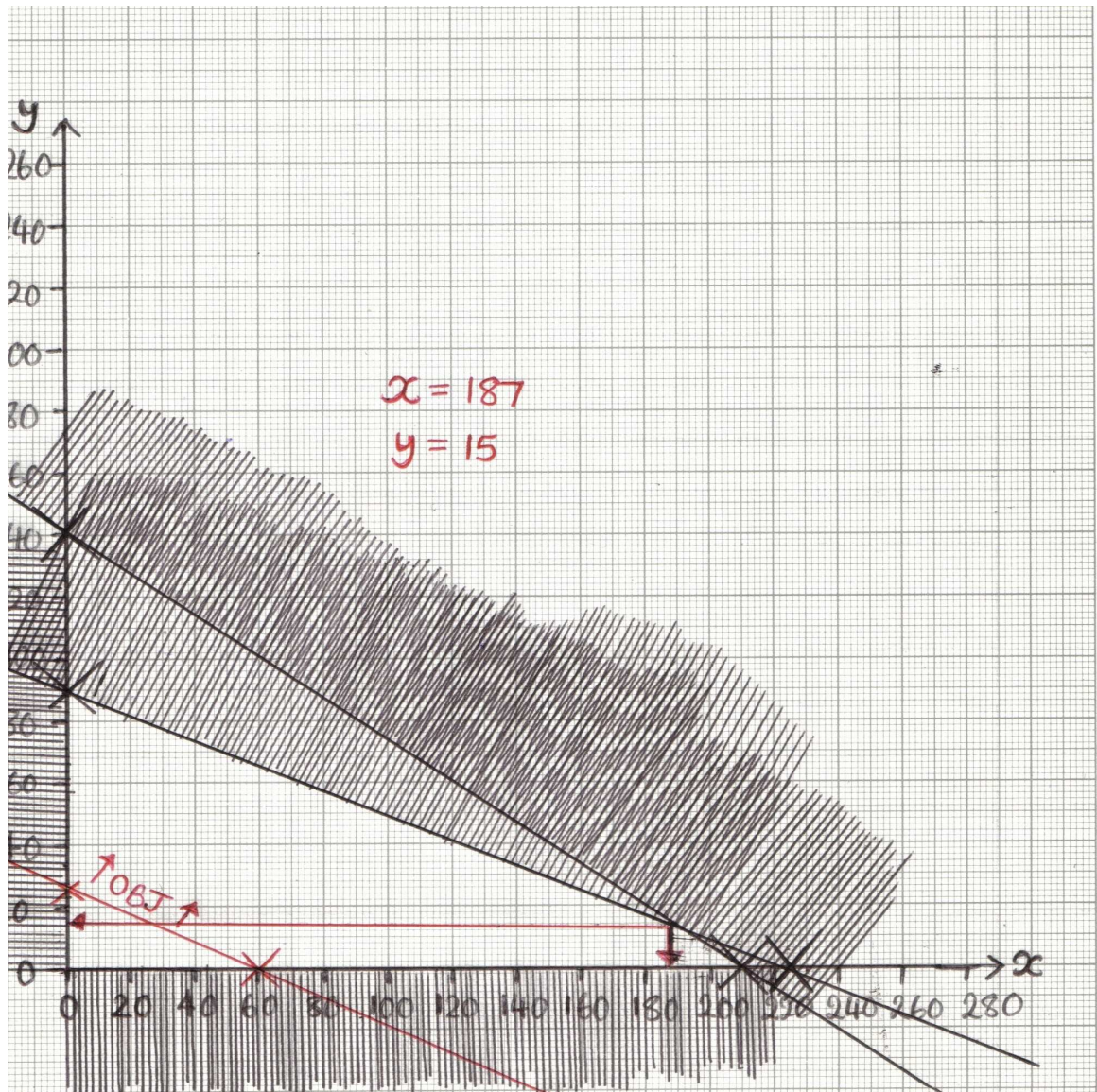


Figure 10.6: Feasible region and objective function for the book publisher's problem

We now need to plot the objective function. The objective is to maximise profit, P , where

$$P = 25x + 60y.$$

Of course, we don't yet know what P is, so we choose a starting value. As before, it's convenient to choose a value which is a multiple of both the x and y coefficients. The x coefficient is 25 and the y coefficient is 60, so we choose $25 \times 60 = 1500$. Thus, our objective function is

$$25x + 60y = 1500.$$

To plot this line, we consider what happens when x and y are zero.

- When $x = 0$, we have

$$\begin{aligned} 25 \times 0 + 60y &= 1500 && \text{i.e.} \\ 60y &= 1500 && \text{i.e.} \\ y &= 25. \end{aligned}$$

- Similarly, when $y = 0$ we have

$$\begin{aligned} 25x + 60 \times 0 &= 1500 && \text{i.e.} \\ 25x &= 1500 && \text{i.e.} \\ x &= 60. \end{aligned}$$

This line is plotted on figure 10.6. As before, we need to move this line as far away from the origin as possible in order to maximise profits, but keep the line inside the feasible region. You can see that the point in the feasible region which is furthest away from the origin, but parallel to the objective line, is the intersection between the lines with equations $2x + 3y = 420$ and $4x + 10y = 900$. As before, we can “read off” our solutions from the graph. Doing so, gives $x = 187.5$ and $y = 15$. Thus, the book publisher should make 187.5 paperback bindings and 15 library bindings in order to maximise profit. Since it doesn’t really make sense to make 187.5 paperback bindings, we round down to 187. Thus, the publisher should make 187 paperback bindings and 15 library bindings in order to maximise profit. This will give

$$\begin{aligned} P &= 25x + 60y \\ &= 25 \times 187 + 60 \times 15 \\ &= 5575, \end{aligned}$$

i.e. 5575 pence profit, or £55.75.

10.3 Exercises

1. Indicate on a diagram the region for which

$$\begin{aligned} 4x + 3y &\leq 12, \\ 2x + 5y &\leq 10, \\ x &\geq 0 \quad \text{and} \\ y &\geq 0. \end{aligned}$$

2. Solve the linear programming problem formulated in question 1 of exercises 9. [*Hint: Try a starting value for the profit of £600*]. Verify your solution algebraically.
3. Kuddly Pals Co. Ltd make two types of giant soft toy: bears and cats. The quantity of material needed and the time taken to make each type of toy is given in the table below.

Toy	Material (m ²)	Time (minutes)
Bear	5	12
Cat	8	8

Each day the company can process up to 2000m² of material and there are 48 worker hours available to assemble the toys.

The profit made on each bear is £1.50 and on each cat is £1.75. Kuddly Pals Co. Ltd wishes to maximise its daily profit.

- Formulate the company's situation as a linear programming problem.
- Draw a suitable diagram to enable the problem to be solved graphically, indicating the feasible region and the direction of the objective line.
- Use your diagram to find the company's maximum profit, £*P*. Verify your solution algebraically.