Unit 13

Applications

Introduction

This short unit considers a number of datasets, one per section (except the last), to illustrate the application of some of the statistical methods developed in the module. Each application is explored through a set of activities designed to give you practice in answering questions of the sort you are likely to meet in the examination. As such, this module also serves as useful revision, although obviously there is not enough space in a single unit to illustrate everything that has been covered in the module.

All the calculations, and even graphical representations, that you are asked to make in this unit can, and should, be done 'by hand', perhaps using your calculator and/or statistical tables; you do not need to use your computer at all.

1 Chocolates

Cadbury Heroes are miniature chocolate bars. At the time of writing, a 712 g tub of Cadbury Heroes contains seven different types of chocolate bar: Caramel, Creme Egg Twisted, Dairy Milk, Eclair, Fudge, Twirl and Wispa. In this section, we will investigate the distribution of the different types of chocolate bar found in 712 g tubs of Cadbury Heroes. Because it is convenient for random variables to take numerical values, we will code the different types of chocolate bars so that each has a numerical value. So, let Caramel be coded as 1, Creme Egg Twisted as 2, Dairy Milk as 3, Eclair as 4, Fudge as 5, Twirl as 6, and Wispa as 7.

Activity 1 *Which distribution?*

For each chocolate bar in a tub of Cadbury Heroes, let X be its type.

(a) What is the range of X?

(b) It seems reasonable to expect there to be equal numbers of each type of chocolate bar within a 712 g tub of Cadbury Heroes. In this case, suggest a suitable distribution for X and write down its probability mass function and cumulative distribution function.

The solution to Activity 1 argued that if a tub of Heroes has equal numbers of each type of chocolate bar, then a suitable distribution for X is the discrete uniform distribution with parameters $m = 1$ and $n = 7$. The numbers of the different types of chocolate bar in one particular 712 g tub of Cadbury Heroes (purchased in December 2016) are given in Table 1 (overleaf).

Table 1 Number of chocolates of different types

Type of chocolate	Frequency
1 (Caramel)	17
2 (Creme Egg Twisted)	10
3 (Dairy Milk)	12
4 (Eclair)	15
5 (Fudge)	5
6 (Twirl)	4
7 (Wispa)	8

(Source: C.M. Queen, The Open University)

This tub certainly didn't have equal numbers of each type of chocolate bar. Does that mean that the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is not a suitable model for X? Or could the frequencies observed in Table 1 have arisen by chance from this discrete uniform distribution? You will use a chi-squared goodness-of-fit test to answer this question in the next activity.

The chi-squared goodness-of-fit test was covered in Section 2 of Unit 10.

Activity 2 *Does the discrete uniform distribution fit the data?*

(a) There were a total of 71 chocolate bars in the tub of Cadbury Heroes which provided the data in Table 1. Calculate the expected frequencies for X taking values $1, 2, \ldots, 7$, for a tub of 71 Cadbury Heroes when X is modelled by the discrete uniform distribution with parameters $m = 1$ and $n = 7$.

(b) Carry out a chi-squared goodness-of-fit test to test whether the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is a suitable model for the data in Table 1.

The solution to Activity 2 concluded that there is moderate evidence to suggest that the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is not a suitable model for X. This means that there is evidence to suggest that the different types of chocolate bars in a 712 g tub of Cadbury Heroes do not occur in equal numbers. Of course, these data were only for one single tub of chocolates, and the result from this test doesn't necessarily mean that *all* such tubs do not have equal numbers of each type of chocolate bar. To investigate this further, we would need more data and would need to see the contents of more 712 g tubs of the chocolates. This might be a simpler and more attractive exercise in gathering further data than statisticians are usually faced with!

2 Paralympic Games 2016

Since 2004, five-a-side football has been an event in the Paralympic Games played by athletes with visual impairment, including blindness. A special football is used which makes a noise when it moves so that players can locate it through sound. In order that no player has an unfair advantage, all players apart from the goalkeepers wear eye shades. Each team has 5 players, including the goalkeeper, who may be sighted and is also allowed to act as a guide during the game.

The Rio 2016 Paralympic Games saw eight five-a-side football teams compete. The teams were divided into two groups, A and B, of four teams each. All teams within a group played each other; the two best teams in each group went on to play in the semi-finals, and the winners of the two semi-finals played in the final. There were also matches to decide who came in 3rd, 4th, . . . , 8th place. In all, 18 matches were played.

For each match, the number of goals for each team was recorded: these ranged from 0 goals to 3 goals. Table 2 gives the number of times that a team scored 0, 1, 2 and 3 goals for the 18 matches (including the scores of both teams for each match). Games in the knockout stage of the competition which ended in a draw were decided by having a 'penalty shoot-out', in which each team takes the same number of penalties and the team which scores the most penalties is declared the winner. Only the original scores for the games before any penalty shoot-outs are given in Table 2.

Table 2 Number of goals scored

Number of goals	Frequency
0	21
1	8
2	6
3	1

(Source: Wikipedia, http://en.wikipedia.org/wiki/Football_5-a-side_at_the_2016_Summer_Paralympics, 5 October 2016)

We are interested in finding a suitable probability model for the random variable X, the number of goals scored in individual matches by individual teams in five-a-side football matches at the 2016 Paralympics. We immediately make one strong – and clearly wrong – assumption: that the number of goals scored by any team in any match is independent of the numbers of goals scored by any team (including the same one) in any other match (or indeed of the number scored by the opposing team in the same match). Accommodating dependence between the numbers of goals scored by different teams (some are stronger, some are weaker), accounting for different opposing teams (again, some are stronger, some are weaker) and other factors, makes the modelling exercise much more difficult and beyond the scope of this module. However, models being models, a model developed under an assumption of independence can still be useful for some purposes, provided that we remember the strong assumptions made.

Activity 3 *A model for goals scored*

(a) Present the data in Table 2 by means of a suitable simple graphical display.

(b) Suggest a modelling distribution for random variable X. Give two reasons to support your choice of distribution.

The solution to Activity 3 suggested the Poisson distribution as a plausible model for X, the number of goals scored in individual matches by individual teams in five-a-side football matches at the 2016 Paralympics. Let $\theta > 0$ denote the parameter of this distribution. In the next activity, you will obtain the likelihood of θ for these data assuming a Poisson model, and use this to find the maximum likelihood estimate $\widehat{\theta}$ of θ.

The Poisson distribution was covered in Section 4 of Unit 3, and the likelihood and maximum likelihood estimation were covered in Unit 7.

Activity 4 *Maximum likelihood estimate*

(a) Show that the likelihood of θ for these data is

$$L(\theta) = \frac{e^{-36\theta}\theta^{23}}{384}.$$

(b) By differentiating the likelihood, find the maximum likelihood estimate $\widehat{\theta}$ of θ, giving its value correct to three decimal places.

You showed in Activity 4 that the MLE of θ for these data is $\widehat{\theta} \simeq 0.639$. So if a Poisson model is a sensible model for the data, then the Poisson model which fits the data best is Poisson(0.639). In the next activity you will test whether in fact the Poisson(0.639) distribution is a good fit to the data in Table 2.

Activity 5 *Is the Poisson model a good fit?*

(a) Calculate the expected frequencies for scoring 0, 1 and ≥ 2 goals in 36 scores when the number of goals scored is modelled by the Poisson(0.639) distribution.

(b) A chi-squared goodness-of-fit test is to be carried out using the categories: ‘0 goals’, ‘1 goal’ and ‘$\geq$ 2 goals’. Explain why the category ‘$\geq$ 2 goals’ is to be used rather than separate categories for ‘2 goals’, ‘3 goals’, and so on.

The chi-squared goodness-of-fit test was covered in Section 2 of Unit 10.

(c) Carry out a chi-squared goodness-of-fit test at the 5% significance level, using the data categorised as in part (b), to test whether the Poisson(0.639) model is a suitable model for the data in Table 2.

The solution to Activity 5 concluded that the Poisson(0.639) distribution seems to be a suitable model for the number of goals scored in individual matches by individual teams in five-a-side football matches at the Paralympics, using the 2016 Rio event for data. It would be interesting to investigate how widely such a Poisson model remains applicable, and how much there is to be gained by more sophisticated modelling taking into account strengths of teams, who's playing whom, etc.

3 Times between major tsunamis

On its website, *National Geographic* magazine describes a tsunami as: 'a series of ocean waves that sends surges of water, sometimes reaching heights of over 100 feet (30.5 meters), onto land. These walls of water can cause widespread destruction when they crash ashore' (http://www.nationalgeographic.com/environment/natural-disasters/tsunamis). Tsunamis are typically caused by underwater earthquakes, landslides or volcanic eruptions.

Table 3 contains the 29 'waiting times', in months, between 30 major tsunamis occurring worldwide between January 1950 and December 2015. The numbers should be read across the rows.

Table 3 Time interval between major tsunamis (months)

44	24	22	41	5	3	8	48	90	40	5	36	112	10	11
49	13	64	19	4	5	8	21	5	8	4	1	23	31	

(Source: Wikipedia, http://en.wikipedia.org/wiki/List_of_historical_tsunamis, 5 October 2016)

Because these data are waiting times, the two most likely distributions to model the data are the exponential distribution and the geometric distribution.

The exponential and geometric distributions for modelling waiting times were covered in Section 2 of Unit 5.

Activity 6 *A model for the waiting times between major tsunamis*

Explain why an exponential distribution would be more appropriate than a geometric distribution for modelling the variation in the waiting times between major tsunamis.

In the solution to Activity 6, the case was made that the exponential distribution is more appropriate than the geometric for modelling the variation in waiting times between major tsunamis. In the following activity, you will explore whether the exponential distribution seems to be a reasonable model for these data.

Activity 7 *Is an exponential model reasonable?*

Figure 1 shows a frequency histogram of the 29 waiting times between major tsunamis given in Table 3. For these data, the sample mean is (exactly) 26 months, and the sample standard deviation is (approximately) 27.01 months.

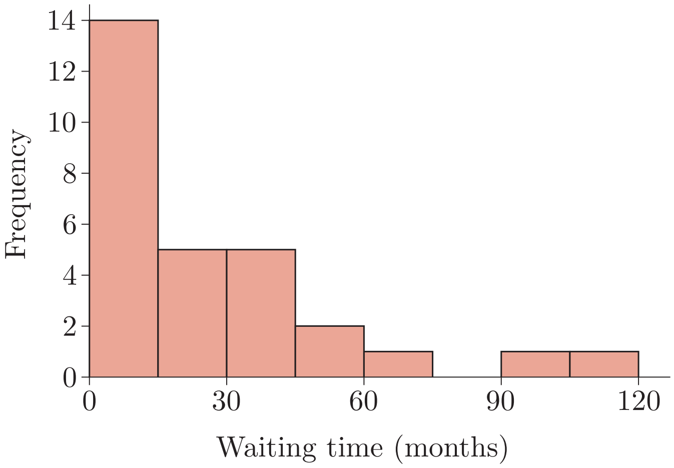


Figure 1 A histogram of waiting times between major tsunamis

Explain briefly whether or not an exponential distribution seems to be a reasonable model for the waiting times between major tsunamis.

The solution to Activity 7 suggested that an exponential distribution would appear to be a reasonable model for the variation in waiting times between major tsunamis. But which exponential distribution do the data suggest?

Activity 8 *Which exponential distribution?*

Use the information given in Activity 7 to find an estimate for the parameter λ of an exponential distribution to model the variation in waiting times between major tsunamis.

The solution to Activity 8 suggested an exponential distribution with parameter $\lambda = 1/26 \simeq 0.038$ for modelling the waiting times between major tsunamis. You will use this model to calculate some probabilities in the next activity.

Activity 9 *Calculating probabilities*

In this activity, use an exponential distribution with parameter $\lambda = 1/26$ to model the waiting times between major tsunamis.

(a) According to the model, what is the probability that the waiting time between two successive major tsunamis is at least one year?

(b) According to the model, what is the probability that the waiting time between two successive major tsunamis is less than 6 months?

(c) For the assumed exponential model, out of 29 waiting times, how many waiting times would be expected to be at least one year? How many would be expected to be less than 6 months? How do these expected values compare with the observed numbers of waiting times that were at least one year and less than 6 months, respectively?

In the next activity, we will make the assumption that the occurrences of major tsunamis may be modelled by a Poisson process, and in the final activity of this section you will investigate whether the data in Table 3 justify this assumption.

The Poisson process was covered in Section 3 of Unit 5.

Activity 10 *Assuming a Poisson process*

Assume that the occurrences of major tsunamis may be modelled by a Poisson process.

(a) Given that the mean time between major tsunamis for the data in Table 3 was 26 months, what is an estimate of the rate λ of the process per month?

(b) Let X be the number of major tsunamis that occur in one year. What is the distribution of X?

(c) Hence calculate the probability that:

(i) exactly two major tsunamis will occur in one year

(ii) at least one major tsunami will occur in one year.

Activity 11 *Is the assumption of a Poisson process reasonable?*

(a) If the occurrences of major tsunamis are modelled by a Poisson process, what assumptions about the occurrences of major tsunamis are being made?

(b) In Figure 2, the number of major tsunamis that had occurred since the start of the period of observation is plotted against the times at which the major tsunamis occurred. By considering all of the Poisson process model assumptions that you identified in part (a), is it reasonable to assume that the occurrence of major tsunamis may be modelled by a Poisson process?

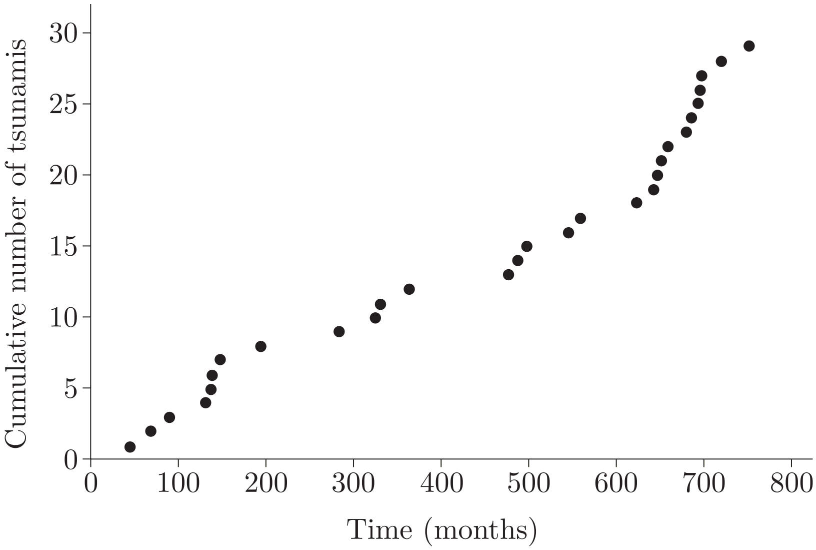


Figure 2 Scatterplot of cumulative number of major tsunamis against their times of occurrence

Activity 11 questioned whether a Poisson process is a reasonable model for the occurrences of major tsunamis. On the one hand, the data suggest that it may not be a reasonable model because Figure 2 suggests that the occurrences are becoming more frequent over time; if so, it is to the earth sciences that we must turn to find an explanation if such an increase in rate is real. But on the other hand, it is possible that the Poisson process *is* a reasonable model for the occurrences of major tsunamis more generally, but other factors concerning these particular data (such as having data for waiting times only in months rather than days, or the possibility that recording of tsunamis is improving over time which makes them appear to occur more frequently) are casting doubt on the model. This illustrates the need for a statistician to keep an open mind about the data available when carrying out a statistical analysis.

4 Pneumonia risk for smokers with chickenpox

Pneumonia can be a serious complication of chickenpox in adults. A study was conducted to determine whether smoking is a risk factor for pneumonia for patients with chickenpox. The data in Table 4 are measurements of the carbon monoxide (CO) transfer factor levels in seven smokers with chickenpox who were admitted to a hospital. CO transfer factor is a measure of lung function; high values are good. CO transfer factor levels were recorded with a view to determining each patient's risk of contracting pneumonia. The measurements were taken when the patients entered the hospital and were repeated one week later. Here we are interested in investigating whether the data suggest that there is a difference in the CO transfer factor levels on entry and one week later. (On admission, patients were treated with intravenous acyclovir at 10 mg/kg, eight-hourly for five days. It is not recorded whether they were required to abstain from smoking.)

CO transfer being measured using modern equipment

Table 4 CO transfer factor levels in smokers with chickenpox

Patient	On entry	One week later
1	40	73
2	50	52
3	56	80
4	58	85
5	60	64
6	62	63
7	66	60

(Source: Ellis, M.E., Neal, K.R. and Webb, A.K. (1987) 'Is smoking a risk factor for pneumonia in patients with chickenpox?', *British Medical Journal*, vol. 294, no. 6578, p. 1002)

With so few data points, a histogram cannot usefully be drawn, but boxplots may be informative; you will investigate these in the next activity.

Activity 12 *A comparative boxplot*

(a) Calculate the median and the interquartile range for each set of CO transfer factor levels in Table 4.

(b) A comparative boxplot for the data is shown in Figure 3.

The median and interquartile range were covered in Section 4 of Unit 1, and comparative boxplots were covered in Subsection 5.3 of that unit.

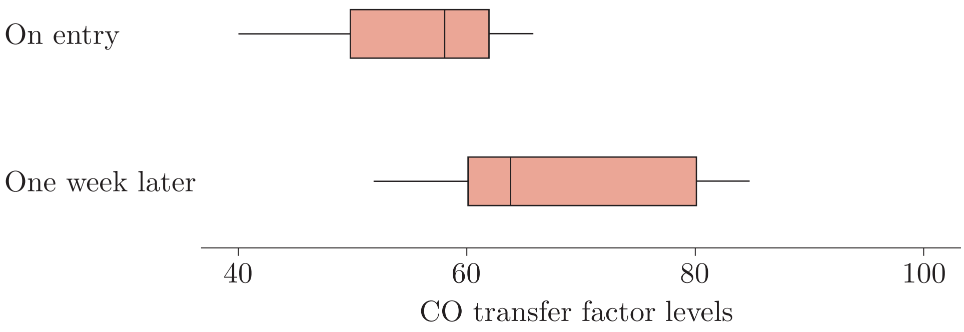


Figure 3 A comparative boxplot for CO transfer factors

Describe two main features of these data as revealed by the comparative boxplot in Figure 3.

As concluded in the solution to Activity 12, the comparative boxplot in Figure 3 suggests that the CO transfer levels are higher one week later than on entry. We can test more formally whether the levels are in fact different. Measurements have been repeated on the same patients, so the data are paired. Therefore, in order to test whether the CO transfer factor levels on entry are different to the CO transfer factor levels one week later, tests involving the differences between the measurements are required.

Activity 13 *Which tests?*

(a) Obtain the differences between measurements in Table 4, calculated as CO transfer factor level 'one week later' minus CO transfer factor level 'on entry'.

(b) Explain how you would investigate whether a normal model for the differences between the CO transfer levels on entry and one week later is plausible.

(c) If the assumption of a normal model is plausible, what test would you use to test whether the CO transfer factor levels on entry and one week later are different? For this test, what are the null and alternative hypotheses?

(d) If the assumption of a normal model is untenable, what alternative test would you use to test whether the CO transfer levels on entry and one week later are different? What are the null and alternative hypotheses in this case?

The solution to Activity 13 suggested two tests for testing whether the CO transfer factor levels on entry and one week later are different: the (one-sample) t-test when the assumption of normality of the differences is plausible, and the Wilcoxon signed rank test when the normality assumption is untenable. In Activities 14 and 15, you will carry out the t-test and Wilcoxon signed rank test, respectively. You will investigate whether the normality assumption is indeed plausible or not in Activity 16.

The t-test was covered in Subsection 3.1 of Unit 9, and the Wilcoxon signed rank test in Subsection 1.1 of Unit 10. Investigating normality was covered in Section 5 of Unit 6.

Activity 14 *Testing when the assumption of normality is justified*

In this activity, we assume that the assumption of normality of differences is plausible.

The differences themselves are in Table 8 in the solution to Activity 13.

(a) The mean of the differences for the data in Table 4 (taking the values 'one week later' minus 'on entry') is 12.14, and the standard deviation of the differences is 15.38. Obtain the value of the test statistic for the

t-test for testing the hypotheses

$$H_0 : \mu_D = 0, \quad H_1 : \mu_D \neq 0,$$

where μ_D is the (population) mean of the differences.

(b) What is the null distribution of the test statistic?

(c) The p-value for the test as calculated by Minitab is 0.082. What do you conclude?

Activity 15 *Testing when the assumption of normality is untenable*

In this activity, we assume that the assumption of normality of differences is untenable.

The differences between measurements in Table 4 (calculated as 'one week later' minus 'on entry') were given in Table 8 in the solution to Activity 13, and are repeated in Table 5 for convenience.

Table 5 Differences 'one week later' minus 'on entry'

Patient	1	2	3	4	5	6	7
Difference	33	2	24	27	4	1	−6

(a) Calculate the Wilcoxon signed rank test statistic for testing the hypotheses

$$H_0 : m_D = 0, \quad H_1 : m_D \neq 0,$$

where m_D is the (population) median of the differences.

(b) The p-value for the test as calculated by Minitab is 0.108. What do you conclude?

The t-test in Activity 14 and the Wilcoxon signed rank test in Activity 15 concluded, respectively, that there was only weak, and little or no, evidence that the CO transfer factor levels were different on entry and one week later. The t-test, however, assumed that the differences were normally distributed, and if that assumption is untenable, then the test is not valid. So is the assumption of normality plausible or not?

Activity 16 *Is the assumption of normality plausible?*

The normal probability plot for the seven differences is given in Figure 4 (overleaf).

Do you think that the assumption that the differences in CO transfer factor levels are normally distributed is plausible or not?

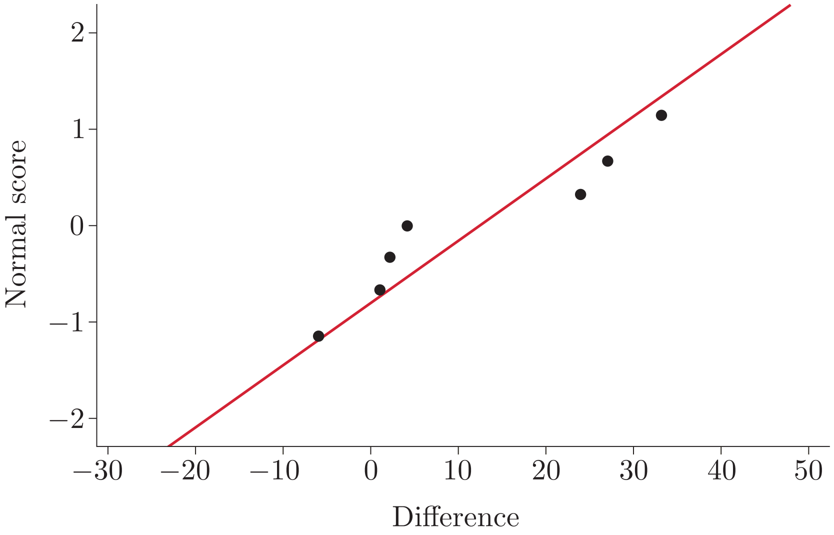


Figure 4 Normal probability plot for CO transfer factor differences

The solution to Activity 16 failed to come to a satisfactory conclusion as to whether the assumption of normality is plausible or not, so we are unsure of the validity of the t-test. In such cases, it is sensible to consider both tests: if both tests give the same conclusion, then you can be fairly confident that you have come to the right conclusion. In the case being considered here, neither test provided convincing evidence to suggest that there is any difference in CO transfer factor levels on entry and one week later.

5 The teleportation parameter

Computer scientists are interested in how many web links are followed by individuals surfing the internet. One parameter of interest is the so-called 'teleportation parameter', α, which can be defined as the probability that someone follows up on the information given on a web page by clicking on one of the links on that page.

An article considered, amongst other things, information collected automatically on web-user behaviour which gave a dataset of proportions of websites with links followed by each user. (Source: Gleich, D.F. et al. (2010) 'Tracking the random surfer: empirically measured teleportation parameters in PageRank', *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pp. 381–90.) As the data are collected automatically, the sample size is very large: $n = 257\,664$. Using these data, the authors of the cited work modelled the distribution of the proportion of web page visits from which a link was followed, X say, by the distribution with probability density function (p.d.f.) of the form

$$f(x) = 12x^2(1-x), \quad 0 < x < 1. \tag{1}$$

Here, the range of X is determined by the fact that it is a proportion. This p.d.f. is shown in Figure 5.

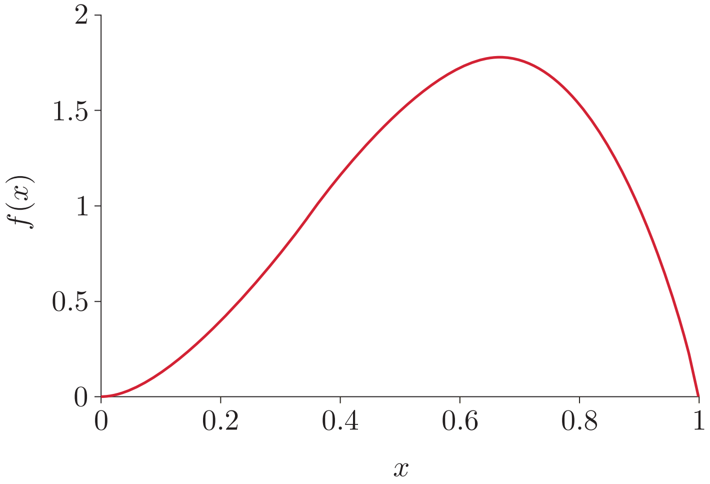
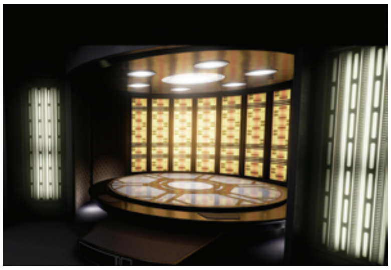


Figure 5 The p.d.f. $f(x) = 12x^2(1-x)$ on $0 < x < 1$



Teleportation is really the theoretical notion of transferring matter from one place to another without going through the space between them; this is the fictional *Star Trek* transporter/teleporter!

The teleportation parameter, α, can be defined as the mean of this distribution. In this section, you will use this model to calculate several properties of the distribution of proportions of websites with links followed, including the value of α associated with this model. The integration techniques that you will need in this section were reviewed in Subsection 3.1 of Unit 2.

The first task is to check that the claimed p.d.f., $f(x)$, really is a valid p.d.f. This you will do in the next activity.

Activity 17 *Checking that f is a p.d.f.*

Show that $f(x)$ given by Equation (1) really is a p.d.f.

The requirements for a function to be a p.d.f. were given in Subsection 3.3 of Unit 2.

Now that we are sure that $f(x)$ given by Equation (1) really is a p.d.f., we can find its corresponding cumulative distribution function (c.d.f.). This you will do in the next activity.

Activity 18 *Finding the c.d.f.*

Find the formula for the c.d.f. $F(x)$ when the p.d.f. $f(x)$ is given by Equation (1).

The c.d.f. of a continuous distribution was discussed in Subsection 4.3 of Unit 2.

The c.d.f. of a distribution can be used to find probabilities connected with that model. Recall that X is the proportion of web page visits from which a link is followed. Find the probabilities required in the following activity by using the c.d.f. associated with the p.d.f. in Equation (1), which was found in the solution to Activity 18 to be

$$F(x) = x^3(4 - 3x), \quad 0 < x < 1. \tag{2}$$

Activity 19 *Finding probabilities*

Use the c.d.f. $F(x)$ given by Equation (2) to calculate the following probabilities.

(a) $P(X < 0.5)$

(b) $P(0.3 \leq X \leq 0.7)$

(c) The probability that the proportion of web page visits from which a link was followed is at least 0.6.

Now, the teleportation parameter itself, α, is defined as the mean of the distribution of proportions of website visits from which a link is followed, that is, $\alpha = E(X)$. In the next activity, you will find the value of α for the model given by Equation (1).

Activity 20 *The value of the teleportation parameter*

The mean of a continuous distribution was covered in Section 2 of Unit 4.

Calculate the value of the teleportation parameter $\alpha = E(X)$ for the model given by Equation (1).

In addition to the mean of the distribution whose p.d.f. is given by Equation (1), you can also evaluate some measures of the spread of the distribution; this is the topic of Activity 21.

Activity 21 *Measures of spread*

The variance and standard deviation of a continuous distribution were covered in Section 4 and especially Subsection 5.2 of Unit 4.

For the model given by Equation (1), calculate the value of:

(a) the variance, $V(X)$

(b) the standard deviation, $S(X)$.

You should make use of the value $E(X) = 0.6$ that you found for this model in Activity 20.

Remember that, in this section, we have been using a model obtained, from data, by others in order to make some statements concerning the distribution of the proportions of website visits for which a link on that web page is followed. We did this assuming that the model those researchers obtained is a good representation of the practical situation. In addition to using the models that they or others come up with, statisticians spend a lot of their time producing, fitting, checking and refining the details of appropriate models for the data in the first place.

6 Daily steps

Along with the recent development of smartphones and smartwatches – but not, at the time of writing, teleportation devices! – there has been a surge in electronic devices and software for tracking personal activity. These range from specialist fitness tracking devices, many of which look rather like watches, to software which can be installed on smartwatches or smartphones.

One of the fitness trackers on the market in 2016

The dataset used in this section comprises some data collected by the author of this section about herself! As such, it has been written in the first person.

I don't have a special fitness tracking device, but I do have software on my phone which counts my steps. Obviously, steps are counted only when I carry my phone. I have extracted the data from my phone of the daily steps for 51 days between 1 October and 30 November 2015. This was a period when I regularly carried my phone for most of the day, although there were some days when I didn't carry my phone, resulting in very few or no steps being recorded for those days. I have excluded the data for such days from the analysis, since these very low counts would be unrepresentative of my usual number of daily steps.

Activity 22 *Histogram of the data*

Figure 6 shows a frequency histogram of daily steps on 51 days.

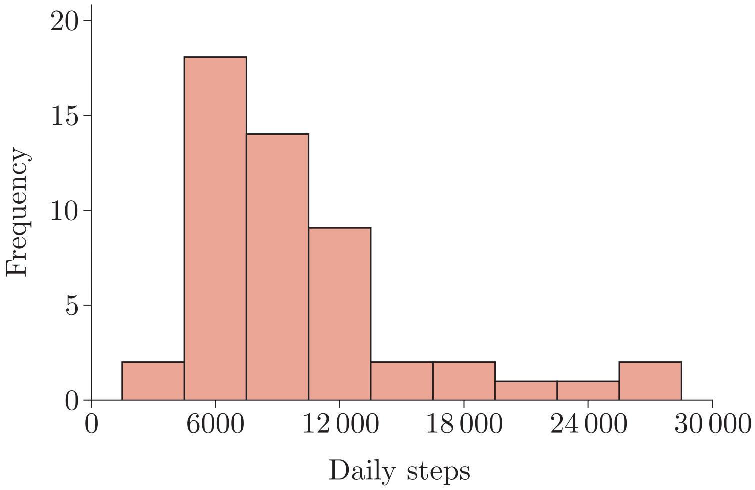


Figure 6 A histogram of daily steps

(Source: data provided by C.M. Queen, The Open University)

(a) Briefly describe the shape of the histogram.

(b) I would like to calculate a confidence interval to give a plausible range of values for my average number of daily steps. Which confidence interval would be more appropriate: a z-interval or a t-interval? Explain your answer.

z-intervals were considered in Subsection 1.2 of Unit 8, and t-intervals in Subsection 4.3 of that unit.

The solution to Activity 22 argued that a z-interval would be an appropriate confidence interval for my average number of daily steps. You will calculate the z-interval in the next activity.

Activity 23 *Confidence interval for the mean daily steps*

The sample mean number of steps for the 51 days with data was 9820, and the sample standard deviation was 5415. Each day's activity can be considered to be independent of any other day's activity.

(a) Calculate an approximate 95% confidence interval for my average number of daily steps.

(b) Fitness trackers usually set a default goal of 10 000 steps per day. Can I conclude that, on average, I am meeting the daily goal of 10 000 steps?

You saw in the solution to Activity 23 that it's plausible that, on average, I meet the goal of 10 000 steps per day. (Hooray!) However, as is clear from Figure 6, I don't achieve this goal every day. I would, however, like to meet the daily goal at least half of the time. For the 51 daily step counts shown in the histogram of Figure 6, the step count was at least 10 000 on 22 of the days: this is less than half of the daily counts. (Oh dear!) Does this mean that the proportion of days that my daily steps meet the goal of 10 000 is actually less than 0.5? A hypothesis test will be used to test this in the next activity.

Activity 24 *Proportion of days goal not met*

(a) Letting p be the proportion of days that I meet the daily goal of 10 000 steps, specify suitable null and alternative hypotheses for testing whether this proportion is less than 0.5.

(b) Given that I met the goal of 10 000 steps on 22 of the 51 days, use a p-value to carry out a large-sample test of the hypotheses that you specified in part (a). State your conclusions clearly.

Testing a proportion was covered in Subsection 3.2 of Unit 9, and p-values were covered in Section 4 of that unit.

Well, it's a bit of a relief that the solution to Activity 24 concluded that the data do not suggest that the proportion of days that I meet the daily goal of 10 000 steps is less than 0.5. However, this analysis has shown me that I am actually less active than I thought I was.

Since writing this analysis (in December 2016), I have made a conscious effort to be more active, and I recorded my daily steps for 1–31 January 2017.

Activity 25 *Comparing daily step data*

Figure 7 shows a comparative boxplot for my daily steps in October–November 2015 and in January 2017.

Comparative boxplots were covered in Subsection 5.3 of Unit 1.

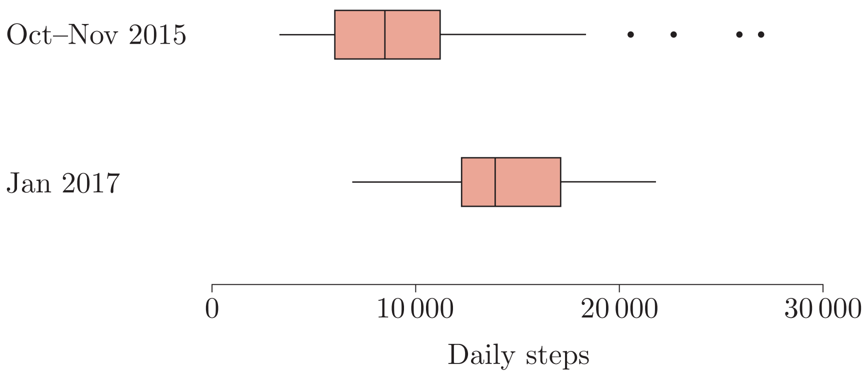


Figure 7 A comparative boxplot of daily steps

Use the comparative boxplot to compare the daily steps for the two time periods.

From the solution to Activity 25, it looks like I have indeed become more active. (Hooray!) Since I have made a conscious effort to be more active, I am confident that the average number of daily steps is greater than 10 000. However, I have now become more ambitious and I would like to test whether my average number of daily steps is actually now greater than 12 000. You will investigate this test problem in the next activity.

Activity 26 *More than 12 000 daily steps on average?*

Figure 8 shows a histogram of daily steps for the 31 days in January 2017.

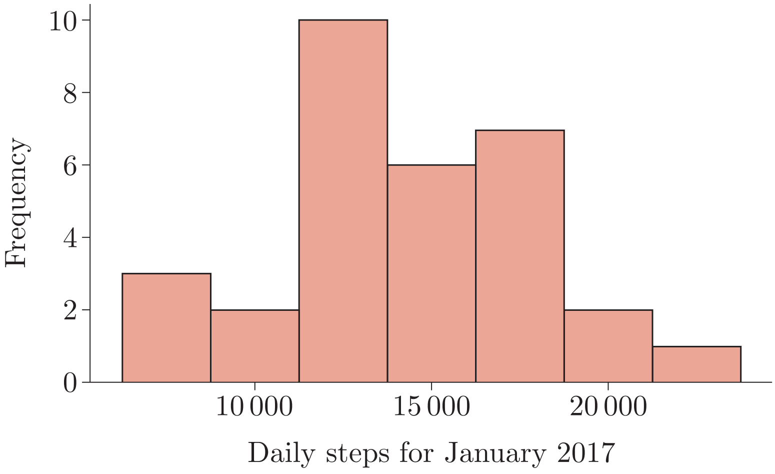


Figure 8 A histogram of daily steps for January 2017

The t-test and z-test were covered in Subsection 3.1 of Unit 9, while Wilcoxon's signed rank test was covered in Subsection 1.1 of Unit 10.

(a) In my view, the histogram in Figure 8 and the comparative boxplot in Figure 7 suggest that neither the t-test nor Wilcoxon's signed rank test may be appropriate for testing whether my average number of daily steps is now greater than 12 000. What aspect of the distribution of the data as portrayed in Figures 7 and 8 have I perceived to lead me to this conclusion? Why would I be content to employ the z-test instead?

p-values were covered in Section 4 of Unit 9.

(b) The sample mean for the daily step data for January 2017 was 14 121 while the sample standard deviation was 3719. Use a p-value to carry out a z-test of the hypotheses

$$H_0 : \mu = 12\,000, \quad H_1 : \mu > 12\,000,$$

where μ is my (current) mean number of daily steps. State your conclusions from this test clearly.

(c) A colleague looked at the comparative boxplot in Figure 7 and the histogram in Figure 8, and took a different view: he thought that any skewness in the distribution of the data is sufficiently small that an assumption of normality of these data is justified. This implies that a t-test may be appropriate for testing whether my average number of daily steps is now greater than 12 000.

You should be able to obtain a range of possible values rather than an exact value for the p-value of this test.

Carry out a t-test of the hypotheses

$$H_0 : \mu = 12\,000, \quad H_1 : \mu > 12\,000,$$

where μ is my (current) mean number of daily steps. State your conclusions from this test clearly.

Activity 26 highlights a couple of points that bear repeating about statistical analysis:

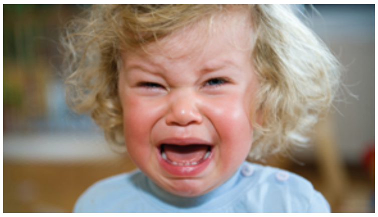
- there is an element of subjectivity in statistical modelling such that even experienced statisticians will sometimes make different modelling assumptions
- results of a statistical analysis often do not depend crucially on precise modelling assumptions, especially when, as above, the choice between different assumptions is somewhat marginal.

In addition, in the case of Activity 26, results using z- and t-tests are especially similar because the sample size, $n = 31$, is not too small. In this particular situation, the effect of different modelling assumptions on the outcome of interest is minor and would have remained so even if the data were more clearly non-normal.

Anyway, I'm delighted to see that there is strong evidence to suggest that my average number of daily steps is now greater than 12 000! And then, quite genuinely, in February 2017, this happened!

7 Expressed emotion

The *expressed emotion index* is a measure of the emotional climate of families with mentally ill members. Expressed emotion can be high or low: in families with high expressed emotion there is yelling, shouting, fighting, or critical or hostile comments. Studies suggest that patients living with



Toddlers frequently express emotion!

relatives scoring low on the expressed emotion index are less likely to relapse than those living with relatives who score high.

In a study of the relationship between expressed emotion and schizophrenia in Spain, a sample of 60 patients was followed up for two years after a psychiatric evaluation. One patient dropped out of the study after twelve months, leaving only 59.

At the initial evaluation, the families of the patients were scored on an expressed emotion scale. Families were then categorised as being either high expressed emotion families or low expressed emotion families. Table 6 shows the number of patients who relapsed during the two-year follow-up period for each of the groups of high and low expressed emotion families.

Table 6 Family expressed emotion and patient relapse

	Family expressed emotion	
	High	Low
Relapse	16	17
No relapse	12	14

(Source: Montero, I. et al. (1992) 'The influence of family expressed emotion on the course of schizophrenia in a sample of Spanish patients', *British Journal of Psychiatry*, vol. 161, no. 2, pp. 217–22)

We are interested both in the overall proportion of patients that relapse within two years of evaluation, and in the difference between the proportions relapsing in the two types of families. In Activity 27, you will calculate an approximate confidence interval to find a plausible range of values for the overall proportion of patients who relapse. You will then consider the difference between the two proportions in Activity 28.

Activity 27 *Overall proportion relapsing*

Confidence intervals for proportions were covered in Subsection 3.2 of Unit 8.

Estimate the overall proportion of patients that relapse, and calculate an approximate 95% confidence interval for this proportion. Is it plausible that half of all patients relapse?

Activity 28 *Difference between proportions relapsing*

(a) Calculate the observed proportion of patients relapsing in the high expressed emotion families, and the observed proportion relapsing in the low expressed emotion families.

(b) Calculate an approximate 95% confidence interval for the difference between the proportions of patients who relapse in the two types of families. What do you conclude from this interval about the effect of family expressed emotion on propensity to relapse in people with schizophrenia?

Confidence intervals for differences between two proportions were covered in Subsection 3.3 of Unit 8.

From the solution to Activity 28 we could not conclude from these data that the proportions of patients relapsing is not the same for the high and low expressed emotion families. However, as mentioned at the start of this section, other studies suggest that patients are less likely to relapse in families with low expressed emotion, which would lead to the proportion relapsing for low expressed emotion being lower than that for high expressed emotion families. Differences in conclusions between different studies do sometimes occur. This can be purely due to random variation, especially if the sample sizes are not very large – after all, one sample is unlikely to be exactly identical to another. But differences in conclusions may also be due to underlying differences between the studies. For example, perhaps the age of patients is relevant to the outcome, or perhaps the patient's gender, or some other factor, is relevant. Without further information, we cannot draw any further conclusions.

8 Norway spruce

The Norway spruce is a large evergreen coniferous tree native to Northern, Central and Eastern Europe. The data in Table 7 were collected as part of a study into how the heights and ages of trees are related. The heights (in feet) and the ages (in years) of 25 Norway spruce are given. In this section we will use linear regression to predict the height of a Norway spruce from its age.

Cones developing on a Norway spruce tree

Table 7 Age and height of Norway spruce

Age (years)	Height (feet)	Age (years)	Height (feet)
15	18	10	13
12	24	13	16
12	15	6	11
10	12	4	9
14	27	4	14
8	15	20	29
16	23	25	39
5	14	24	30
6	5	36	35
9	28	38	41
9	12	22	26
11	20	20	34
10	24		

Activity 29 *Response and explanatory variable*

If we are interested in predicting the height of a Norway spruce given its age, which is the response variable and which is the explanatory variable?

A scatterplot of height against age is shown in Figure 9.

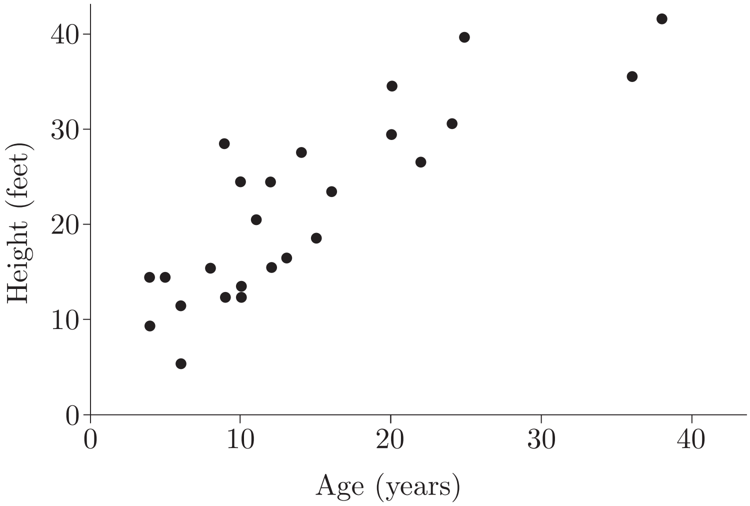


Figure 9 A scatterplot of height against age

The pattern in the plot suggests that it might be reasonable to fit a straight line to the data. You are asked to calculate the equation of the least squares line in Activity 30.

Activity 30 *The least squares line*

Suppose we wish to model the relationship between height Y and age x by a linear regression model. The summary statistics for the data in Table 7 are given by

$$n = 25, \quad \sum x_i = 359, \quad \sum y_i = 534,$$

$$\sum x_i^2 = 7135, \quad \sum x_i y_i = 9485.$$

Details of how to calculate the least squares line were given in Subsection 2.3 of Unit 11.

(a) Use the summary statistics to calculate S_{xx} and S_{xy}.

(b) Calculate the equation of the least squares line for the data.

In the solution to Activity 30, you saw that the fitted regression model for the data in Table 7 is

$$\text{height} = 8.18 + 0.92 \times \text{age}. \qquad (3)$$

After fitting a regression model like this, it is important to check that the model assumptions are reasonable. You will do this in the next activity.

Activity 31 *Checking the model assumptions*

A residual plot is shown in Figure 10(a) and a normal probability plot of the residuals is shown in Figure 10(b).

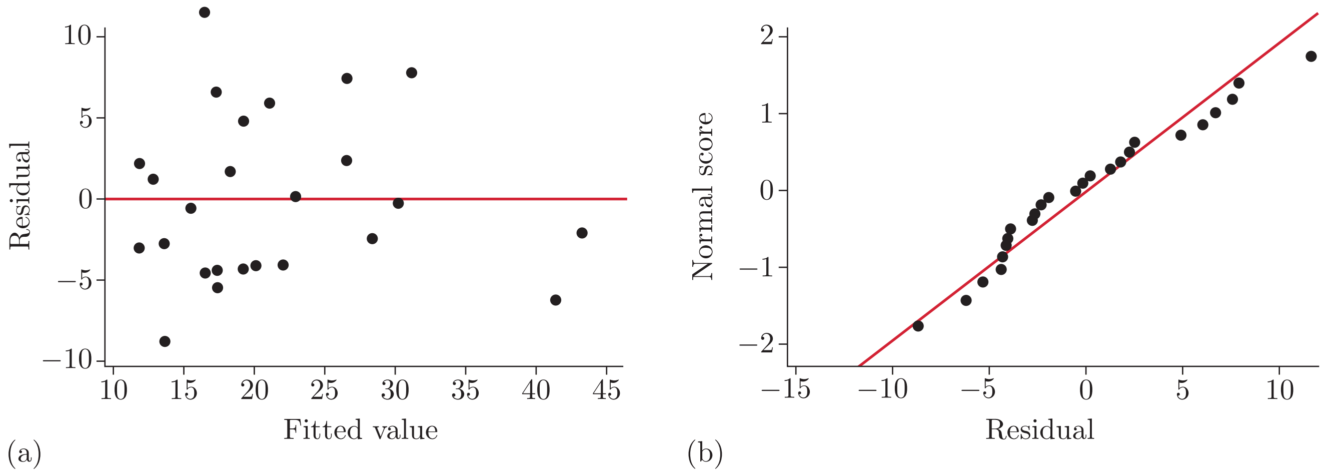


Figure 10 Checking the model assumptions: (a) residual plot; (b) normal probability plot

Comment on what these plots tell you.

Checking the linear regression model assumptions is covered in Section 3 of Unit 11.

The 'point' estimate of the slope parameter β in Equation (3) is 0.92. In the next activity, you will calculate a confidence interval for this parameter.

Activity 32 *A confidence interval for the slope β*

For the fitted regression model given in Equation (3), the residual sum of squares is given by

$$\sum (y_i - \widehat{y}_i)^2 = 626.580.$$

(a) Find an estimate of σ^2.

(b) Hence calculate a 95% confidence interval for the slope parameter β of the line.

How to find an estimate of σ^2 and a confidence interval for the slope parameter were covered in Subsections 4.1 and 4.3, respectively, of Unit 11.

One of the uses of a fitted regression model is to make predictions for specific values of the explanatory variable. You will use the model for this purpose next.

Activity 33 *Predicting the height*

According to the fitted regression model given in Equation (3), what is the predicted height of a 20-year-old Norway spruce?

In Activity 33, you used the least squares line for the data to predict the height of a 20-year-old Norway spruce. However, this doesn't give any indication of the uncertainty concerning the prediction. You will consider this uncertainty in the next activity.

Activity 34 *Confidence intervals and prediction intervals*

Confidence and prediction intervals in linear regression were covered in Subsection 4.3 of Unit 11.

(a) Calculate a 95% confidence interval for the mean height of 20-year-old Norway spruce trees.

(b) Calculate a 95% prediction interval for the height of an individual 20-year-old Norway spruce tree.

(c) Explain why, in the context of regression, a 95% prediction interval for the response must always be wider than a 95% confidence interval for the mean response at a specified value of the explanatory variable.

Activities 33 and 34 calculated a 'point' prediction and prediction interval, respectively, for a 20-year-old Norway spruce. So could the fitted model be used to predict the height of any Norway spruce given the tree's age? You will consider this question in relation to predicting the height of a particular Norway spruce in the next activity.

Activity 35 *Trafalgar Square Christmas tree*

The Trafalgar Square Christmas tree

Every Christmas since 1947, the city of Oslo, Norway, has given a Norway spruce tree as a gift to the people of Britain as a token of gratitude for British support for Norway in the Second World War. The tree is displayed in Trafalgar Square, London, and is decorated with hundreds of lights. The Trafalgar Square Christmas tree is typically 50–60 years old and over 20 metres (roughly $65\frac{1}{2}$ feet) tall.

(a) Suppose that the next Trafalgar Square Christmas tree is 58 years old. Explain why it may not be appropriate to use the fitted linear regression model for the data in Table 7 to predict the height of this tree.

(b) Now use the least squares line given by Equation (3) to make a point prediction of the height of a 58-year-old Norway spruce. Given what we said about the height of the Trafalgar Square Christmas tree, does the prediction surprise you?

As noted in the solution to Activity 35, it may not be appropriate to use the fitted linear regression model to predict the height for Norway spruce which are older than those in the sample used to fit the model. Equally, it may not be appropriate to use the model to predict the height of a Norway spruce which is younger than those in the sample. In particular, since the predicted height of a zero-year-old tree would be 8.18 feet according to the model, the model is clearly unreliable for small ages!

9 School performance

In England, school pupils aged 15–16 years take examinations giving GCSE (General Certificate of Secondary Education) or equivalent qualifications in a number of subjects. As well as their obvious importance to the pupils taking the exams, the results of these qualifications are used to measure school performance. In this section, we will consider two performance measures for schools in England which were introduced in 2016, called 'Attainment 8' and 'Progress 8'.

GCSEs with a numerical grading system have since been introduced in England.

Attainment 8 measures the achievement of 15–16-year-old pupils across 8 GCSE or equivalent qualifications including mathematics and English (both double-weighted), 3 further qualifications in specific subjects (namely, science subjects, computer science, history, geography and languages), and 3 further qualifications from a list of Department for Education approved qualifications. For the data considered here, each individual qualification is graded by a single letter, and each grade corresponds to a numerical 'score'. A school's Attainment 8 is calculated as the average of these grade scores across all pupils using their best 8 qualifications satisfying the subject criteria described above. Schools can calculate their Attainment 8 from their pupils' results.

Progress 8 is a type of value-added measure. Each individual pupil's Attainment 8 score is compared with the average Attainment 8 score of all pupils who had similar prior achievement: the higher the Progress 8 score, the greater the progress made by the pupil in comparison to other pupils with similar prior achievement. The school's Progress 8 score is the average of the individual pupil Progress 8 scores. Individual schools do not have the data available to be able to calculate their Progress 8 score, and the Progress 8 scores for all schools are calculated by England's Department for Education and are published around 3 months after the GCSE results are published.

In this section, we will investigate whether it is possible to use multiple regression to predict a school's Progress 8 score (the response variable Y) using two explanatory variables. We will use the Attainment 8 score as the first explanatory variable x_1, and the second explanatory variable, x_2, is the prior attainment of the cohort of 15–16-year-old pupils in each school as measured by the average point score obtained by these pupils at age 10–11 in national tests (the 'Key Stage 2 SATs').

The data used here were obtained from the Department for Education in England. (Source: https://www.compare-school-performance.service.gov.uk/download-data, data downloaded in February 2017.) We will consider data for 2016 for three types of state school:

- selective schools, which select their pupils based on academic achievement or aptitude
- comprehensive schools, which do not select their pupils based on academic achievement or aptitude
- secondary modern schools, which also do not select their pupils based on academic achievement or aptitude, but are usually found in areas where there are selective state schools – as such, secondary modern schools generally have a lower average prior attainment than comprehensive schools because some of the pupils with high prior attainment attend the local selective school instead.

Some schools do not have data available for all three variables (for example, the results for small schools are not published because individual pupils could be identified), and these schools have been excluded from the analysis.

Activity 36 *Scatterplots*

Figures 11 and 12 show scatterplots of the response Y (Progress 8) against x_1 (Attainment 8) and x_2 (attainment at age 10–11), respectively, for all three types of school. In each scatterplot, different colours are used for data points relating to the three different types of school.

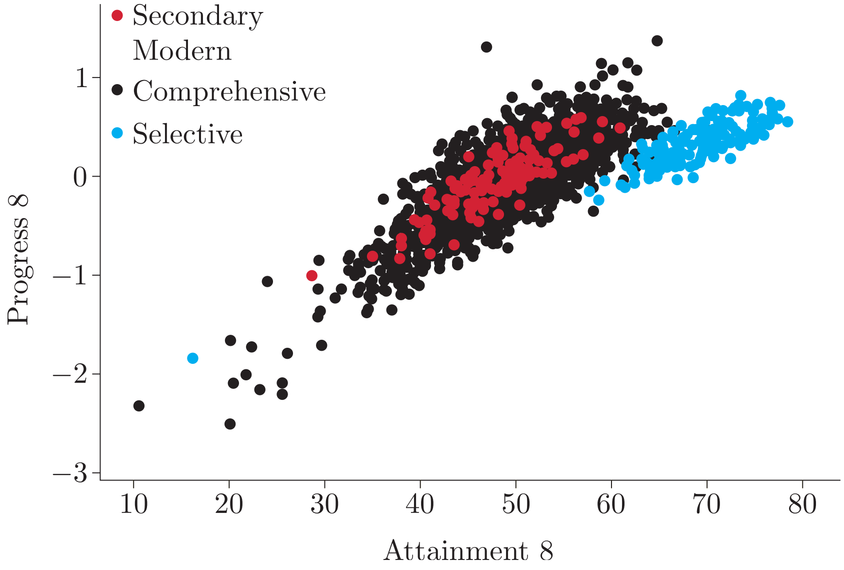


Figure 11 Scatterplot of Y (Progress 8) against x_1 (Attainment 8)

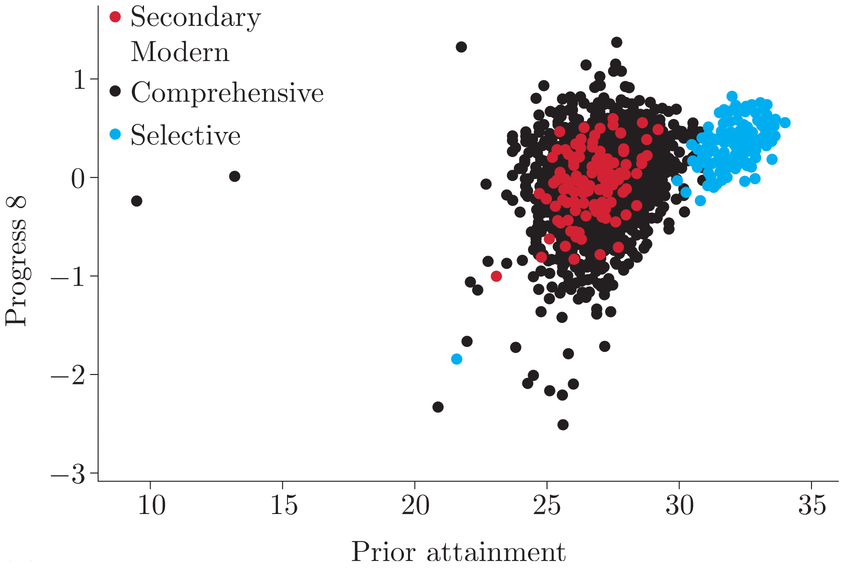


Figure 12 Scatterplot of Y against x_2 (prior attainment)

Explain why fitting a single multiple regression model using data from all three types of school wouldn't be appropriate.

You saw in the solution to Activity 36 that it would not be appropriate to use a single multiple regression model for all three types of school. We could, however, try fitting separate multiple regression models for the different school types. We will start by considering selective schools only: there are 164 such schools in the dataset.

Figure 13 shows scatterplots of the response Y (Progress 8) against x_1 (Attainment 8) and x_2 (prior attainment) for selective schools.

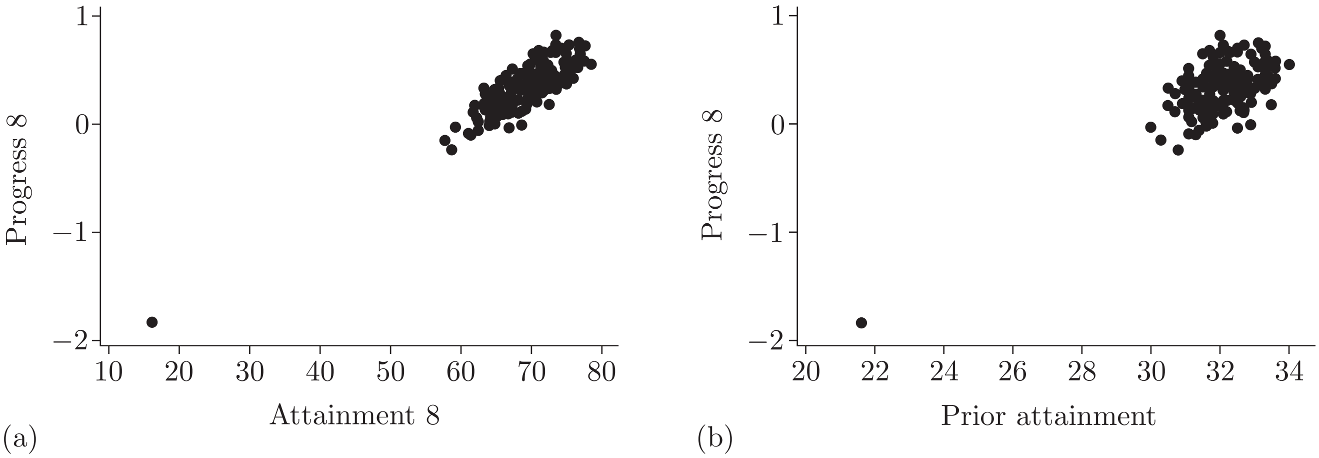


Figure 13 Scatterplots of (a) Y (Progress 8) against x_1 (Attainment 8), and (b) Y against x_2 (prior attainment), for selective schools only

Dominating both plots is an outlier in the bottom left-hand corner of each: this is a school with very low scores for each of Y, x_1 and x_2 in comparison to the other selective schools. Given that this school is so unlike the other selective schools, we will drop this outlier from the analysis in case this single anomalous data point influences the model unduly. (There turns out to be a good reason why this school is so unlike the others so that, unlike in Subsection 3.3 of Unit 12, it does not seem necessary to report results both with and without this school.)

Multiple regression was covered in Section 5 of Unit 11.

Some of the Minitab output when fitting the multiple regression model to the data for the remaining 163 selective schools is as follows.

```
Coefficients

Term                   Coef  SE Coef  T-Value  P-Value   VIF
Constant              5.104    0.139    36.68    0.000
Attainment 8        0.09831  0.00122    80.61    0.000  5.14
Prior attainment   -0.36004  0.00653   -55.16    0.000  5.14

Regression Equation

Progress 8 = 5.104 + 0.09831 Attainment 8 - 0.36004 Prior attainment
```

You will interpret what this output tells you about the multiple regression model in the next activity.

Activity 37 *Multiple regression for selective schools*

(a) Use the Minitab output to write down the fitted multiple regression model for selective schools in terms of x_1, x_2 and Y.

(b) Explain why this analysis suggests that Attainment 8 and prior attainment together influence Progress 8.

(c) Interpret the regression coefficients.

We now have a fitted multiple regression model and we established in Activity 37 that both explanatory variables have regression coefficients which are not zero and therefore should remain in the model. Next the model will be used to predict the Progress 8 scores for several schools.

Activity 38 *Fitted values and residuals*

(a) The 78th school in the dataset achieved a Progress 8 score of 0.00, which means that, on average, the attainment of pupils at this school is as expected compared with pupils of similar prior attainment. The Attainment 8 score was 64.8 and the prior attainment for this school was 31.8. Calculate the fitted value of Progress 8 for this school and its associated residual. Did the model over- or under-predict the Progress 8 score for this school?

(b) The first school in the dataset achieved the highest Attainment 8 score of 78.5 and also had the highest prior attainment of 34.0. The Progress 8 score was 0.55. How well did the model predict the Progress 8 score for this school?

(c) The Attainment 8 score for the seventh school in the dataset is 69.4, which is the same as the median Attainment 8 score across selective schools. The prior attainment and Progress 8 scores for this school were 32.2 and 0.35, respectively. Calculate the residual for this school and comment on your result.

Of course, exam results can't measure *all* student achievement. Here, school pupils are on an expedition associated with the Duke of Edinburgh award scheme.

In Activity 39, you will consider the residual plot and normal probability plot of residuals for the fitted model to decide whether the model assumptions seem reasonable.

Activity 39 *Are the model assumptions reasonable?*

Figure 14 shows the residual plot and normal probability plot of residuals for the fitted model for selective schools.

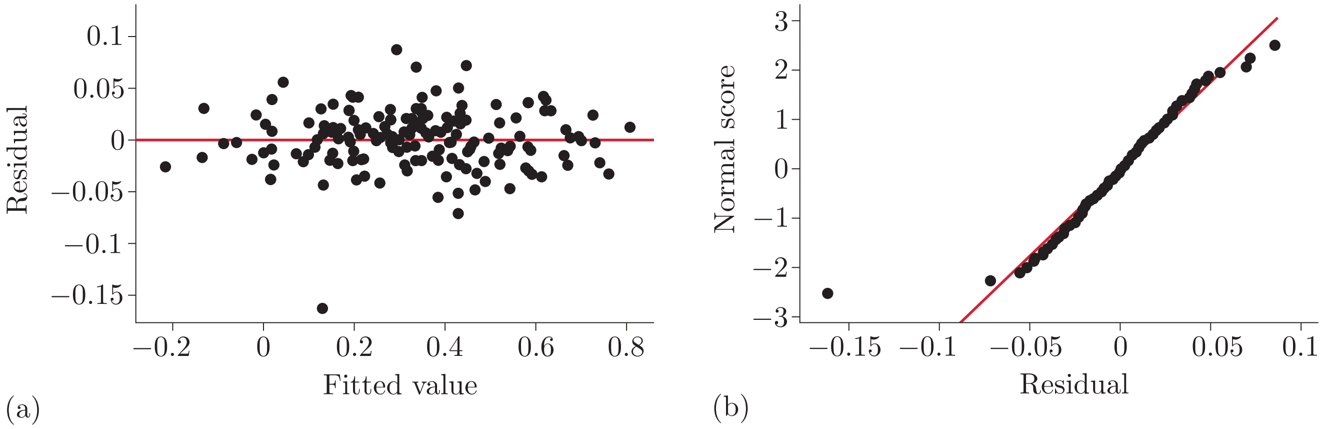


Figure 14 Checking the model assumptions: (a) residual plot; (b) normal probability plot of residuals

Do the model assumptions seem reasonable?

You have seen in Activities 37–39 how multiple regression can be used to model Progress 8 score using the Attainment 8 score and prior attainment as explanatory variables for selective schools. What about comprehensive schools and secondary modern schools? Do similar multiple regression models work well for those types of school as well?

Part of the Minitab output when a multiple regression model is fitted for the 2791 comprehensive schools in the dataset is as follows.

```
Coefficients

Term                    Coef   SE Coef  T-Value  P-Value   VIF
Constant              1.6982    0.0538    31.58    0.000
Attainment 8        0.088236  0.000569   155.11    0.000  2.17
Prior attainment    -0.22425   0.00260   -86.16    0.000  2.17

Regression Equation

Progress 8 = 1.6982 + 0.088236 Attainment 8 - 0.22425 Prior attainment
```

For the 117 secondary modern schools in the dataset, the corresponding Minitab output is as follows.

```
Coefficients

Term                    Coef  SE Coef  T-Value  P-Value   VIF
Constant               2.826    0.129    21.90    0.000
Attainment 8         0.09898  0.00127    77.95    0.000  2.82
Prior attainment    -0.28432  0.00645   -44.07    0.000  2.82

Regression Equation

Progress 8 = 2.826 + 0.09898 Attainment 8 - 0.28432 Prior attainment
```

You will consider the models for the comprehensive and secondary modern schools in the next activity.

Activity 40 *Fitted models for comprehensive and secondary modern schools*

Comment on the fitted multiple regression models for comprehensive and secondary modern schools in comparison to the fitted model for selective schools. In which ways are the models similar, and in which ways are they different?

Despite the similarities between the fitted models for comprehensive and secondary modern schools that were observed in the solution to Activity 40, there is actually a big difference between them: residual plots and probability plots of residuals (not shown) show that while the multiple regression model fits well for the secondary modern dataset, it does not for the comprehensive schools. We will therefore proceed, in the next activity, to consider further the fitted multiple regression models only for the secondary modern and selective schools.

Now, as you saw in the solution to Activity 40, the fitted multiple regression model for secondary modern schools is similar, although not identical, to that for selective schools. This means that schools with the same or similar Attainment 8 and prior attainment scores could have *different* predicted Progress 8 scores depending on the type of school. This is illustrated in the next activity.

Activity 41 *Predicting Progress 8 scores: secondary modern and selective schools*

The fitted multiple regression models for secondary modern and selective schools are as follows.

Secondary modern schools:

$$y = 2.826 + 0.09898\,x_1 - 0.28432\,x_2.$$

Selective schools:

$$y = 5.104 + 0.09831\,x_1 - 0.36004\,x_2.$$

(a) One of the secondary modern schools has a prior attainment score of 29.2. The Attainment 8 score for this school is 60.9, and the predicted Progress 8 score is approximately 0.552. Calculate the predicted Progress 8 score for a selective school with the same prior attainment and Attainment 8 scores.

(b) One of the selective schools has a prior attainment score of 30.0. The Attainment 8 score for this school is 59.3, and the predicted Progress 8 score is 0.133. Calculate the predicted Progress 8 score for a secondary modern school with the same prior attainment and Attainment 8 scores.

(c) Comment on your results.

The situation in which there is a slightly different model for the same variables under different circumstances (in our case here, different types of school) arises frequently in practice. We coped with it here by fitting three separate models for the three different types of school. There is, however, a neater way of doing things by using a single model which allows for differing parameter values for the different types of school by treating the type of school as a so-called factor. This model is beyond the scope of this module, but is something that you are likely to meet if you go on to study regression further.

10 And finally . . .

So this is the end of the unit, and indeed the end of the module! We very much hope that you have enjoyed it. The module has explored the fundamental statistical techniques and ideas used for analysing and interpreting data, and you should now have the basic skills required to start to make sense of data. The module also provided the necessary foundations required for studying statistics further if you wish to do so. In this age of data, there are a huge number of exciting datasets out there ready to be gathered and explored, and we hope that we have enthused you to do so!

Summary

In this unit, you have used various statistical methods developed in the module to explore several datasets. You have:

- considered various probability models for datasets, including the discrete uniform, Poisson, exponential and normal models, and one with a particular polynomial p.d.f., as well as the Poisson process
- derived a likelihood function and the associated maximum likelihood estimate
- calculated several confidence intervals, including a z-interval and confidence intervals for a proportion and the difference between two proportions
- carried out several tests, including the t-test, Wilcoxon signed rank test, chi-squared goodness-of-fit test and testing a proportion
- used linear regression, including obtaining a least squares line, checking model assumptions, calculating confidence and prediction intervals, and using multiple regression.

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate the wide variety of possible applications that the techniques developed in the module can be applied to
- be more confident in your application of the techniques developed in the module
- appreciate that there may be uncertainty as to the most suitable statistical technique to use when there is ambiguity over the validity of normality or other model assumptions
- appreciate that in the real world, a statistical analysis doesn't always give a clear-cut result
- appreciate that in the real world, it may not be possible to fully answer the question of interest with the data available.

Solutions to activities

Solution to Activity 1

(a) Each chocolate bar can be one of 7 types coded $1, 2, \dots, 7$. The range of X is therefore $\{1, 2, 3, 4, 5, 6, 7\}$.

(b) If it is reasonable to expect equal numbers of each type of chocolate bar, then, since the range of X is $\{1, 2, 3, 4, 5, 6, 7\}$, a suitable distribution is the discrete uniform distribution with parameters $m = 1$ and $n = 7$. Therefore, from Equations (7) and (8) in Unit 3, respectively, X has probability mass function

$$p(x) = \frac{1}{7-1+1} = \frac{1}{7}, \quad x = 1, 2, \dots, 7,$$

and cumulative distribution function

$$F(x) = \frac{x-1+1}{7-1+1} = \frac{x}{7}, \quad x = 1, 2, \dots, 7.$$

Solution to Activity 2

(a) From the solution to Activity 1(b),

$$p(x) = \frac{1}{7}, \quad x = 1, 2, \dots, 7.$$

So in a tub of 71 chocolate bars, the expected frequency of each type of chocolate bar is

$$E_i = 71 \times \frac{1}{7} \simeq 10.14, \quad i = 1, 2, \dots, 7.$$

(b) The chi-squared goodness-of-fit test uses the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where k is the number of categories, which in this case is 7. So the observed value of the test statistic is

$$\begin{aligned}\chi^2 &= \frac{(17-10.14)^2}{10.14} + \frac{(10-10.14)^2}{10.14} + \frac{(12-10.14)^2}{10.14} + \frac{(15-10.14)^2}{10.14} \\ &\quad + \frac{(5-10.14)^2}{10.14} + \frac{(4-10.14)^2}{10.14} + \frac{(8-10.14)^2}{10.14} \\ &\simeq 4.641 + 0.002 + 0.341 + 2.329 + 2.605 + 3.718 + 0.452 \\ &= 14.088 \simeq 14.09.\end{aligned}$$

If the discrete uniform model is correct, then the distribution of the test statistic is approximately $\chi^2(k - p - 1) = \chi^2(6)$, since k is the number of categories (in this case 7) and p is the number of estimated parameters (in this case 0, since the parameters of this discrete uniform distribution are known).

Comparing the observed value of 14.09 with the quantiles of the $\chi^2(6)$ distribution, from the table of chi-squared distribution quantiles in the Handbook, we see that the test statistic lies between the 0.95-quantile (which is 12.59) and the 0.975-quantile (which is 14.45). So the p-value lies between 0.025 and 0.05. (The exact p-value happens to be 0.029, but you cannot work this out from the table, nor do you need to.) There is therefore moderate evidence against the null hypothesis that the discrete uniform distribution with parameters $m = 1$ and $n = 7$ is a suitable model for X.

In the interpretational terms of Table 3 of Unit 9, it is therefore also the case that $0.01 < p \leq 0.05$.

Solution to Activity 3

(a) A suitable graph is a bar chart such as that in Figure 15. (This is because the data are discrete.) It is easy to draw by hand.

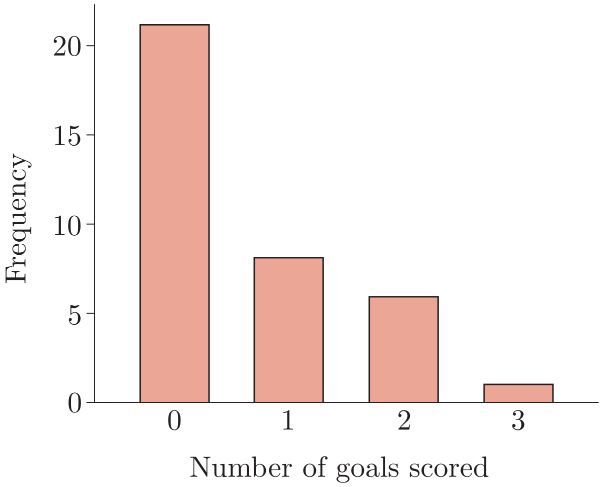


Figure 15 A bar chart for goals scored

(b) A Poisson model seems appropriate. Reasons include the following:

- the possible values for X are counts
- the range of X starts at 0 and doesn't have any theoretical upper value
- the bar chart of the data has a shape which is consistent with a Poisson distribution with a small mean; it is decreasing and right-skew.

Solution to Activity 4

(a) From Equation (6) of Unit 3, the Poisson p.m.f. is

$$p(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}.$$

So, following Example 17 of Unit 7, the likelihood is

$$\begin{aligned}L(\theta) &= p(0;\theta)^{21} \times p(1;\theta)^{8} \times p(2;\theta)^{6} \times p(3;\theta)^{1}\\ &= \left(\frac{e^{-\theta}\theta^0}{0!}\right)^{21} \times \left(\frac{e^{-\theta}\theta^1}{1!}\right)^{8} \times \left(\frac{e^{-\theta}\theta^2}{2!}\right)^{6} \times \frac{e^{-\theta}\theta^3}{3!}\\ &= e^{-21\theta} \times e^{-8\theta}\theta^8 \times \frac{e^{-6\theta}\theta^{12}}{2^6} \times \frac{e^{-\theta}\theta^3}{6}\\ &= \frac{e^{-36\theta}\theta^{23}}{384},\end{aligned}$$

as required.

(b) Using Equation (9) of Unit 7 to differentiate a product, we have

$$\begin{aligned}L'(\theta) &= \frac{1}{384}\left(-36e^{-36\theta} \times \theta^{23} + e^{-36\theta} \times 23\theta^{22}\right)\\ &= \frac{e^{-36\theta}\theta^{22}}{384}(-36\,\theta + 23).\end{aligned}$$

To find $\widehat{\theta}$, we now need to solve $L'(\theta) = 0$. All but the linear term in brackets is irrelevant to solving $L'(\theta) = 0$ because, for $\theta > 0$, $e^{-36\theta}\theta^{22}/384 > 0$. The linear term has a single value of θ at which it is zero, so that value must be the MLE $\widehat{\theta}$: $\widehat{\theta}$ satisfies

$$-36\,\theta + 23 = 0,$$

so

$$\widehat{\theta} = \frac{23}{36} \simeq 0.639.$$

Solution to Activity 5

(a) The Poisson(0.639) p.m.f. is

$$p(x) = \frac{e^{-0.639}0.639^x}{x!}.$$

So

$$p(X = 0) = \frac{e^{-0.639}0.639^0}{0!} \simeq 0.5278,$$

and the expected frequency of scoring 0 goals in 36 scores, E_0, is, to two decimal places,

$$E_0 \simeq 36 \times 0.5278 \simeq 19.00.$$

Similarly,

$$p(X = 1) = \frac{e^{-0.639}0.639^1}{1!} \simeq 0.3373,$$

and the expected frequency of scoring 1 goal in 36 scores, E_1, is, to two decimal places,

$$E_1 \simeq 36 \times 0.3373 \simeq 12.14.$$

Finally,

$$\begin{aligned} P(X \geq 2) &= 1 - \{P(X = 0) + P(X = 1)\} \\ &= 1 - e^{-0.639} - 0.639e^{-0.639} \simeq 0.1349. \end{aligned}$$

So the expected frequency of scoring ≥ 2 goals in 36 scores, E_2, is, to two decimal places,

$$E_2 \simeq 36 \times 0.1349 \simeq 4.86.$$

(b) For the chi-squared goodness-of-fit test to be valid, a rough rule of thumb is that the expected frequencies for each category need to be 5 or more. The expected frequency for the category '≥ 2 goals' is 4.86, so if there were separate categories '2 goals', '3 goals', and so on, the expected frequencies for these categories would be too small. As it is, the expected frequency is just under the rule of thumb value of 5. However, as this is a rule of thumb and the expected frequency is only just under 5, the test should be just about valid. (Further combining categories into just two groups would not work; we will mention why at the end of the solution to the next part of this activity.)

(c) The chi-squared goodness-of-fit test uses the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where k is the number of categories. So the observed value of the test statistic is

$$\begin{aligned} \chi^2 &= \frac{(21 - 19.00)^2}{19.00} + \frac{(8 - 12.14)^2}{12.14} + \frac{(7 - 4.86)^2}{4.86} \\ &\simeq 0.211 + 1.412 + 0.942 = 2.565 \simeq 2.57. \end{aligned}$$

If the Poisson model is correct, then the distribution of the test statistic is approximately $\chi^2(k - p - 1) = \chi^2(1)$, since k is the number of categories (in this case 3) and p is the number of estimated parameters (in this case 1, since the parameter of the Poisson distribution has been estimated as 0.639).

We are using a 5% significance level, so the 0.95-quantile of the $\chi^2(1)$ distribution is required: from Table 18 in Unit 10 or the table in the Handbook, this is 3.84. Since the observed value of the test statistic is 2.56, which is less than the 0.95-quantile of $\chi^2(1)$, there isn't sufficient evidence to reject the hypothesis that Poisson(0.639) is a suitable model for the data in Table 2.

It seems that the Poisson model is a suitable model for the data in Table 2.

(Had we combined categories into just $k = 2$ groups, the number of degrees of freedom for the χ^2 test would have reduced to $k - p - 1 = 2 - 1 - 1 = 0$. This reflects the fact that there would be too little data left to test goodness-of-fit appropriately.)

Solution to Activity 6

Major tsunamis can occur at any time, and the waiting time between tsunamis need not be a whole number of months. Therefore an exponential distribution, which is continuous, is more appropriate for modelling the waiting times than a geometric distribution, which is discrete.

Solution to Activity 7

An exponential model does seem reasonable because:

- the shape of the histogram is not inconsistent with the data coming from an exponential distribution: there is a single peak at 0 and the frequencies generally decrease with increasing waiting times
- the sample mean and sample standard deviation are close together in value.

Solution to Activity 8

Since the mean time between major tsunamis is 26 months, an estimate of the parameter λ is

$$\widehat{\lambda} = \frac{1}{26} \simeq 0.038.$$

(This is, in fact, the maximum likelihood estimate of λ.)

Solution to Activity 9

(a) There are 12 months in one year, so, letting X denote the waiting time between two successive major tsunamis,

$$\begin{aligned} P(\text{waiting time is at least one year}) &= P(X \geq 12) = 1 - F(12) \\ &= 1 - \left(1 - e^{-\frac{12}{26}}\right) = e^{-\frac{12}{26}} \\ &\simeq 0.630. \end{aligned}$$

(The exponential distribution's cumulative distribution function is given in Equation (2) of Unit 5.)

(b) $$\begin{aligned} P(\text{waiting time is less than 6 months}) &= P(X < 6) = F(6) \\ &= 1 - e^{-\frac{6}{26}} \simeq 0.206. \end{aligned}$$

(c) The expected number of waiting times of at least one year is

$$e^{-\frac{12}{26}} \times 29 \simeq 18.28,$$

and the expected number of waiting times of less than 6 months is

$$\left(1 - e^{-\frac{6}{26}}\right) \times 29 \simeq 5.98.$$

For the observed data, 16 were waiting times of at least one year, and 8 were waiting times of less than 6 months. The expected values from the model are therefore not totally out of line with the data.

Solution to Activity 10

(a) The estimate of the rate λ of occurrence of major tsunamis per month is

$$\lambda = \frac{1}{26} \simeq 0.038 \text{ per month.}$$

(This is the same value as you calculated in Activity 8.)

(b) X has a Poisson distribution with parameter

$$\lambda t = \left(\frac{1}{26} \text{ per month}\right) \times (12 \text{ months}) = \frac{12}{26} = \frac{6}{13}.$$

(c) (i) From Equation (6) of Unit 3, the probability that exactly two major tsunamis will occur in one year is given by

$$P(X = 2) = \frac{e^{-\frac{6}{13}}\left(\frac{6}{13}\right)^2}{2!} \simeq 0.067.$$

(ii) The probability that at least one major tsunami will occur in one year is given by

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\frac{6}{13}} \simeq 0.370.$$

Solution to Activity 11

(a) There are three assumptions being made:

- major tsunamis occur singly
- the rate of occurrence of major tsunamis remains constant
- the incidence of future major tsunamis is independent of the past.

(b) It is not impossible that more than one major tsunami may occur simultaneously, but the probability of such an event seems to be so small that this assumption may well be reasonable. (The coarseness of the discretisation of the data in Table 3 is such that more than one tsunami might happen within a period of one month. If so, something needs to be done about this, but it does not mean that a Poisson process on the underlying continuous scale is not still a good model.)

However, Figure 2 suggests that the rate at which major tsunamis are occurring may not be constant since the plot doesn't follow a very straight line. It appears from Figure 2 that major tsunamis seem to be occurring more frequently towards the end of the time period. However, it is possible that the rate at which major tsunamis occur is in fact constant, but that the recording of them has improved in recent years, leading to major tsunamis being recorded more frequently.

It seems reasonable, but not unarguable, that the incidence of future major tsunamis is independent of the past, so the third assumption for a Poisson process identified in the solution to part (a) seems appropriate.

Overall, there is some doubt about the reasonableness of a Poisson process as a model for the occurrence of major tsunamis.

Solution to Activity 12

(a) On entry, the median is

$$m = x_{\left(\frac{1}{2}(n+1)\right)} = x_{(4)} = 58.$$

The quartiles are

$$q_L = x_{\left(\frac{1}{4}(n+1)\right)} = x_{(2)} = 50,$$

$$q_U = x_{\left(\frac{3}{4}(n+1)\right)} = x_{(6)} = 62.$$

So the sample interquartile range on entry is

$$q_U - q_L = 62 - 50 = 12.$$

One week later, the median is

$$m = x_{(4)} = 64.$$

The quartiles are

$$q_L = x_{(2)} = 60,$$

$$q_U = x_{(6)} = 80.$$

So the sample interquartile range one week later is

$$q_U - q_L = 80 - 60 = 20.$$

(b) The measurements after one week are generally higher than on entry since the boxplot for this set of measurements is located to the right of the other boxplot. The spread is also greater one week later. (These two features are apparent from the numerical work of part (a) also.)

Solution to Activity 13

(a) The required differences are given in Table 8.

Table 8 Differences 'one week later' minus 'on entry'

Patient	1	2	3	4	5	6	7
Difference	33	2	24	27	4	1	−6

(b) The normality of the differences can be investigated using a normal probability plot. If the points lie close to a straight line, then the normality assumption is plausible.

(c) A (one-sample) t-test on the differences could be used if the assumption of a normal model is plausible. For this test, because we're interested in detecting a difference between the two measurements in either direction, the hypotheses to be tested are

$$H_0: \mu_D = 0, \quad H_1: \mu_D \neq 0,$$

where μ_D is the population mean of the differences.

(d) If a normal model is untenable, then the Wilcoxon signed rank test could be used. (Another potential test if a normal model is untenable is the z-test, but that would require a larger sample than we have here – the rule of thumb we have been using is that we would need $n \geq 25$.)

Again, we are interested in detecting a difference in either direction, but for this nonparametric test the hypotheses involve the population median differences m_D (rather than the population mean differences as considered in the t-test). The hypotheses to be tested are

$$H_0 : m_D = 0, \quad H_1 : m_D \neq 0.$$

Solution to Activity 14

(a) The value of the test statistic is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{12.14 - 0}{15.38/\sqrt{7}} \simeq 2.088.$$

(b) The null distribution of the t-test is $t(n-1)$, where n is the sample size. Therefore, for the data with sample size $n = 7$, the null distribution is $t(6)$.

(c) Since $0.05 < p < 0.1$, there is only weak evidence against H_0. We conclude that the data provide weak evidence to suggest that, on average, there is a difference between CO transfer factor levels on entry and one week later, and since t is positive, any difference is such that CO transfer factor levels are higher one week later than on entry.

Solution to Activity 15

(a) The Wilcoxon signed rank test statistic can be calculated using Table 9.

Table 9 Differences 'one week later' minus 'on entry'

Patient	1	2	3	4	5	6	7
Difference	33	2	24	27	4	1	−6
Sign of difference	+	+	+	+	+	+	−
Absolute value of difference	33	2	24	27	4	1	6
Rank of absolute value	7	2	5	6	3	1	4

The test statistic w_+ is the sum of the ranks associated with the positive differences, so

$$w_+ = 7 + 2 + 5 + 6 + 3 + 1 = 24.$$

(b) Since $p > 0.1$, there is little or no evidence to suggest that the median difference between CO transfer factor levels on entry and one week later is not zero. The data therefore suggest that there is little or no evidence that there is any difference in the CO transfer factor levels.

Solution to Activity 16

With only seven data points, it is very difficult to say whether the assumption of normality of the differences is plausible or not. The data do roughly lie about the straight line, so normality can't be ruled out, but equally, the points are certainly not so close to the line as to give overwhelming support to the assumption of normality either!

Solution to Activity 17

For a function f to be a valid p.d.f., it needs to be non-negative and to integrate to 1 over its range.

Non-negativity over its range is clear from Figure 5. Alternatively, notice that every element that is multiplied together to form f is itself non-negative for $0 < x < 1$.

As for its integral over its range, we have

$$\begin{aligned}\int_0^1 f(x)\,dx &= \int_0^1 12x^2(1-x)\,dx = 12\int_0^1 x^2(1-x)\,dx\\ &= 12\int_0^1 (x^2-x^3)\,dx = 12\left[\frac{x^3}{3}-\frac{x^4}{4}\right]_0^1\\ &= 12\left(\frac{1}{3}-\frac{1}{4}-(0-0)\right) = 12\times\frac{1}{12} = 1,\end{aligned}$$

as required.

Solution to Activity 18

For $0 < x < 1$, the c.d.f. is given by

$$\begin{aligned}F(x) &= \int_0^x f(y)\,dy = \int_0^x 12y^2(1-y)\,dy = 12\int_0^x (y^2-y^3)\,dy\\ &= 12\left[\frac{y^3}{3}-\frac{y^4}{4}\right]_0^x = 12\left(\frac{x^3}{3}-\frac{x^4}{4}-(0-0)\right)\\ &= 12x^3\left(\frac{1}{3}-\frac{x}{4}\right) = x^3(4-3x).\end{aligned}$$

Solution to Activity 19

(a) Because X is continuous,

$$P(X < 0.5) = P(X \leq 0.5) = F(0.5) = (0.5)^3(4-3\times 0.5) = 0.3125.$$

(b)
$$\begin{aligned}P(0.3 \leq X \leq 0.7) &= F(0.7) - F(0.3)\\ &= (0.7)^3(4-3\times 0.7) - (0.3)^3(4-3\times 0.3)\\ &= 0.6517 - 0.0837 = 0.568.\end{aligned}$$

(c) The required probability is $P(X \geq 0.6)$:

$$\begin{aligned}P(X \geq 0.6) &= 1 - P(X < 0.6) = 1 - F(0.6)\\ &= 1 - (0.6)^3(4-3\times 0.6) = 1 - 0.4752 = 0.5248.\end{aligned}$$

Solution to Activity 20

$$\begin{aligned}\alpha = E(X) &= \int_0^1 x\,f(x)\,dx = \int_0^1 12x^3(1-x)\,dx = 12\int_0^1 (x^3-x^4)\,dx\\ &= 12\left[\frac{x^4}{4}-\frac{x^5}{5}\right]_0^1 = 12\left(\frac{1}{4}-\frac{1}{5}-(0-0)\right)\\ &= 12\times\frac{1}{20} = \frac{3}{5} = 0.6.\end{aligned}$$

Solution to Activity 21

(a) We will use the relationship $V(X) = E(X^2) - \{E(X)\}^2$. To do so, we first need to calculate $E(X^2)$:

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 f(x)\,dx = \int_0^1 12x^4(1-x)\,dx = 12\int_0^1 (x^4 - x^5)\,dx \\ &= 12\left[\frac{x^5}{5} - \frac{x^6}{6}\right]_0^1 = 12\left(\frac{1}{5} - \frac{1}{6} - (0-0)\right) \\ &= 12 \times \frac{1}{30} = \frac{2}{5} = 0.4. \end{aligned}$$

Therefore

$$V(X) = E(X^2) - \{E(X)\}^2 = 0.4 - (0.6)^2 = 0.04.$$

(b) We will use the relationship $S(X) = \sqrt{V(X)}$:

$$S(X) = \sqrt{V(X)} = \sqrt{0.04} = 0.2.$$

Solution to Activity 22

(a) The histogram is unimodal with a single peak around 6000 steps. It is also right-skew, since the bars on the right-hand side of the histogram fall away more slowly than those on the left-hand side.

(b) In order to use a t-interval, the data need to be normally distributed. However, a normal model is not plausible for these data because they are right-skew.

On the other hand, a z-interval can be used for any data regardless of their distribution, as long as the sample size is large enough. The sample size of $n = 51$ is reasonably large, so a z-interval would be more appropriate for these data.

Solution to Activity 23

(a) From Equation (3) of Unit 8, an approximate 95% confidence interval for the mean is the z-interval given by

$$(\mu^-, \mu^+) = \left(\overline{x} - z\,\frac{s}{\sqrt{n}},\ \overline{x} + z\,\frac{s}{\sqrt{n}}\right),$$

where z is the 0.975-quantile of $N(0,1)$. So

$$\mu^- = 9820 - 1.96 \times \frac{5415}{\sqrt{51}} \simeq 8334,$$

$$\mu^+ = 9820 + 1.96 \times \frac{5415}{\sqrt{51}} \simeq 11\,306.$$

Thus an approximate 95% confidence interval for my mean number of daily steps is $(8334, 11\,306)$.

Because the numbers are in the 1000s, the confidence interval limits were rounded to the nearest whole number.

(b) The confidence interval calculated in part (a) contains the value 10 000, so it is one of the plausible values for my mean number of daily steps. Therefore it is plausible that, on average, I am indeed meeting the daily goal of 10 000 steps.

Solution to Activity 24

(a) The hypotheses to be tested are

$$H_0: p = 0.5, \quad H_1: p < 0.5.$$

(b) The sample size of $n = 51$ is reasonably large, so the hypotheses specified in part (a) can be tested using the test statistic

$$Z_p = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where the null distribution for Z_p is $N(0,1)$. For these data, $\widehat{p} = 22/51$ so

$$z_p = \frac{\frac{22}{51} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{51}}} \simeq -0.980.$$

This is a one-sided test, so the p-value is calculated as

$$p = P(Z_p \leq -0.980),$$

where $Z_p \sim N(0,1)$. So

$$\begin{aligned} p &= P(Z_p \leq -0.980) = P(Z_p \geq 0.980) = 1 - P(Z_p < 0.980) \\ &= 1 - \Phi(0.98) = 1 - 0.8365 = 0.1635. \end{aligned}$$

Since $p = 0.1635 > 0.1$, the p-value provides little or no evidence against H_0. We conclude that the data do not suggest that the proportion of days that I meet the goal of 10 000 steps is less than 0.5.

Solution to Activity 25

The daily steps for January 2017 are generally higher than those for October–November 2015 since the boxplot for January 2017 is located to the right of the other boxplot. The spread for January 2017 is generally similar to the spread for the first set of data, with the exception of four large outliers for the 2015 data.

Solution to Activity 26

(a) Both the histogram and the boxplot suggest to me that a symmetric model for the data may not be appropriate. However, symmetry of the underlying distribution is an assumption of Wilcoxon's signed rank test, while a normal model (and hence symmetry) is an assumption of the t-test. Thus neither of these tests may be appropriate for testing whether the average number of daily steps is now greater than 12 000.

The z-test doesn't have any such assumptions, and can therefore be used for these data because the sample size is large enough (i.e. ≥ 25 by the rule of thumb).

(b) The observed value of the z-test statistic is

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{14\,121 - 12\,000}{3719/\sqrt{31}} \simeq 3.175.$$

This is a one-sided test, so the p-value is calculated as

$$p = P(Z \geq 3.175),$$

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

$$\begin{aligned} p &= P(Z \geq 3.175) = 1 - P(Z < 3.175) = 1 - \Phi(3.175) \\ &\simeq 1 - \Phi(3.18) = 1 - 0.9993 = 0.0007. \end{aligned}$$

Since $p < 0.01$ there is strong evidence against H_0, so there is strong evidence that the average number of daily steps is greater than 12 000.

(c) The observed value of the t-test statistic is the same as the z-test statistic of part (b): $t \simeq 3.175$. But in this case the p-value is calculated as

$$p = P(t \geq 3.175),$$

where $T \sim t(30)$; here, the degrees of freedom parameter of the null distribution has been calculated as $n - 1 = 31 - 1 = 30$. Now, from the table of quantiles of the t-distribution in the Handbook, the observed value of t lies between the 0.995-quantile of the $t(30)$ distribution (which is 2.750) and the 0.999-quantile of the $t(30)$ distribution (which is 3.385). As we are conducting a one-sided test, the p-value lies between 0.001 and 0.005.

Both z- and t-test statistics have the formula $(\overline{x} - \mu_0)/(s/\sqrt{n})$.

Again, since $p < 0.01$ there is strong evidence against H_0, so there is strong evidence that the average number of daily steps is greater than 12 000.

Solution to Activity 27

The overall proportion relapsing is

$$\widehat{p} = \frac{16 + 17}{16 + 12 + 17 + 14} = \frac{33}{59} \simeq 0.559.$$

An approximate 95% confidence interval for p, the underlying proportion that relapse, is then

$$\left(\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = \left(0.559 \pm 1.96\sqrt{\frac{0.559 \times 0.441}{59}}\right)$$

$$\simeq (0.559 - 0.127, 0.559 + 0.127) \simeq (0.432, 0.686).$$

Since the confidence interval contains the value 0.5, it is indeed plausible that half of all patients relapse.

Solution to Activity 28

(a) The proportion of patients in high expressed emotion families who relapsed was

$$\frac{16}{16 + 12} = \frac{16}{28} \simeq 0.571,$$

and the proportion in low expressed emotion families was

$$\frac{17}{17 + 14} = \frac{17}{31} \simeq 0.548.$$

(b) An estimate for $d = p_1 - p_2$, the difference between the proportions, where p_1 and p_2 are the proportions for high and low expressed emotion families, respectively, is

$$\widehat{d} = \widehat{p}_1 - \widehat{p}_2 = 0.571 - 0.548 = 0.023.$$

An approximate 95% confidence interval for d is

$$\left(\widehat{d} \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}\right)$$

$$= \left(0.023 \pm 1.96\sqrt{\frac{0.571 \times 0.429}{28} + \frac{0.548 \times 0.452}{31}}\right)$$

$$\simeq (0.023 - 0.254, 0.023 + 0.254) \simeq (-0.231, 0.277).$$

This interval gives a range of plausible values for the true difference between the proportions relapsing. Since the interval contains zero, as well as both positive and negative values, we cannot conclude from these data that family expressed emotion is related to propensity to relapse.

Solution to Activity 29

If we wish to predict height from age, then age should be regarded as the explanatory variable and height as the response variable.

Solution to Activity 30

(a) Using the summary statistics provided, S_{xx} and S_{xy} are given by

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 7135 - \frac{359^2}{25} = 1979.76,$$

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 9485 - \frac{359 \times 534}{25} = 1816.76.$$

(b) The least squares estimates of β and α are

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1816.76}{1979.76} \simeq 0.9177 \simeq 0.92$$

and

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} = \frac{534}{25} - \frac{1816.76}{1979.76} \times \frac{359}{25} \simeq 8.1823 \simeq 8.18.$$

The equation of the least squares line is therefore

$$y = 8.18 + 0.92\,x.$$

That is, the fitted model is

$$\text{height} = 8.18 + 0.92 \times \text{age}.$$

Solution to Activity 31

The points in the residual plot seem to be scattered about zero in a random, unpatterned fashion. Therefore the assumption that the residuals have constant, zero mean and constant variance seems reasonable.

The points lie very roughly along the straight line in the normal probability plot, indicating that the assumption that the data come from a normal sample might be reasonable. It could, on the other hand, be argued that there is a hint of some systematic variation around the line; however, this doesn't look serious enough to rule out the normality assumption.

Solution to Activity 32

(a) An estimate of σ^2 is given by

$$s^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{626.580}{23} \simeq 27.243.$$

(b) A 95% confidence interval for β is given by

$$\left(\widehat{\beta} - t\frac{s}{\sqrt{S_{xx}}}, \widehat{\beta} + t\frac{s}{\sqrt{S_{xx}}}\right),$$

where t is the 0.975-quantile of $t(n-2)$.

Here, $n = 25$ and the 0.975-quantile of $t(23)$ is 2.069. So the confidence interval is (using the unrounded value of $\widehat{\beta}$)

$$\left(0.9177 - 2.069\frac{\sqrt{27.243}}{\sqrt{1979.76}}, 0.9177 + 2.069\frac{\sqrt{27.243}}{\sqrt{1979.76}}\right)$$

$$\simeq (0.9177 - 0.2427, 0.9177 + 0.2427) \simeq (0.68, 1.16).$$

Solution to Activity 33

The predicted height (in feet) of a 20-year-old Norway spruce is

$$8.18 + 0.92 \times 20 = 26.58.$$

Solution to Activity 34

(a) A 95% confidence interval for the mean response at the value x_0 is given by

$$\left(\widehat{\alpha} + \widehat{\beta}\,x_0 - t\,s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}}, \widehat{\alpha} + \widehat{\beta}\,x_0 + t\,s\sqrt{\frac{(x_0 - \overline{x})^2}{S_{xx}} + \frac{1}{n}}\right),$$

where t is the 0.975-quantile of $t(n-2)$.

When $x_0 = 20$,

$$\widehat{\alpha} + \widehat{\beta}\,x_0 = 8.1823 + 0.9177 \times 20 = 26.5363.$$

Also, $t = 2.069$ and $s = \sqrt{27.243}$ (from the solution to Activity 32(a)), $S_{xx} = 1979.76$ (from the solution to Activity 30(a)) and

$$\overline{x} = \frac{359}{25} = 14.36.$$

Hence the required confidence interval is

$$\left(26.5363 \pm 2.069\sqrt{27.243}\sqrt{\frac{(20 - 14.36)^2}{1979.76} + \frac{1}{25}}\right)$$

$$\simeq (26.5363 \pm 2.5571) \simeq (23.98, 29.09).$$

More decimal places were retained in the estimated regression coefficients here than in the solution to Activity 33 because of the greater complexity of the current calculations. The estimates to 4 decimal places can be found in the solution to Activity 30(b).

(b) The 95% prediction interval for the response at x_0 is given by

$$\left(\widehat{\alpha}+\widehat{\beta}\,x_0-t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}}+\frac{1}{n}+1},\widehat{\alpha}+\widehat{\beta}\,x_0+t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}}+\frac{1}{n}+1}\right).$$

So the required prediction interval is

$$\left(26.5363 \pm 2.069\sqrt{27.243}\sqrt{\frac{(20-14.36)^2}{1979.76}+\frac{1}{25}+1}\right)$$

$$\simeq (26.5363 \pm 11.0977) \simeq (15.44, 37.63).$$

(c) A confidence interval is for the mean response at a given value of the predictor, whereas a prediction interval is for an individual response. Since the latter involves the additional variability of individual observations around the mean, the prediction interval must be wider.

Solution to Activity 35

(a) The ages of the trees in Table 7 range from 4 to 38 years. At 58 years, the particular Norway spruce being considered is much older than the trees in the dataset used to fit the linear regression model. As such, we do not know whether the fitted linear regression model will continue to be appropriate for larger values of x than those used to fit the model. For example, do older Norway spruce grow at a slower rate than younger Norway spruce? If so, then it's possible that the fitted model may not be appropriate for older trees.

(b) The predicted height (in feet) of a 58-year-old Norway spruce is

$$8.18 + 0.92 \times 58 = 61.54.$$

Perhaps surprisingly, the predicted height is rather less than the observed height. An explanation might be that the Trafalgar Square Christmas tree is, presumably, not any old, randomly chosen, 58-year-old Norway spruce but one carefully nurtured and/or chosen for its shape and height. (A height of $65\frac{1}{2}$ feet is, however, well within the 95% prediction interval associated with this point prediction, which is very wide.)

Solution to Activity 36

In the scatterplots in Figures 11 and 12, the data points for selective schools are in a different location to those for comprehensive and secondary modern schools. As such, it looks like it might not be appropriate to use a single multiple regression model for all three types of school, since the relationships between the response and explanatory variables are not the same for all types of school.

Solution to Activity 37

(a) The fitted multiple regression model is

$$y = 5.104 + 0.09831\,x_1 - 0.36004\,x_2.$$

(b) From the `Coefficients` table, the p-value for each individual two-sided test of the null hypothesis $H_0 : \beta_j = 0$, for $j = 1, 2$, is 0.000, which means that for each regression coefficient $p < 0.01$. There is therefore strong evidence that each regression coefficient is non-zero, which in turn implies that together Attainment 8 (x_1) and prior attainment (x_2) influence Progress 8 (Y).

(c) The regression coefficients can be interpreted as follows.

- Regression coefficient for x_1: If the value of Attainment 8 increases by one unit, and the value of prior attainment remains fixed, then Progress 8 would be expected to increase by 0.09831. The fact that there would be expected to be an increase in Progress 8 for fixed prior attainment makes sense because if two schools have the same prior attainment, then the school with the higher Attainment 8 score will have made more progress than the school with the lower Attainment 8 score.
- Regression coefficient for x_2: If the value of prior attainment increases by one unit, and the value of Attainment 8 remains fixed, then Progress 8 would be expected to decrease by 0.36004. The decrease is indicated by the negative coefficient. This makes sense because if two schools have the same Attainment 8, then the school with the higher prior attainment will have made less progress than the school with the lower prior attainment.

Solution to Activity 38

(a) The fitted value of Progress 8 for the 78th school is

$$\widehat{y}_{78} = 5.104 + 0.09831 \times 64.8 - 0.36004 \times 31.8 \simeq 0.0252 \simeq 0.03.$$

Therefore the residual for this school is

$$w_{78} \simeq 0.00 - 0.03 = -0.03.$$

Since the residual is negative, the model over-predicted the Progress 8 score for this school, although only slightly, since the residual is small.

(b) The fitted value of Progress 8 for the first school is

$$\widehat{y}_1 = 5.104 + 0.09831 \times 78.5 - 0.36004 \times 34.0 \simeq 0.5800 \simeq 0.58.$$

Therefore the residual for this school is

$$w_1 \simeq 0.55 - 0.58 = -0.03.$$

Since the residual is negative, the model over-predicted the Progress 8 score for this school as well, although also only by a small amount (in fact, by the same amount).

(c) The fitted value of Progress 8 for the seventh school is

$$\widehat{y}_7 = 5.104 + 0.09831 \times 69.4 - 0.36004 \times 32.2 \simeq 0.3334 \simeq 0.33.$$

Therefore the residual for this school is

$$w_7 \simeq 0.35 - 0.33 = 0.02.$$

Since the residual is positive, the model under-predicted the Progress 8 score for this school, but not by very much.

Solution to Activity 39

There is one very clear outlier with a low residual value in both of these plots. This school's actual Progress 8 score is much lower than that predicted by the model, given its Attainment 8 and prior attainment scores. Other than this value, the points in the residual plot seem to be scattered about zero in a random, unpatterned fashion, and in the normal probability plot the points lie quite close to the straight line (although there is perhaps the hint of a slight systematic 'S' shape in the normal probability plot). Overall, the model assumptions do seem reasonable.

Solution to Activity 40

The fitted multiple regression models for both comprehensive and secondary modern schools are similar to that for selective schools in that there is still a positive regression coefficient for x_1 (Attainment 8) and a negative one for x_2 (prior attainment), and these again are both significantly different from zero because their p-values are reported as being 0.000 for each model. However, the actual values of the regression parameters are different for each type of school to those in the fitted model for selective schools.

Solution to Activity 41

(a) The predicted Progress 8 score for a selective school with the same Attainment 8 and prior attainment scores is

$$\widehat{y} = 5.104 + 0.09831 \times 60.9 - 0.36004 \times 29.2 \simeq 0.578.$$

(b) The predicted Progress 8 score for a secondary modern school with the same Attainment 8 and prior attainment scores is

$$\widehat{y} = 2.826 + 0.09898 \times 59.3 - 0.28432 \times 30.0 \simeq 0.166.$$

(c) For the Attainment 8 and prior attainment scores of the secondary modern school considered in part (a), the predicted Progress 8 score increases from that given by the model for secondary modern schools when made under the model that assumes it to be a selective school.

For the Attainment 8 and prior attainment scores of the selective school considered in part (b), the predicted Progress 8 score increases from that given by the model for selective schools when made under the model that assumes it to be a secondary modern school.

It seems that both schools would benefit a little from swapping their statuses! There is nothing here to suggest that either secondary modern or selective status has a uniformly helpful effect on pupils from schools with these particular, similar, levels of Attainment 8 and prior attainment scores.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 149: © Nathan Jones

Page 151: © Vasilis Ververidis / www.123rf.com

Page 153: © FlyAkwa / https://commons.wikimedia.org/wiki/File:Tsunami_Phuket.jpg This file is licensed under the Creative Commons Attribution-Noncommercial-ShareAlike Licence http://creativecommons.org/licenses/by-sa/4.0/

Page 157: © BSIP / Universal Images Group

Page 161: © DVersiga84 / https://www.flickr.com/photos/61505200@N02/32531292440/in/photostream

Page 163: © Stephen VanHorn / www.123rf.com

Page 168: © arnoaltix / iStock / Getty Images Plus

Page 169: © Miroslav Pinkava / www.123rf.com

Page 172: © Diliff / https://commons.wikimedia.org This file is licensed under the Creative Commons Attribution-Noncommercial-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 173: © Photofusion / UIG

Page 177: © 2017 South Bromsgrove High Academy Trust. Reproduced by permission

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.