Unit 12

Transformations and the modelling process

Introduction

You have now met some of the most important ideas of statistics:

- you can summarise the key features of a dataset and can represent it graphically in different ways
- you have seen that variability in a population can be represented by a probability distribution
- with a few assumptions, you are able to use a variety of distributions to model the behaviour of both discrete and continuous random variables
- you can use data and models to perform inference and hence to answer practical questions.

Each unit so far has dealt with particular topics or methods in statistics, which have been illustrated by examples. You might think of the techniques you have encountered in the module as statistical tools. In the first part of this unit (Sections 1 and 2), you will meet the final tools to be added to your statistical toolbox in this module: the use of *transformations* of the data, first in a one-sample context, in Section 1, and then in a regression context, in Section 2.

In contrast, the second part of this unit focuses on the overall statistical *modelling process*. Having assembled your toolbox, the aim now is to work out how to use it when confronted with a statistical problem. An introduction to the modelling process is provided in Section 3, along with some reminders of some of the basic tools at your disposal. In Section 4, you will practise undertaking a complete analysis using a variety of tools; the section consists of a chapter of Computer Book C. When you have finished your analysis, you must be able to summarise what you have done and tell interested parties about it: writing a statistical report is discussed in Section 5.

1 Transforming the data: the one-sample case

The continuous distributions available to you that are most commonly used for modelling purposes are the continuous uniform, exponential and normal distributions. These were introduced in Units 3, 5 and 6, respectively. On the next page is a summary of the ranges and shapes of these three distributions followed, in Figure 1, by graphs of examples of their p.d.f.s.

The chi-squared and t-distributions are continuous also, but are more rarely used directly for modelling purposes.

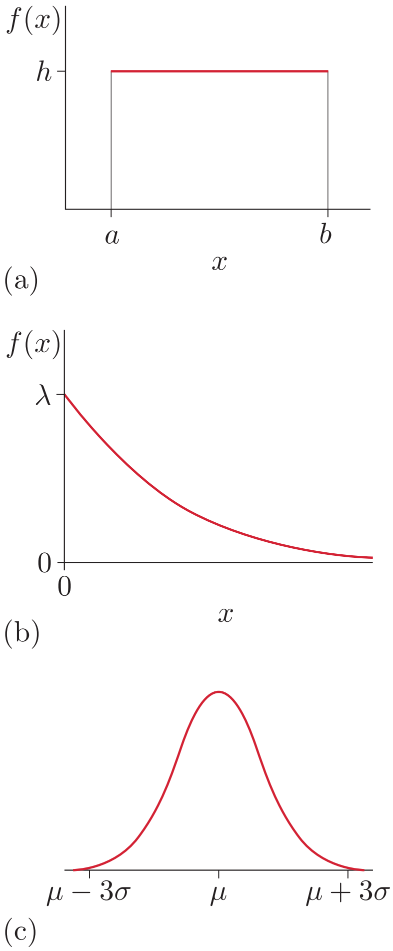


Figure 1 P.d.f.s of the (a) continuous uniform, (b) exponential and (c) normal distributions

Continuous models: range and shape

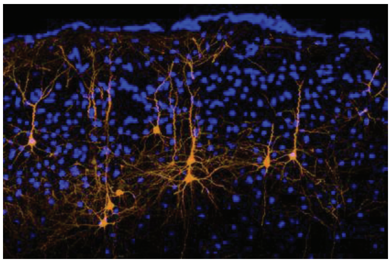
- The continuous uniform distribution has finite range $a < x < b$ and its p.d.f. is flat.
- The exponential distribution has range $0 < x < \infty$, unbounded to the right, and its p.d.f. is a decreasing function of x.
- The normal distribution has unbounded range $-\infty < x < \infty$ and is symmetric about a single mode that coincides with the mean. Values far from the mean have low probability.

At first sight, this presents a serious problem, since the choice of shapes covered by these models is very restricted.

Example 1 *Interspike intervals*

The motor cortex is the part of the brain concerned with movement. Neurons are cells that experience momentary electric potential changes, or 'spikes', the occurrence of which can be tracked over time. In the experiment of interest here, $n = 100$ 'interspike' intervals of motor cortex neurons of a monkey were measured (in milliseconds). The aims of the study were to describe the distribution of waiting times between spikes, and to estimate the firing rate of neurons.

Since the data in this case are waiting times between spikes, it seems a reasonable first assumption to assume that these spikes arise at random. Hence, a reasonable first model for the waiting times between spikes is the exponential distribution. An exponential p.d.f. was shown in Figure 1(b).



Motor cortex neurons (the orange structures looking a little like root systems) in a mouse

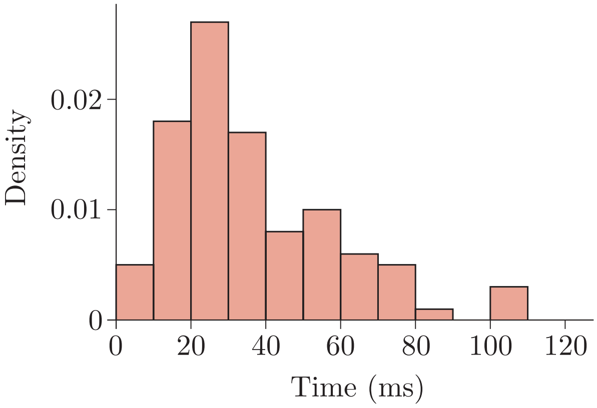


Figure 2 Histogram of interspike intervals

(Source: Zeger, S.L. and Qaqish, B. (1988) 'Markov regression models for time series: a quasi-likelihood approach', *Biometrics*, vol. 44, no. 4, pp. 1019–31)

Figure 2, on the other hand, shows a unit-area histogram of the data. The data are certainly skew and display a long right tail, as would be expected of an exponential distribution. However, it is noticeable that the mode is not at zero, but somewhere in the range 20 to 30. This could be due to random variation, but it is more likely that it reflects a failure of the exponential model.

Activity 1 *Other inappropriate models for interspike intervals*

Why does neither the continuous uniform distribution nor the normal distribution provide a good model for the data of Example 1?

1.1 Transformations: some general considerations

When dealing with continuous data, one method by which the available collection of modelling distributions can be extended enormously is by using transformations. The idea is simple: if the data do not have the shape required, you can try to transform them, that is, take a function of them, so that the transformed data do have the shape required.

Transformations can be used for certain purposes with discrete data too, but we will not consider that here.

You met *linear* transformations of data in Units 4 and 6. Unfortunately for our current purposes, linear transformations do not change the shape of a distribution. For example, you saw in Subsection 3.1 of Unit 6 that a linear transformation of normally distributed data results in another set of normally distributed data. (It is just the mean and variance of the normal distribution that are changed.) We therefore need to consider *non-linear* transformations.

You briefly met the notion of non-linear transformations of *parameters*, not of data values, in the context of confidence intervals in Subsection 3.1 of Unit 8.

The idea behind the use of (non-linear) transformations is illustrated in Example 2 using simulated data.

Example 2 *Transforming data*

Four histograms of datasets, each based on 300 data points, are shown in Figure 3 (overleaf).

The histogram in Figure 3(a) looks as though the data might be normally distributed, but those in Figures 3(b), 3(c) and 3(d) are progressively more skew and the distributions appear to be far from normal. It may surprise you to learn that the same sample of data was used for all four histograms. First, a computer was used to generate a sample of size $n = 300$ from a normal distribution; these data are represented by the histogram in Figure 3(a). Suppose that a typical value in this dataset is denoted x_i. Then the data points used for Figure 3(b) are the values x_i^2; the data points used for Figure 3(c) are the values $e^{2(x_i-1)}$; and the data points used for Figure 3(d) are the values $1/(251x_i)$.

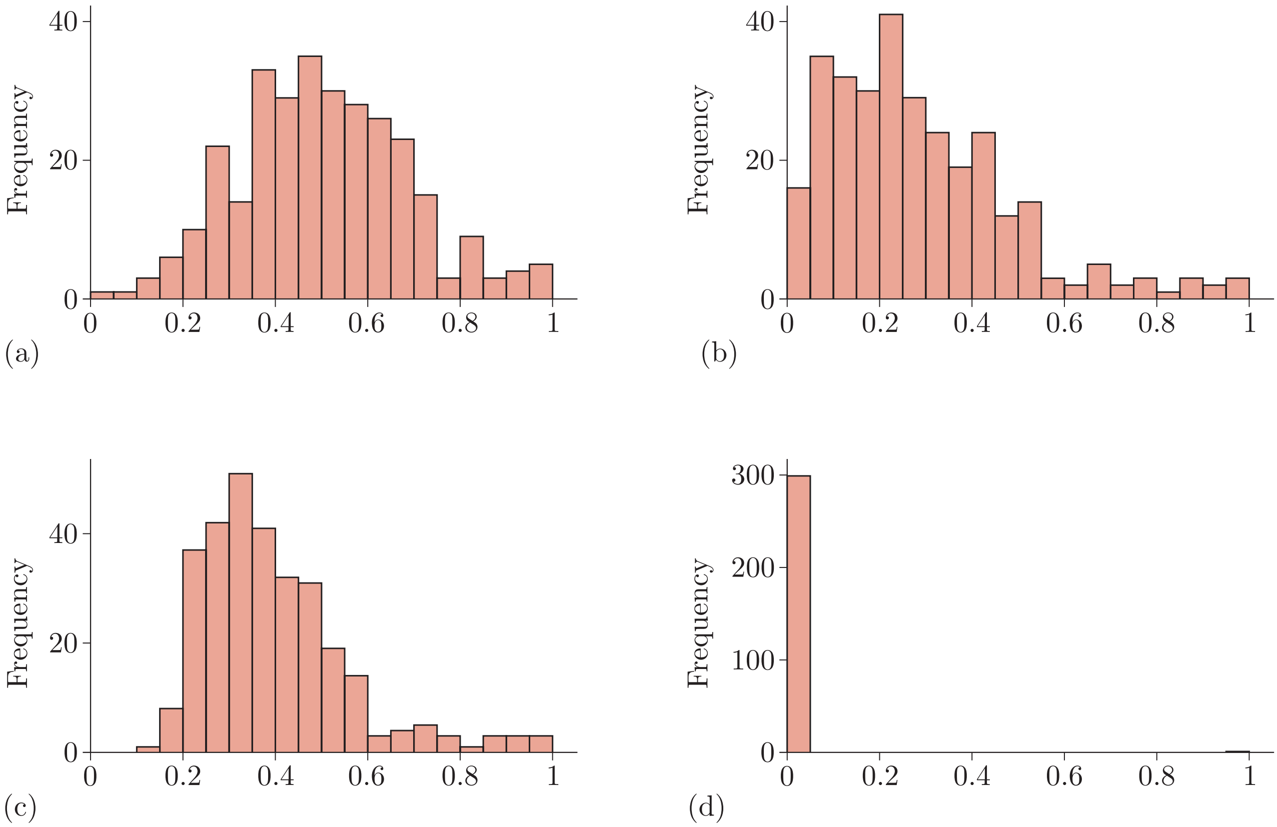


Figure 3 Four histograms

Since the data represented in Figure 3(a) are normally distributed, they could be used to carry out a t-test, for example. However, it would not be legitimate to carry out a t-test using the data in any of Figures 3(b), 3(c) or 3(d) because the variation is far from normal. Suppose now that data resembling those in Figures 3(b), 3(c) or 3(d) were to arise in practice. It would clearly be worth considering transforming them. For example, if the data looked like those in Figure 3(b), then an appropriate procedure to follow would be to take the inverse transformation to the transformation that led to Figure 3(b) from Figure 3(a) in the first place. Since the initial transformation, in this case, was to square the normally distributed data values, then the transformation of the data in Figure 3(b) that leads back to Figure 3(a) must be to take the square root of each value. It would then be appropriate to carry out t-tests on the square-root-transformed data, based on the assumption of normality.

One aim of transforming a set of data values to a different set of values by means of a mathematical transformation is, as in the second half of Example 2, to render the transformed data more plausibly normal. If we are able to do this, then we can perform statistical modelling and inference by applying the techniques we already have available for normally distributed data to the transformed dataset. Information that we obtain via the transformed version of the dataset can then be reinterpreted in terms of questions concerning the original, untransformed, data. So this is

A popular alternative target of transforming data is just to make them more symmetric rather than specifically normal.

the aim of transforming data on which we will concentrate in this section. In this case, assessment of the success or otherwise of proposed transformations can be made using normal probability plots, as introduced in Section 5 of Unit 6 and used for residuals in a regression context in Unit 11.

Wedding day in the rain in Minneapolis–St Paul

Example 3 *March precipitation in Minneapolis–St Paul*

The total precipitation (in inches) in the month of March was recorded in 30 successive years in Minneapolis–St Paul, Minnesota, USA. Figure 4(a) shows a normal probability plot for these data. The normal probability plots in Figures 4(b) and 4(c) were produced after transforming the data using the log transformation $\log x$ and the cube root transformation $x^{1/3}$, respectively. That is, if the original data are denoted by x_i, then the normal probability plot in Figure 4(b) is based on the values $\log x_i$, and the normal probability plot in Figure 4(c) is based on the values $x_i^{1/3}$.

As always in this module, 'log' (without a subscript) denotes the natural logarithm

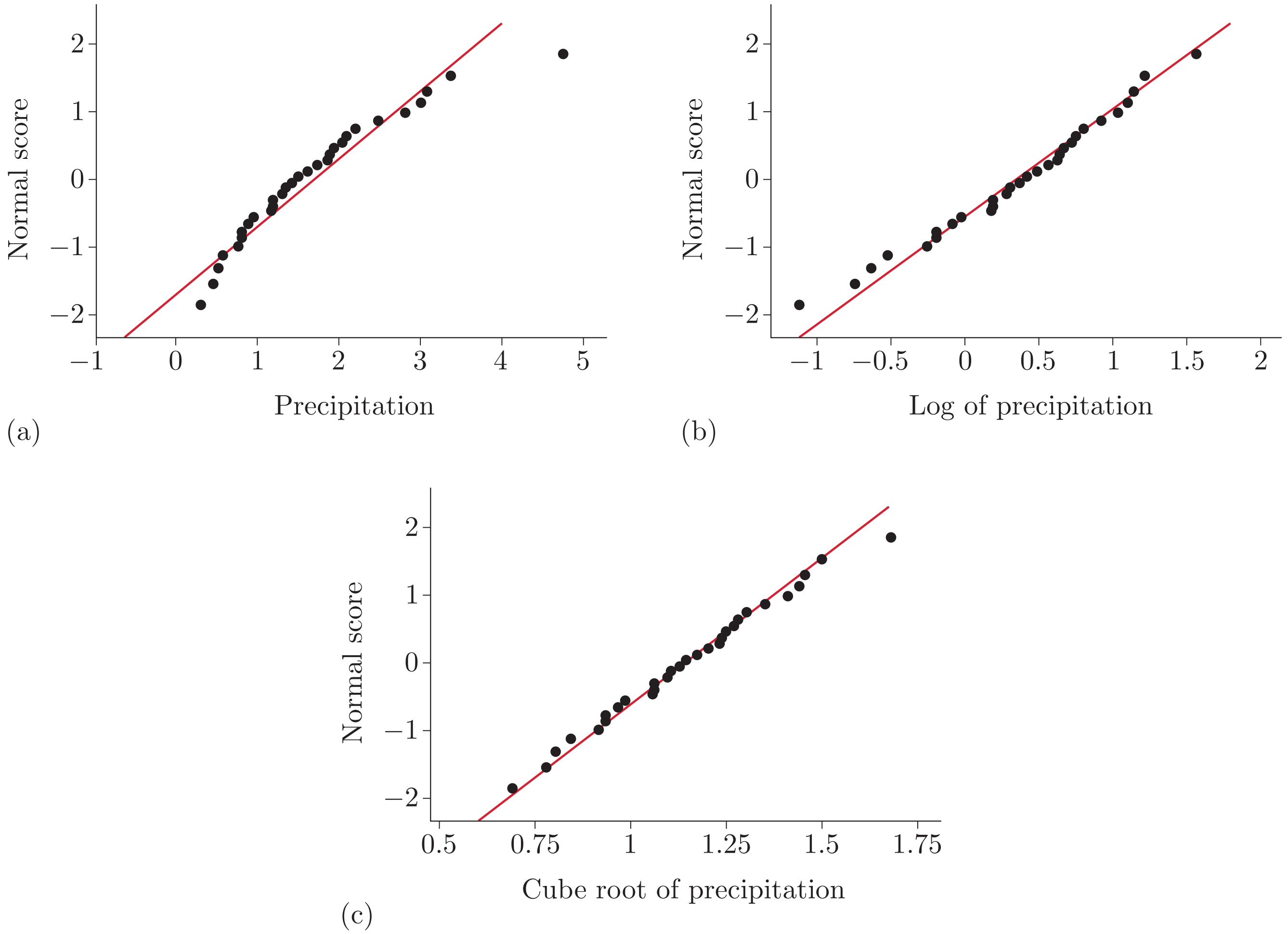


Figure 4 Normal probability plots for precipitation data: (a) untransformed, (b) log transformed, (c) cube root transformed

(Source: Hinkley, D. (1977) 'On quick choice of power transformation', *Applied Statistics*, vol. 26, no. 1, pp. 67–9)

The probability plot for the untransformed data displays a systematic pattern giving an indication of non-normality. The log transformation results in a straighter plot, suggesting that while the x_i values might not be modelled (directly) by a normal distribution, the $\log x_i$ values might be modelled by a normal distribution. Arguably, an even straighter probability plot is obtained using the cube root transformation; it seems reasonable also to model the values of $x_i^{1/3}$ using a normal distribution.

Suppose now that a typical value in a dataset is denoted by x and its transformed value by $y = h(x)$, where h is some transformation function. How do we go about choosing h? We have seen that if we can think of a possible function to act as h, then we can check whether using it results in normality of the transformed data by using a normal probability plot. But possibilities for h include those you have already seen, like $y = \log x$ and $y = x^{1/3}$, and others like $y = x^2$, $y = e^x$ and $y = 1/x$; in fact, the list is endless.

Well, first, we can observe that some transformations are simply not available for use with some data. For example, the transformation $y = \log x$ is not defined for negative (or zero) values of x, so it can be used only when all the data values are positive. The same goes for the transformation $y = \sqrt{x}$. That is, the transformation $y = h(x)$ needs to be defined for all values in the range of the distribution of the data.

Second, it turns out that the approach makes sense only if a transformation is either increasing or decreasing over the range of the distribution of x. Recall that a transformation of x is increasing if its graph rises as you move to the right through the range of x; it is decreasing if its graph falls. Alternatively, as you were reminded in Subsection 3.1 of Unit 8, a transformation $h(x)$ is increasing if its derivative is positive, that is, $h'(x) > 0$, and is decreasing if its derivative is negative, that is, $h'(x) < 0$. So we need to choose h so that either $h'(x) > 0$ or $h'(x) < 0$ over the range of values of the distribution of x.

Example 4 *The transformation $y = x^2$*

Suppose that x can take both positive and negative values. Then the transformation $y = x^2$ is neither increasing nor decreasing over the range of the distribution of x. This is illustrated in Figure 5, where the graph of $y = x^2$ is shown for $-2 < x < 2$: $y = x^2$ is decreasing between -2 and 0, then increasing between 0 and 2. Mathematically, if $h(x) = x^2$, then $h'(x) = 2x$, so $h'(x) < 0$ for $-2 < x < 0$ (in fact, for all $x < 0$) and $h'(x) > 0$ for $0 < x < 2$ (in fact, for all $x > 0$). However, if the range of the distribution of x included only values $x > 0$, then the transformation $y = x^2$ would be appropriate as it is increasing over this range.

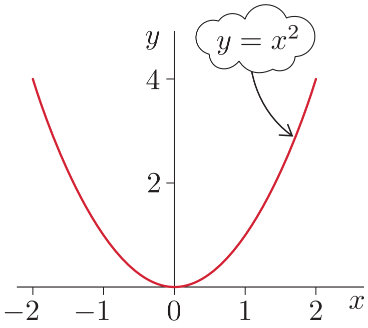


Figure 5 The graph of $y = x^2$ for x between -2 and 2

Activity 2 *Which transformation is possible?*

Suppose that a dataset consists of values x such that $-1 < x < 1$. Only one of the four transformations listed below is both defined and either increasing or decreasing over the range of the distribution of x. Identify which one, and explain why each transformation is or is not available for such data.

$$y = \sqrt{x} = x^{1/2}, \quad y = x^4, \quad y = (2+x)^2, \quad y = -\log x.$$

A summary of the basic requirements of transformations is given in the following box.

Basic requirements of transformations

To be suitable for consideration as a transformation of data x, a transformation $y = h(x)$ has to be defined and either increasing or decreasing over the range of the distribution of x.

The case for transforming data is strengthened if some natural physical interpretation of the transformation is available. This might, for example, have something to do with units of measurement: suppose that x represents a volume, in units of m^3, so that $x^{1/3}$ is in units of metres; then working with the latter *might* be preferable for some purposes. On the other hand, if x actually arises from the product of two terms, $x = tw$ say, though you can't individually observe t and w, then the basic rules of logarithms mean that

$$y = \log x = \log(tw) = \log t + \log w. \tag{1}$$

Since y is now a sum, it is *possible* that its distribution is 'more normal' than that of x, by a kind of Central Limit Theorem effect, albeit for a sum of only two random variables.

Most often, however, as in Example 3, there is no convenient physical interpretation, and the data are transformed simply to satisfy the requirements of the statistical procedure that you wish to use. Moreover, in such cases, again as you saw in Example 3, two (or more) transformations might prove to be broadly equally justifiable, and it doesn't then really matter which of them you choose to use.

Some general indications for choosing a transformation may, however, be given, especially for data x that are positive; it is this scenario on which we concentrate in the remainder of this section. When positive data are highly right-skew, with many relatively small values and fewer higher values, and possibly some very high values, the following transformations – which are all shown in Figure 6 (overleaf) – tend to reduce the spread of higher values more than that of lower values:

$$y = \sqrt{x} = x^{1/2}, \quad y = x^{1/3}, \quad y = \log x, \quad y = \log(1+x).$$

This is because they all, while being increasing transformations, slow down their rate of increase as x increases.

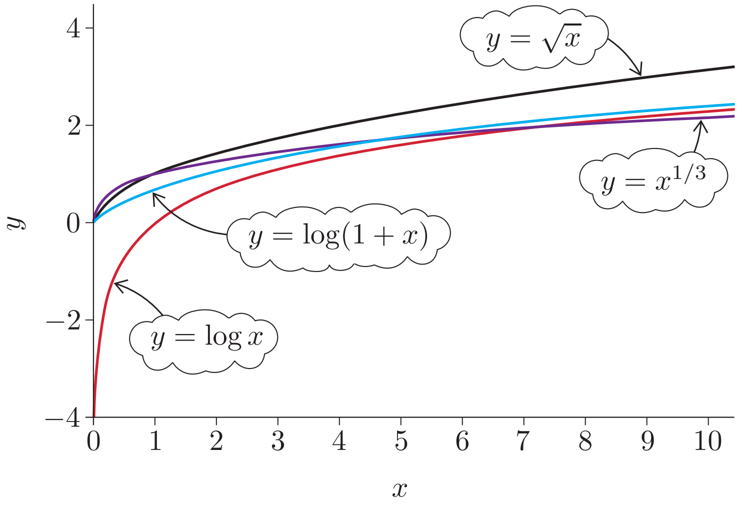


Figure 6 The graphs of $y = \sqrt{x}$, $y = x^{1/3}$, $y = \log x$ and $y = \log(1 + x)$ for $x > 0$

Each of the other transformations in Figure 6 works in the same way.

This is clarified in Figure 7 for the particular case of $y = \sqrt{x}$. You can see from the figure how the right-skew set of data values x_i is transformed to a much less skew set of transformed data values y_i. The overall effect of any one of these transformations is therefore to reduce the right-skew in the data, and potentially to make them symmetric (and even normally distributed!).

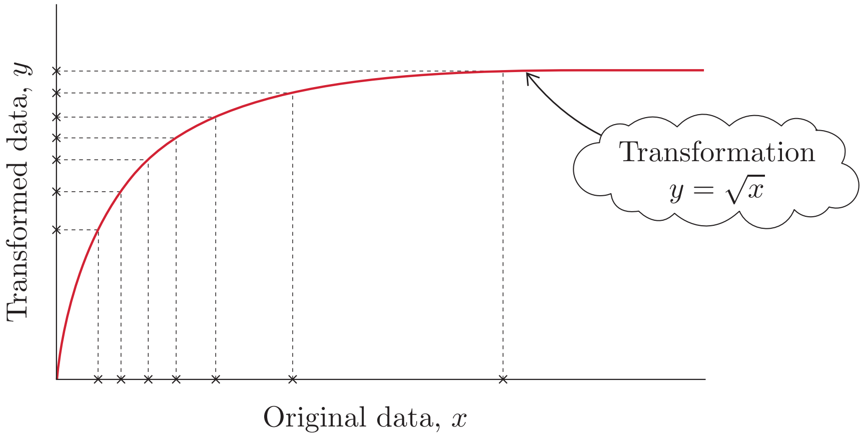


Figure 7 The way the transformation $y = \sqrt{x}$ works

1.2 The ladder of powers

The notion at the end of Subsection 1.1 can be taken a bit further. It is popular to consider a *ladder of powers*, which lists transformations of the form

$$\ldots,\ x^{-2},\ x^{-1},\ x^{-1/2},\ \log x,\ x^{1/2},\ \boxed{x^1},\ x^2,\ x^3,\ x^4,\ \ldots.$$

The transformation $y = x^1 = x$ leaves data values unchanged. Notice the position of $\log x$ in the ladder of powers: although not, in fact, a power transformation, it fills the position of x^0, which is not a valid transformation because it collapses all values to 1.

Transformations corresponding to powers below 1 on the ladder (that is, transformations to the left of x^1 in the list above) all contract high data values relative to low data values. The first two of these transformations, $y = \sqrt{x}$ and $y = \log x$, were shown in Figure 6.

The next three of these transformations,

$$y = x^{-1/2} = \frac{1}{\sqrt{x}}, \quad y = x^{-1} = \frac{1}{x}, \quad y = x^{-2} = \frac{1}{x^2}$$

are shown in Figure 8.

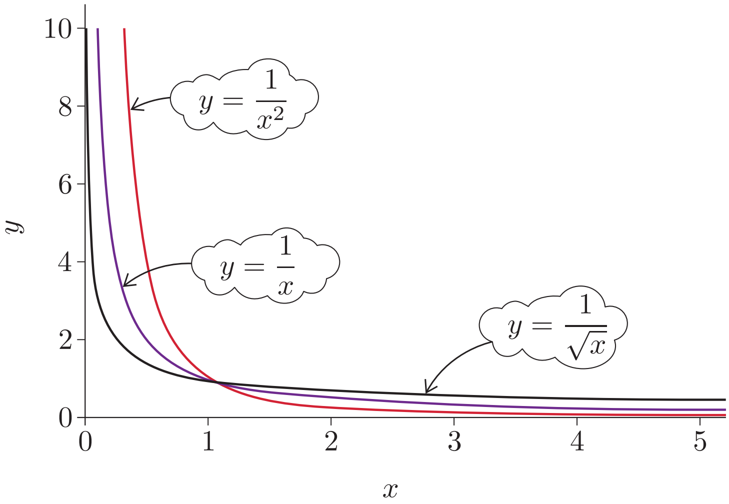


Figure 8 The graphs of $y = 1/\sqrt{x}$, $y = 1/x$ and $y = 1/x^2$ for $x > 0$

Figure 9 (overleaf) shows (via the particular example of $y = 1/x$) how these three transformations work: they too expand low values while contracting high values. However, because these are decreasing transformations, they also flip the values around: high values of x become low values of y, while low values of x become high values of y. This is not a problem in the context of using such a transformation to transform data to normality: all the transformations on the ladder of powers with powers below 1 can reduce any right-skew in the data.

On the other hand, transformations with powers above 1 on the ladder of powers (that is, transformations to the right of x^1 in the list) all expand high data values relative to low data values. These transformations are shown in Figure 10 (overleaf), and the way they work (in the particular case of $y = x^2$) is shown in Figure 11 (overleaf). These transformations are therefore most useful if we need to reduce left-skew in the distribution of the data.

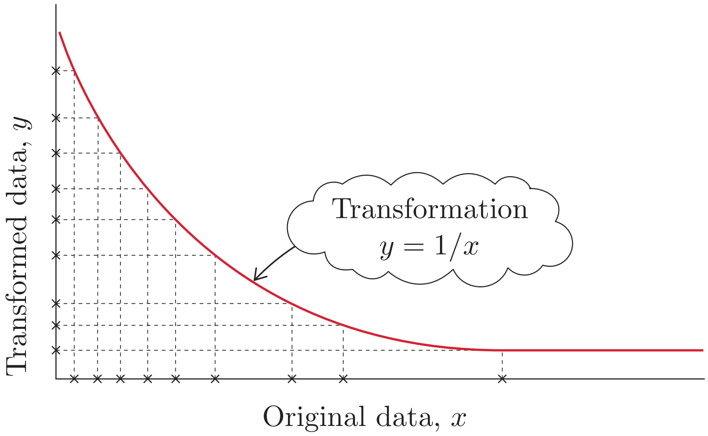


Figure 9 The way the transformation $y = 1/x$ works

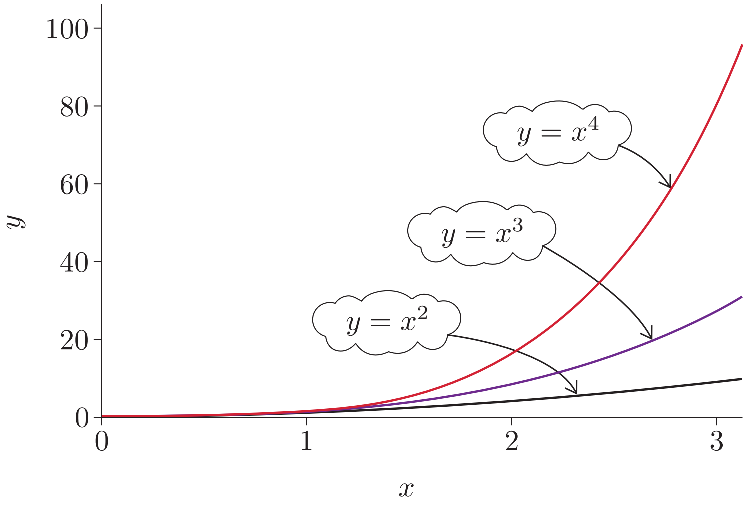


Figure 10 The graphs of $y = x^2$, $y = x^3$ and $y = x^4$ for $x > 0$

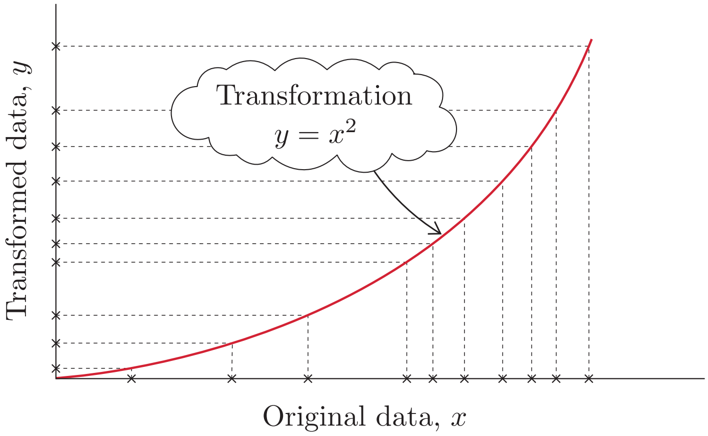


Figure 11 The way the transformation $y = x^2$ works

Screencast 12.1 further explores the effects of transformations on the ladder of powers.

Screencast 12.1 Transformations on the ladder of powers

Transformations: ladder of powers

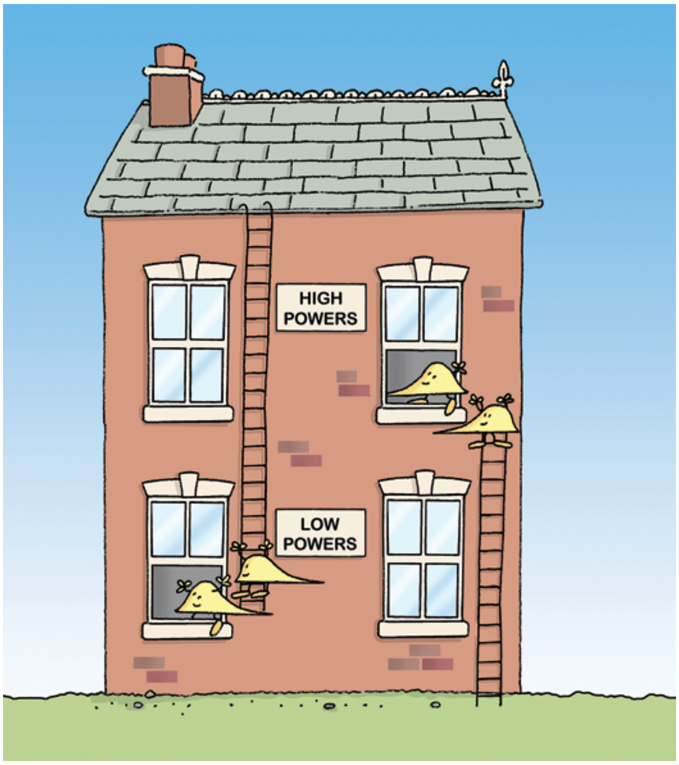
The **ladder of powers** lists transformations of the form

$$\ldots, x^{-2}, x^{-1}, x^{-1/2}, \log x, x^{1/2}, x^{1}, x^{2}, x^{3}, x^{4}, \ldots.$$

The transformation x^1 leaves the data unchanged.

When transforming skew, positive, data to make them more symmetric, and hence more amenable to modelling with a normal distribution:

- for right-skew data, go **down** the ladder of powers
- for left-skew data, go **up** the ladder of powers.



By the way, you might have noticed that, with the exception of $\log x$, which transforms positive data to data which can take any value, the transformations in the ladder of powers transform positive data to a different set of positive data. As mentioned earlier in the module, the normal distribution can still be used as a reasonable model for such data if the probability it assigns to negative values is suitably small.

Activity 3 *Which transformation to use?*

Figure 12 shows histograms for three datasets (which each happen to have been simulated using a computer). For each dataset, suggest a suitable possible transformation that would make the transformed data more symmetric and hence more nearly normally distributed.

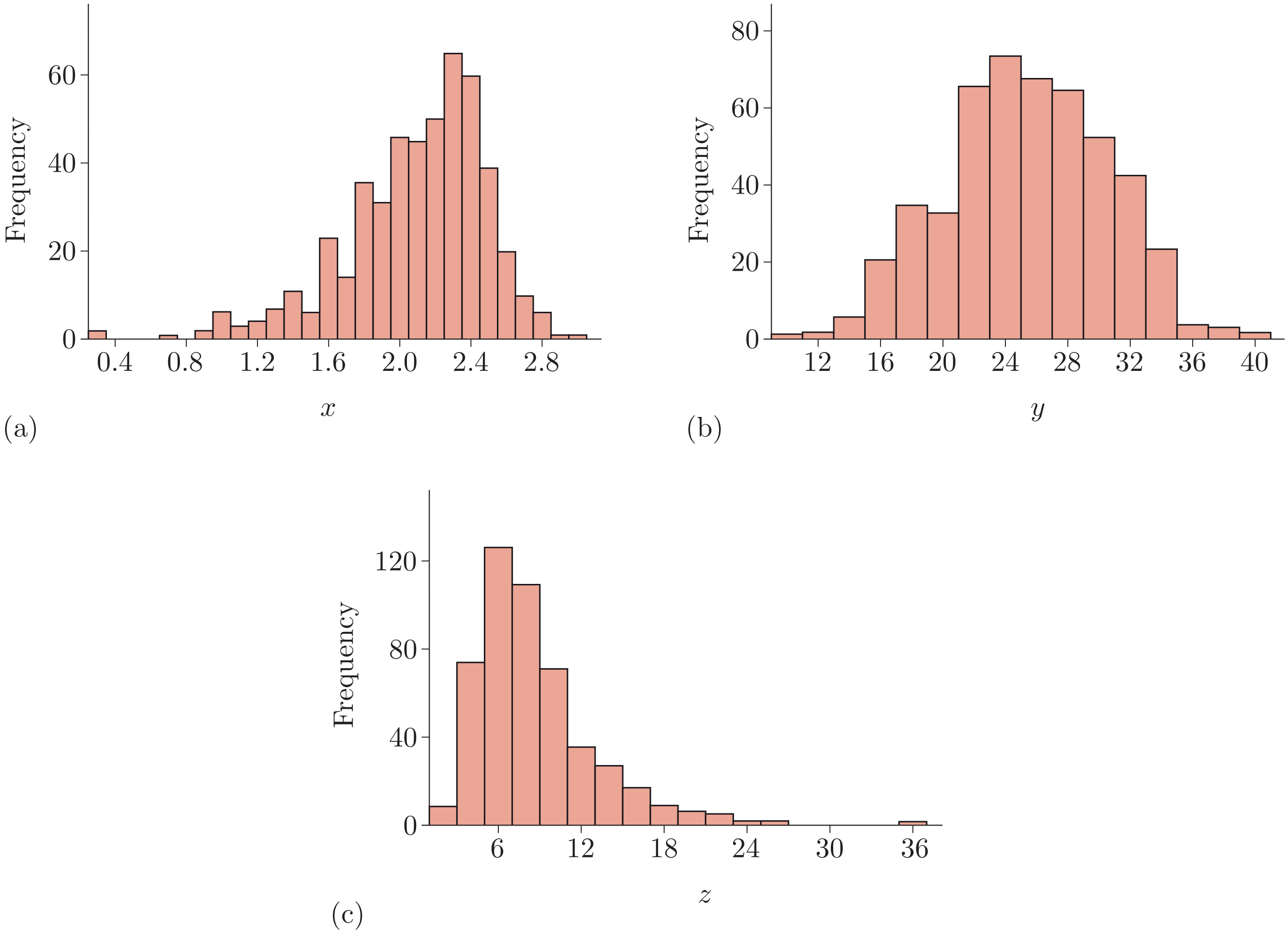


Figure 12 Histograms of three simulated datasets

Activity 4 *Interspike intervals*

In Example 1, it was found that an exponential model is probably not appropriate for the data on interspike intervals. (See the histogram in Figure 2.) In Activity 1, it was argued that a normal model is probably not appropriate for the data on interspike intervals either. This is now confirmed in the normal probability plot of the interspike intervals given in Figure 13: there is a distinct bend in the normal probability plot. In fact, a bend of the type observed in Figure 13 – the points increase first faster than the overall straight line, then slower – is typical when the data reflect a unimodal, right-skew, distribution.

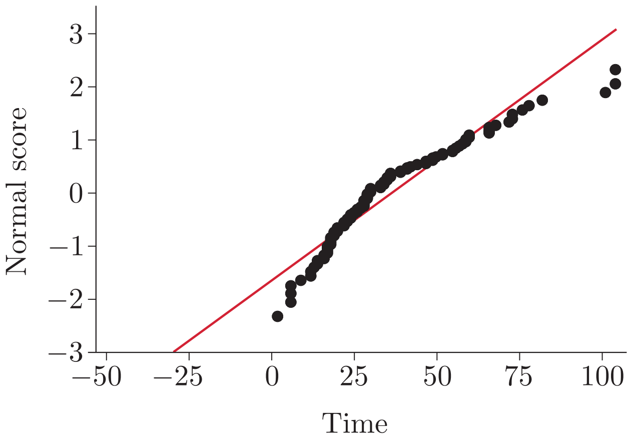


Figure 13 Normal probability plot of interspike intervals

(a) Which transformations from the ladder of powers would you expect to be the most appropriate when attempting to transform the interspike interval data to normality, and why?

(b) Figure 14(a) shows a normal probability plot of the data after a log transformation, and Figure 14(b) shows a normal probability plot of the data after a square root transformation. Comment on the effects of the two transformations. In your view, which transformation has produced the more normally distributed result?

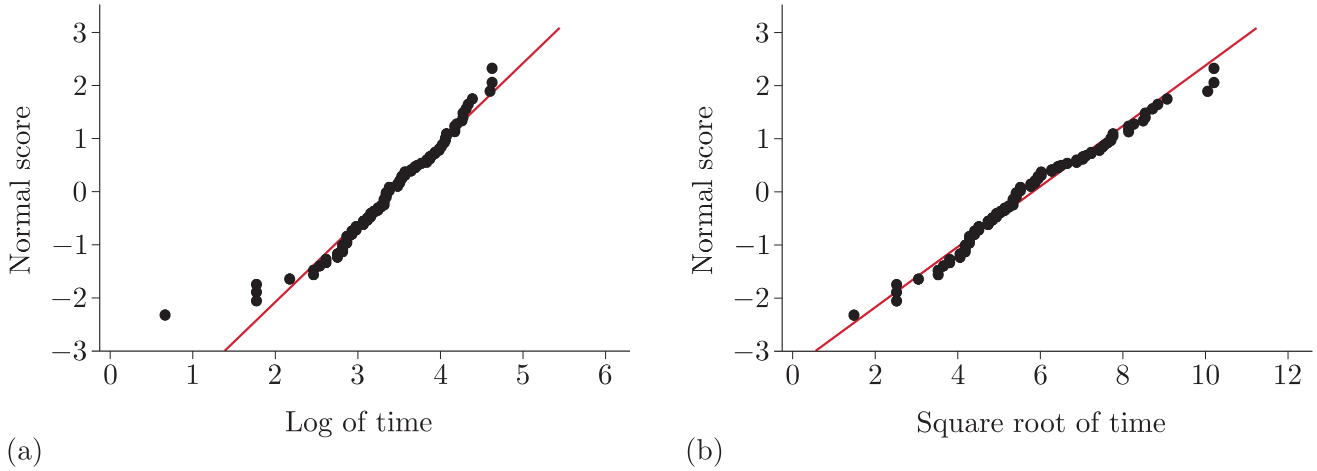


Figure 14 Normal probability plots of interspike intervals: (a) log transformed, (b) square root transformed

It is worth pointing out at this stage that it is not always necessary to obtain a more symmetric or normal distribution! For example, suppose that the aim of the analysis of the interspike intervals data were to calculate the mean interspike interval, together with a 95% confidence interval for the mean. Then, since the sample size is sufficiently large – there are 100 observations – by the Central Limit Theorem, the distribution of the sample mean is approximately normal, even though the underlying distribution is skew. So large-sample methods can be used to find an approximate 95% confidence interval for the mean. Therefore it is not necessary to transform the data to find a confidence interval for the mean. There may, of course, be other reasons to transform the data.

The sample mean of the interspike intervals turns out to be 36.49 ms with 95% large-sample confidence interval for the mean of $(32.20, 40.78)$ ms.

Exercises on Section 1

Exercise 1 *Validity of powers on the ladder*

Implicit in Subsection 1.2 is the claim that all transformations on the ladder of powers,

$$\ldots,\ x^{-2},\ x^{-1},\ x^{-1/2},\ \log x,\ x^{1/2},\ x^{1},\ x^{2},\ x^{3},\ x^{4},\ \ldots,$$

are valid transformations for positive data in the sense that they are defined and either increasing or decreasing functions of positive x. The former claim is obvious – you can take any power, or a log, of a positive value; the latter claim appears to be true from Figures 6, 8 and 10. By writing any member of the ladder of powers other than log in a unified mathematical form, verify mathematically that every member of the ladder of powers (other than log) is either increasing or decreasing, and identify which are which.

Of course, log is also an increasing function of positive x.

Exercise 2 *Strength of glass fibres*

In Example 2 of Unit 8, a dataset comprising $n = 63$ strengths (in unspecified units) of glass fibres, each of length 1.5 cm, was described. In Unit 8, this dataset was used to introduce the ideas underlying the provision of a large-sample, approximate, confidence interval for the mean strength of such glass fibres. For that purpose, the actual distribution of the glass fibre strengths was not needed (because of the Central Limit Theorem). Suppose now, however, that for some other reason there is interest in the distribution of glass fibre strengths.

(a) A histogram of the data is given in Figure 15. Does a normal distribution appear to be a possible model for these data? If not, why not?

(b) Consider the possibility of transforming the glass fibre data to improve their symmetry and hence potential normality. Which transformations from the ladder of powers would you expect to be the most appropriate when attempting to transform the glass fibre data to normality, and why?

(c) Figure 16(a) shows a normal probability plot of the data. The rest of Figure 16 shows normal probability plots after transforming the data: Figure 16(b) after a square transformation, Figure 16(c) after a cube transformation, and Figure 16(d) after a fourth power transformation. Comment on the effects of the three transformations. In your view, which transformation has produced the most normally distributed result?

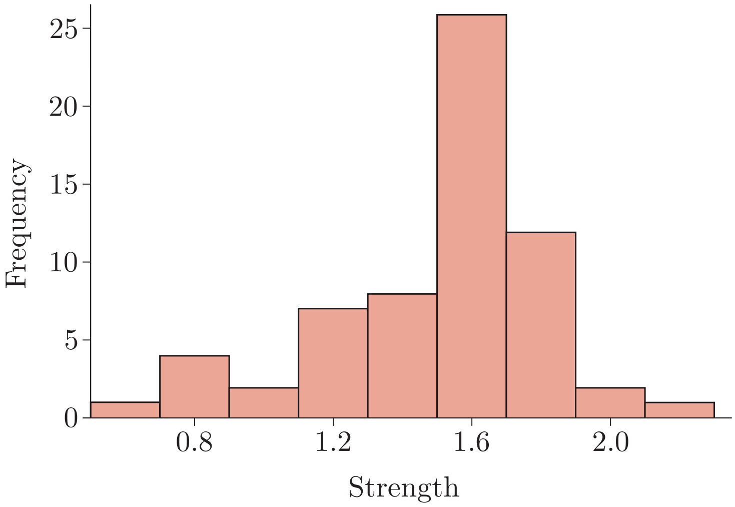


Figure 15 Histogram of glass fibre strengths

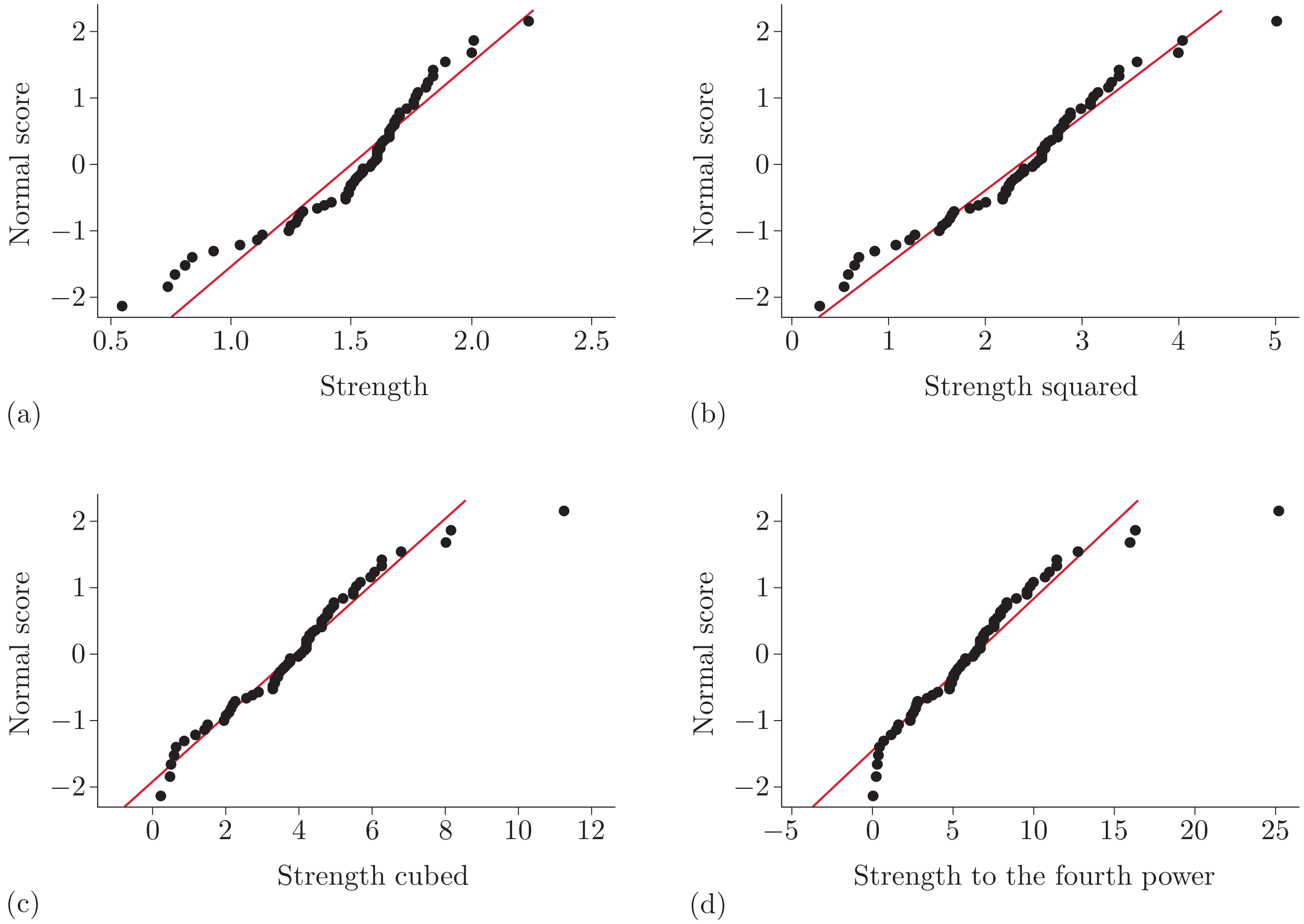


Figure 16 Normal probability plots of glass fibre strengths: (a) untransformed, (b) squared, (c) cubed, (d) taken to the fourth power

2 Transformations in regression

In this section, let us turn our attention back to regression modelling. In Section 1 of Unit 11, you saw that a scatterplot of the data often gives some indication of the relationship between two variables. In particular, when the relationship appears to be linear, the data might be fitted by a simple linear regression model. You then learned, in Sections 2 to 4 of Unit 11, how to fit such a linear model to the data using least squares, how to check the modelling assumptions, and how to perform statistical inference for linear regression models. And you were introduced to multiple regression in Section 5 of Unit 11.

It turns out that transformations can also prove to be very useful – in more ways than one – in regression modelling.

Non-linear to linear? As if by magic . . .

First, in some situations, apparently non-linear relationships can be reformulated as linear relationships, and the techniques of Unit 11 applied to them. This is done by transforming the *explanatory variable* and then modelling the dependence of the response variable on the transformed explanatory variable. This is the topic of Subsection 2.1. The core of the work in that subsection is contained in Computer Book C, however.

Second, the other important reason for transforming regression data has to do with the variation of the random terms. For data where the random terms W_i appear to be non-normal, or where the assumption of constant variance of the random terms does not appear to be reasonable, it is sometimes possible to make an appropriate transformation of the *response variable* such that the assumptions seem to be satisfied for the transformed data. This is the topic of Subsection 2.2.

These two uses of transformations in linear regression are summarised in the following box.

In linear regression, it is sometimes possible to:

- straighten out the regression function by **transforming the explanatory variable**
- make the assumptions associated with the random terms conform to those of the linear regression model by **transforming the response variable**.

It is also possible to apply transformations to both explanatory and response variables at the same time, but this will not be investigated in this module. Transformations are also relevant to multiple regression; this is discussed briefly in Subsection 2.3, only for the main focus of that subsection to become the use of multiple regression to solve a specific type of transformation problem in linear regression with one explanatory variable!

2.1 Linear regression on a function of the explanatory variable

We start this subsection with an example of linear regression with one explanatory variable where, by considering the nature of the explanatory variable, the possibility of working with a transformation is raised.

Example 5 *Removing arsenic from drinking water*

This example concerns the results of an experiment investigating how the effectiveness of a method to remove traces of arsenic from drinking water depends on the water's pH, that is, its degree of acidity or alkalinity. Figure 17 shows a scatterplot of the response variable, the percentage of arsenic removed, against the explanatory variable, the pH level, along with the line fitted by least squares. The fitted line has equation

$$y = 190.3 - 18.03\,z, \tag{2}$$

where y is the percentage of arsenic removed and z is the pH level. (You will see why we have used z instead of the usual x in a moment.) Normal probability and residual plots (not shown) suggest that the linear regression model is reasonable for these data.

Filtering through sand to remove arsenic from drinking water in the Red River delta, Vietnam

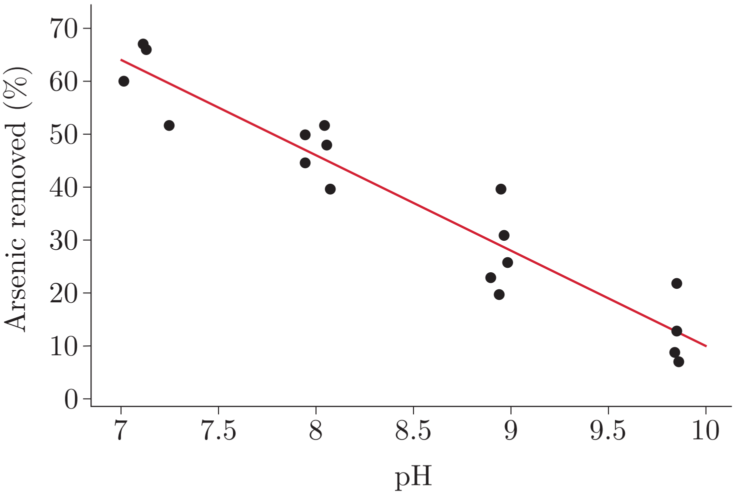


Figure 17 Percentage of arsenic removed, y, against pH, z

(Source: data taken from Devore, J.L. (2014) *Probability and Statistics for Engineering and the Sciences*, 9th edn, Boston, Cengage Learning, p. 490; Devore approximated it from Lytle, D.A., Sorg, T.J. and Snoeyink, V.L. (2005) 'Optimizing arsenic removal during iron removal: theoretical and practical considerations', *Journal of Water Supply Research and Technology – AQUA*, vol. 54, no. 8, pp. 545–60)

Now, values of pH are on a scale running from 1 to 14; the value 7 corresponds to neutrality, and to pure water. Values greater than 7 (as in the experiment reported above) correspond to an alkaline solution; values lower than 7 indicate acidity. This is a standard, internationally agreed, scale for pH, but it is derived from a more basic measured quantity: the pH level is the negative of the logarithm to base 10 of the activity of the hydrogen ion. It is equally meaningful, therefore, to ask how the percentage of arsenic removed, y, depends (on average) on the activity of the hydrogen ion; call the latter x and note that $z = -\log_{10} x$.

We therefore have a relationship between y and x by putting $z = -\log_{10} x$ in Equation (2):

$$y = 190.3 - 18.03 \times (-\log_{10} x) = 190.3 + 18.03 \log_{10} x. \qquad (3)$$

Figure 18 shows a scatterplot of the response variable, the percentage of arsenic removed, against this alternative explanatory variable, the activity of the hydrogen ion, along with the fitted curve given by Equation (3).

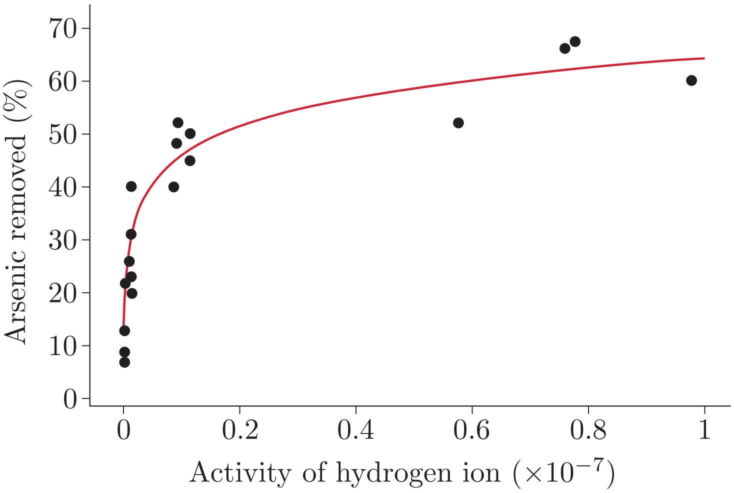


Figure 18 Percentage of arsenic removed, y, against activity of the hydrogen ion, x

Now, Formula (3) is non-linear in x. It therefore seems that we have managed to fit a suitable (very) non-linear relationship between y and x via least squares fitting of a straight-line relationship between y and z! The main point here, therefore, is that there appears to be a straight-line relationship between y and some function, or transformation, of x, namely $-\log_{10} x$.

We could equally well have written $\log_{10} x$ in place of $-\log_{10} x$ here; a straight-line relationship applies to either transformation of x because they are (in this case, very simple) linear functions of one another.

What we have seen in Example 5 is that it is sometimes possible to 'straighten out' or 'linearise' the data by a suitable *transformation of the explanatory variable* so that a linear regression model can be fitted to the transformed data. Then the results from Unit 11 can be used on the transformed data. Explicitly, if h denotes the transformation of x so that $x' = h(x)$, say, then the model for the data becomes

$$Y_i = \alpha + \beta\, h(x_i) + W_i$$

Note that the variation in the random terms W_i is not affected by this transformation.

or equivalently

$$Y_i = \alpha + \beta\, x'_i + W_i.$$

This model is non-linear in x but linear in the new explanatory variable x'.

Naturally, results from an analysis made on the transformed data will refer to the transformed data and *not* directly to the original data. You should always take care to present results with reference to the appropriate set of data in order to avoid misunderstandings or misinterpretations.

Finding an appropriate transformation in this context can be very difficult and it often involves a certain amount of trial and error. Statisticians always use a computer to do this, so almost all the activities for this subsection are in Computer Book C. Unfortunately, apart from including many of the transformations that you might wish to try, the ladder of powers of Section 1 is not really directly useful here.

This non-linear river has linearised itself with the creation of an oxbow lake

Refer to Chapter 4 of Computer Book C for the next part of the work in this section.

Two of the datasets introduced in Section 1 of Unit 11 look as though they might be modelled in such a way that they could be treated by linear regression on a suitable function of x, as above. But can they? The first of these, the duckweed data introduced in Example 5 of Unit 11, will be considered in the following example; the second of these, the paper strength data introduced in Activity 2 of Unit 11, will be considered in Subsection 2.3.

Example 6 *More on the model for the duckweed data*

Figure 19 is a repeat of Figure 6 of Unit 11, showing a scatterplot of the data in this case.

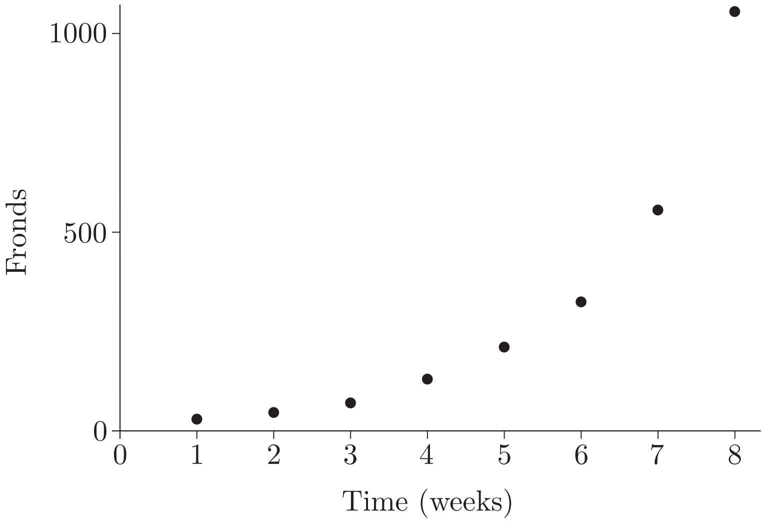


Figure 19 Number of duckweed fronds against time

In Example 7 of Unit 11, it was argued that a possible regression model for the duckweed data might be

$$Y_i = 20e^{\lambda x_i} + W_i,$$

where Y represents the number of duckweed fronds, and x represents the time after the start of the experiment, in weeks. This model is certainly non-linear in x. However, it appears that if we define $x' = h(x) = 20e^{\lambda x}$, then we would have a linear regression model of the form $Y_i = x'_i + W_i$ (which is, in fact, a linear regression model going through the origin). There is a snag, however. Unlike the transformations of x that you have been considering above and in Computer Book C, this transformation involves the unknown parameter λ. If λ were known to be 1, say, then we could indeed proceed by linear regression on the transformed explanatory variable $x' = 20e^{x}$. But λ is not known and it too needs to be estimated. In such cases, the transformation approach is not available and the model is said to be inherently or intrinsically non-linear. It then needs to be treated by the methods of *non-linear regression*, but these are beyond the scope of this module.

Similarly for any other fixed value of λ.

An important aspect, then, of being able to linearise the data by a suitable transformation of the explanatory variable is that such a transformation should be fully specified and not itself depend on an unknown parameter.

Activity 5 *Which of these regression functions can be linearised?*

Which of the following regression functions can be linearised by employing a suitable transformation so that a linear regression model can be fitted to the transformed data?

$$\alpha + \beta x^3, \quad \alpha + \beta \log(x + \lambda), \quad \alpha + \beta \log\left(\frac{x}{1-x}\right), \quad \beta \exp\left(\mu x + e^{-\gamma x^2}\right).$$

After our brief dalliance with logs to the base 10 in Example 5, we revert to natural logarithms from here on.

2.2 Transforming the response variable

Straightening out a non-linear relationship in order to be able to fit a linear relationship to the data, as in Subsection 2.1, is one of the two main reasons for transforming regression data. The second main reason for transforming regression data is to try to make the assumptions associated with the random terms, W_i, conform to those of the linear regression model – constant, zero mean, constant variance, normality – in situations where they don’t. While the method in Subsection 2.1 for linearising the regression function was transformation of the explanatory variable, the method in this subsection for ‘normalising’ the random terms in the model is *transformation of the response variable.* (In this case, we leave the explanatory variable as it is.)

Example 7 *Fuel efficiency and weight of cars*

For the purposes of a competition associated with the 1983 Annual Meeting of the American Statistical Association, a dataset was compiled on attributes of a number of models of car then in use in the USA. This example concerns two of the variables from that dataset: the response variable is fuel efficiency measured as the number of miles per gallon of petrol (mpg) achieved by that model of car, and the explanatory variable is its weight in pounds (lb). There are $n = 398$ data points in the dataset. How does fuel efficiency depend on weight?

A 1982 Ford Mustang GL

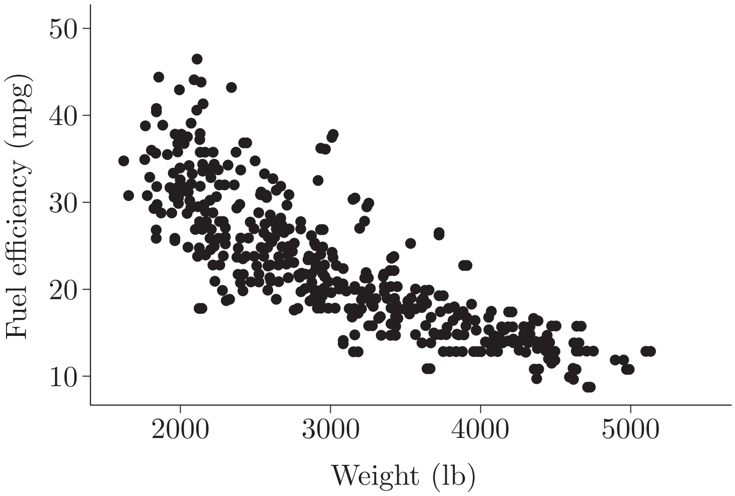


Figure 20 Fuel efficiency against weight

(Source: data of E. Ramos and D. Donoho, extracted from the *Statlib* data archive at Carnegie Mellon University, Pittsburgh, USA)

It is clear from Figure 20 that fuel efficiency decreases as weight increases. The decrease looks a bit non-linear, so maybe a transformation of the explanatory variable, weight, might be considered. However, another, striking feature of Figure 20 is that the amount of variability in the response variable appears not to be constant but also to decrease with increasing weight. A transformation of the response variable, fuel efficiency, might be worth trying, therefore. Figure 21 shows the transformed response variable log(fuel efficiency) plotted against weight. The effect of the transformation is as we might have hoped: the variability in the data appears to be constant (and, as a bonus, the dependence of log(fuel efficiency) on weight appears to be at least approximately linear!).

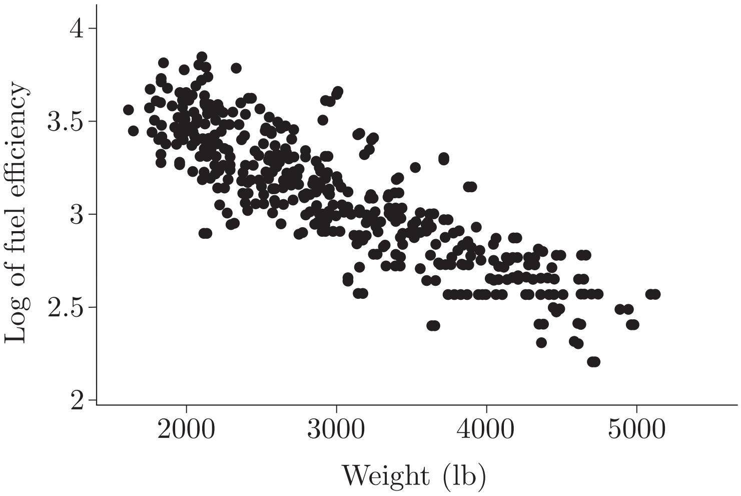


Figure 21 Log(fuel efficiency) against weight

Normal probability and residual plots, shown in Figure 22, suggest that the linear regression model is reasonable for these transformed data. (The only discrepancy from the linear regression model that is evident from these plots is a possible non-normality in the tails of the distribution of residuals in Figure 22(b).)

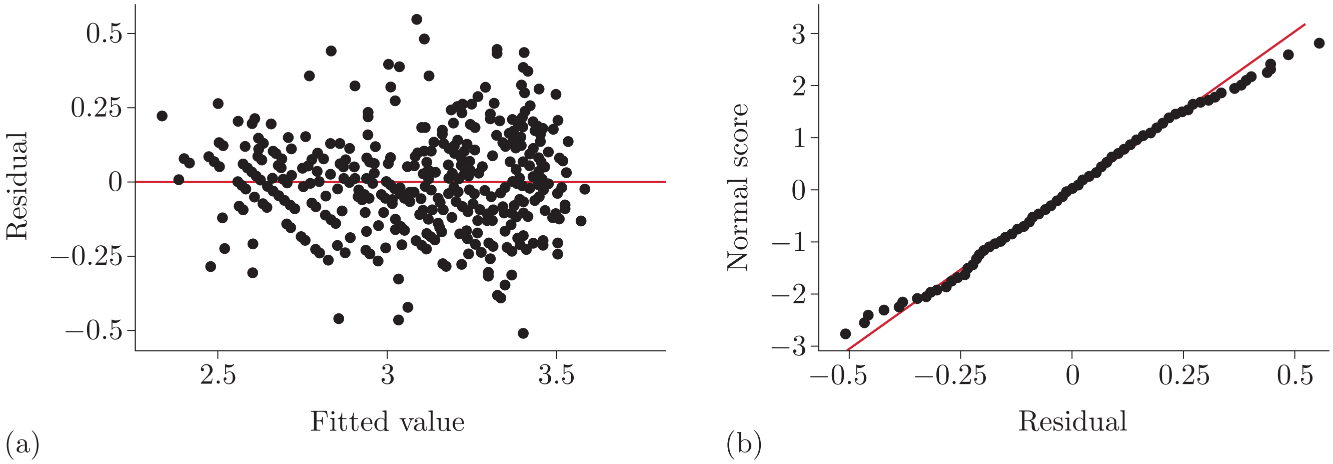


Figure 22 Regression of log(fuel efficiency) on weight: (a) residual plot; (b) normal probability plot of residuals

The line fitted by least squares has been overlaid on the transformed data in Figure 23. The equation of the fitted line is

$$y = 4.1445 - 0.000351\,x,$$

where y is the log of fuel efficiency and x is the weight (in lb). That is,

$$\log(\text{fuel efficiency}) = 4.1445 - 0.000351 \times \text{weight}.$$

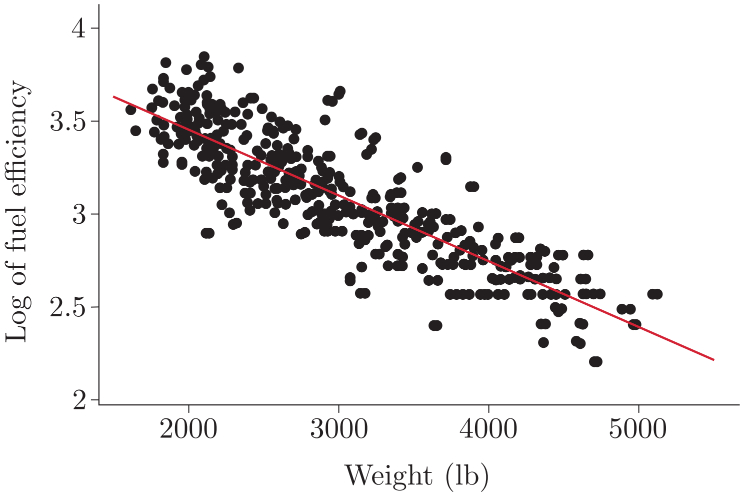


Figure 23 Log(fuel efficiency) against weight, and the least squares line

As in Subsection 2.1 where the explanatory variable was transformed, results from an analysis made on the data with transformed response variable will refer to the transformed data and not the original data. For example, in this case a unit (pound) increase in weight corresponds (on average) to a reduction in log(fuel efficiency) of 0.000351. Moreover, the following calculation allows us to interpret the effect of weight on the original scale. We have that

$$\begin{aligned}\log(\text{fuel efficiency}) - 0.000351 &= \log(\text{fuel efficiency}) + \log(0.9996)\\ &= \log(0.9996 \times \text{fuel efficiency}),\end{aligned}$$

$e^{-0.000351} = 0.9996.$

the last equality corresponding to use of Equation (1). Therefore a unit (pound) increase in weight corresponds (on average) to a reduction in log(fuel efficiency) of 0.000351, which corresponds in turn to multiplying the actual fuel efficiency by 0.9996; this is a reduction in fuel efficiency of 0.04% for each pound increase in weight.

Other transformations of the response variable are possible too, as in the next example.

Example 8 *Male elephants and their matings*

An eight-year study of the mating behaviour of African elephants was reported in 1989. As part of this study, data were provided on the number of matings (the response variable, y) of each of $n = 41$ male African elephants; this is plotted against the elephant's age (x, in years) in Figure 24. Note that because some of the elephants are of the same age and had the same number of matings, some data points in Figure 24 are jittered vertically, that is, displaced slightly from their true position in order to avoid plotting points on top of one another.

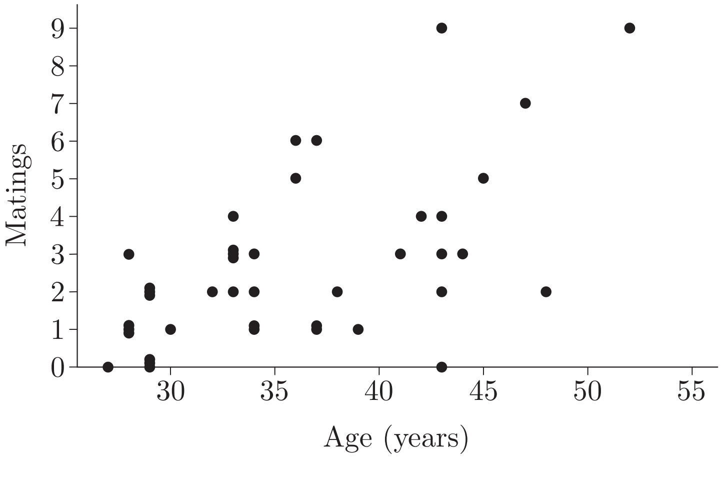


Figure 24 Number of matings against age

(Source: Poole, J.H. (1989) 'Mate guarding, reproductive success and female choice in African elephants', *Animal Behaviour*, vol. 37, no. 5, pp. 842–9)

A male African elephant

The number of matings clearly increases as age increases. But as was the case in Figure 20, a feature of Figure 24 is that the amount of variability in the response variable appears not to be constant, in this case increasing with increasing values of the explanatory variable. A transformation of the response variable again seems worth trying, therefore. You could contemplate taking logs, as in Example 7, but there's an immediate

You *could* remedy this by taking $\log(y + a)$ where $a > 0$, but how do you choose a?

problem: some elephants had no matings, and the log of zero is undefined. An alternative transformation that is defined at zero (and increasing for positive x) is the square root transformation, so let's try that. Figure 25 shows the transformed response variable $\sqrt{y}$ plotted against age (again with a little jittering). The effect of the transformation on the scatterplot seems at least to be an improvement: the amount of variability is certainly more constant across ages than it was in Figure 24 (especially if you think of the elephant aged 43 with no matings as an outlier and ignore him), and the increase is plausibly linear.

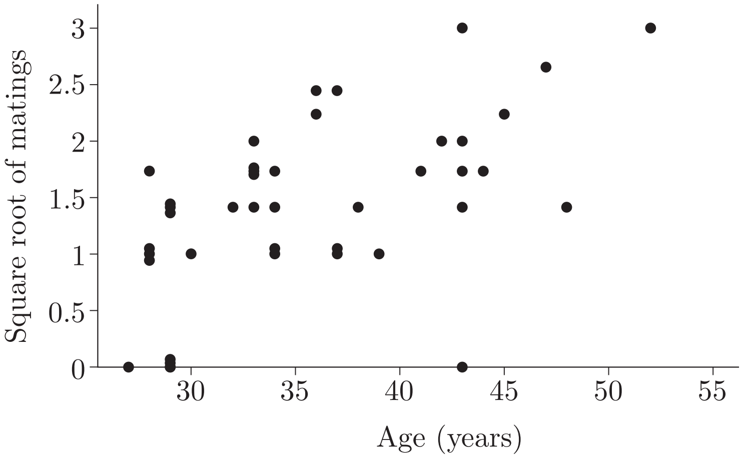


Figure 25 Square root of number of matings against age

A line is fitted by least squares to the data – in original form, no jittering! – and shown on the scatterplot (with jittered points) in Figure 26. In fact, normal probability and residual plots (shown in Figure 27) confirm, in perhaps a slightly surprisingly unequivocal manner, that the linear regression model is reasonable for these transformed data. The fitted line has the formula

$$\sqrt{y} = -0.812 + 0.0632\,x,$$

where y is the number of matings and x is the elephant's age.

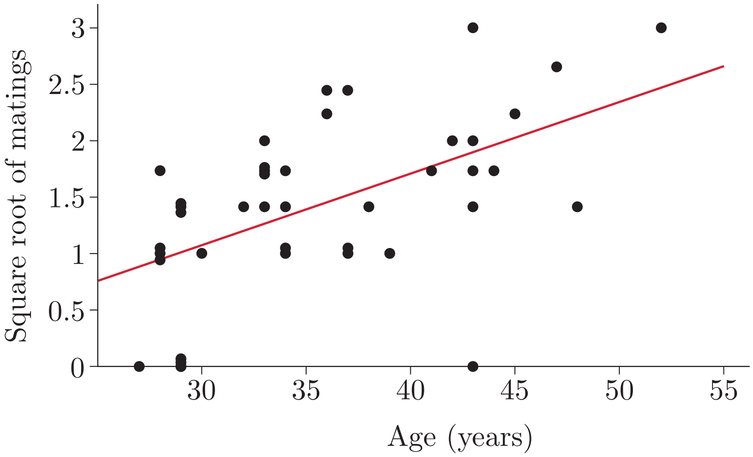


Figure 26 Square root of number of matings against age, and least squares line

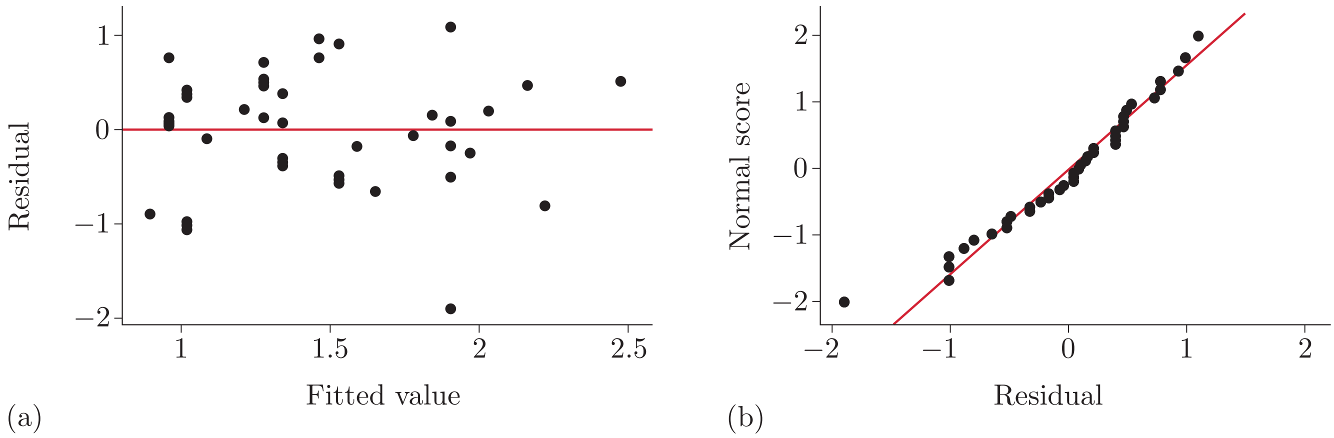


Figure 27 Regression of $\sqrt{\text{matings}}$ on age: (a) residual plot; (b) normal probability plot of residuals

Activity 6 *Predicted number of matings*

Using the model fitted to the data in Example 8, what is the point prediction of the number of matings that might be expected for a male elephant aged 40 years?

It is clear that, as was the case for transformation of the explanatory variable, experimentation with, and comparisons of, different transformations of the response variable might be carried out in practice, with the help of a computer. This can be done in Minitab, but you are not asked to make any such investigations in this subsection.

You might also have an objection to our treatment of Example 8, and you may well be right! The number of matings associated with each elephant is a discrete random variable, not a continuous one. The linear regression model as considered in this unit is designed for use with a continuous response variable. Discrete response variables can be accommodated in a generalisation of the linear regression model that is a topic for a more advanced module. However, in practice, approximating a discrete regression situation by transforming the response and pretending the result is continuous, as we have done in Example 8, is an alternative that remains usefully contained in a statistician's toolbox, at least for occasional use.

'Poisson regression' is one relevant generalisation.

2.3 Multiple regression with transformed variables

So far in this section, we have discussed how and why either the explanatory variable or the response variable can be transformed in linear regression with one explanatory variable. The same principles can be used in multiple regression where any number of the explanatory variables can be transformed (in order to improve the linear dependence on explanatory variables of the regression function) or the response variable can be transformed (in order to make the random terms have the properties required in a multiple linear regression model). The multiplicity of explanatory variables makes their transformation – the choice of both which explanatory variable(s) to transform and how to transform each of them – a more difficult problem than in the one explanatory variable case, so we will not investigate it here. On the other hand, transformation of the response variable adds little or no complication over that in the single explanatory variable case and so, for the opposite reason, also won't be investigated here!

Something of this sort might have underlain the decision to take logs of one of the explanatory variables in the economic growth example of Section 5 of Unit 11.

There is, however, a problem involving just one explanatory variable which linear regression with one explanatory variable can't cope with, but which *can* be tackled using multiple regression. An example of such a problem is provided by the data on the tensile strength of kraft paper, and its relation to the percentage of hardwood pulp used in its manufacture. These data were first considered in Activity 2 of Unit 11.

Example 9 *A model for the paper strength data*

Rolls of kraft paper in a workshop

Figure 7 of Unit 11 is repeated here as Figure 28. It shows values of the tensile strength of kraft paper (the response variable, y, in units of pounds per square inch, or p.s.i.) plotted against the hardwood content of the pulp from which the paper was made (the explanatory variable, x, in %).

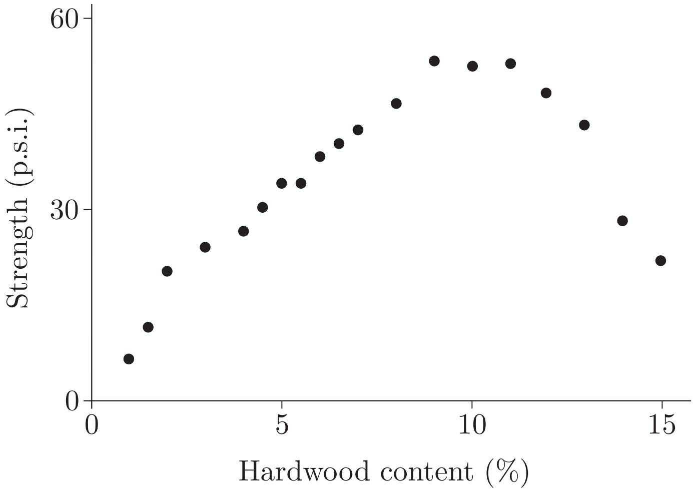


Figure 28 Tensile strength against hardwood content

As was briefly mentioned in the solution to Activity 2 of Unit 11, the up-and-down pattern to these data suggests that a possible model for the data might involve a quadratic function of x or perhaps a cubic function of x. The simpler of these is the quadratic, so let us consider a model like

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + W_i. \tag{4}$$

A less obvious way of writing the same model is as

$$Y_i = \alpha + \beta(x_i + \delta x_i^2) + W_i,$$

where $\delta = \gamma/\beta$. This makes it clear that if we try to transform the function of x in this model by $x' = x + \delta x^2$ we hit the same problem as in Example 6: the transformation involves an unknown parameter, δ. It appears that non-linear regression is needed.

In this case, however, there is an alternative approach: we can use multiple regression rather than non-linear regression! Look back to Equation (4). We can consider x^2 as a *second* explanatory variable in that regression model, alongside x. Let's set $x_1 = x$ and $x_2 = x^2$. Then the model in Equation (4) is a special case of the multiple regression model

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + W_i,$$

where β and γ in Equation (4) have been renamed β_1 and β_2, respectively. In this case, the two explanatory variables happen to be strongly related ... but there was no requirement in Section 5 of Unit 11 that they should be anything different!

Activity 7 *Models for the paper strength data*

(a) A multiple regression model for the paper strength data was fitted, with the tensile strength of kraft paper as the response variable y, the hardwood content of the pulp from which the paper was made as the first explanatory variable, x_1, and the square of the first explanatory variable as the second explanatory variable, x_2. The fitted model is

$$y = -6.67 + 11.76\, x_1 - 0.6345\, x_2.$$

The residual plot for this model is given in Figure 29.

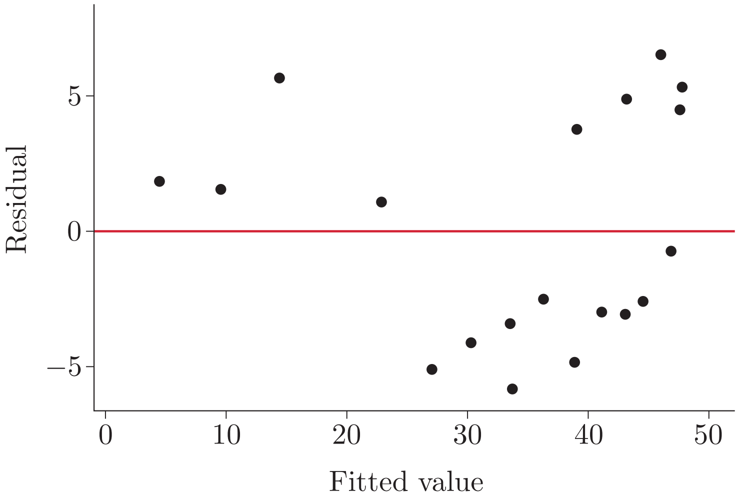


Figure 29 Quadratic model: residual plot

On the basis of the residual plot, do the model assumptions seem reasonable? If not, why not?

(b) Having considered and discarded a quadratic model for the paper strength data in part (a), how about a cubic model instead? In multiple regression terms, the model is of the form

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + W_i,$$

where, in addition to all the ingredients of the multiple regression model using a quadratic curve, $x_{i3} = x_i^3$ represents the cubes of the values of x_i. Such a model was fitted to the data, the fitted model being

$$y = 5.65 + 3.58\,x_1 + 0.654\,x_2 - 0.0552\,x_3.$$

The residual plot and the normal probability plot of the residuals are given for this model in Figure 30.

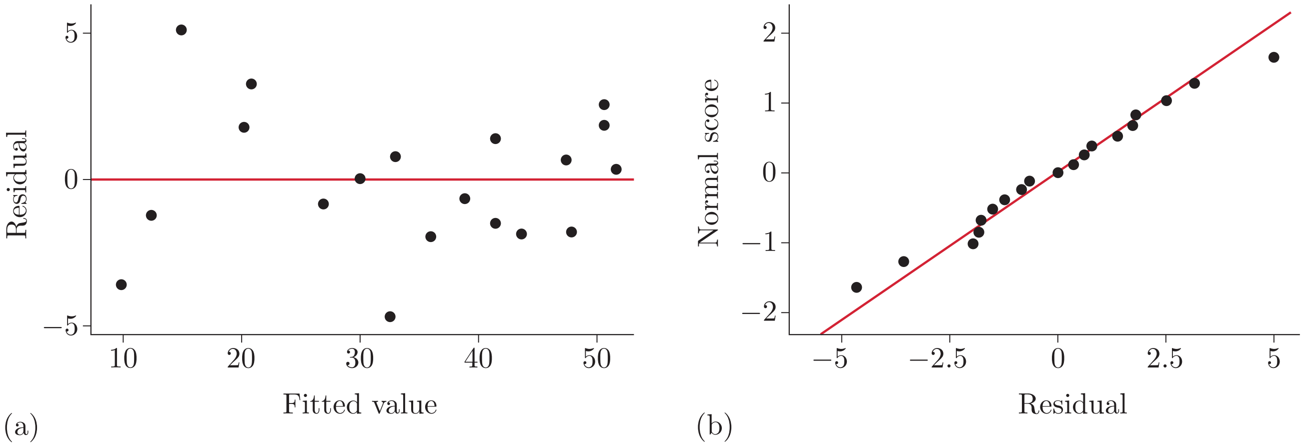


Figure 30 Cubic model: (a) residual plot; (b) normal probability plot of residuals

On the basis of these plots, do the model assumptions seem reasonable? If not, why not?

(c) Suppose that a new batch of kraft paper was to be produced using pulp with a hardwood content of 10%. Using the fitted cubic model, what is your prediction of the tensile strength of that paper?

It is useful to be reminded what has just been achieved in Activity 7. By using multiple *linear* regression, we have managed to fit the *cubic* regression model

$$y = 5.65 + 3.58\,x + 0.654\,x^2 - 0.0552\,x^3,$$

where y is the tensile strength of kraft paper and x is its hardwood content. The non-linearity of the fitted curve is accentuated by plotting it on top of a scatterplot of the data, shown in Figure 31.

Unfortunately, only non-linear functions with a certain specific structure – in fact, those that can be written as linear combinations of functions of x – can be accommodated by multiple linear regression in this way.

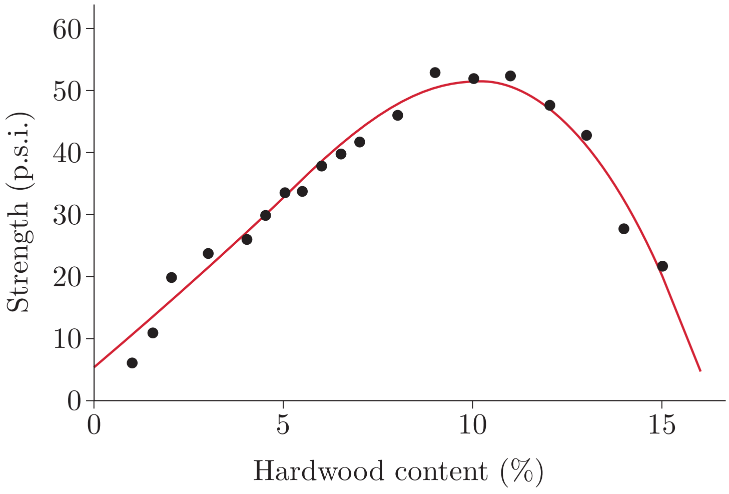


Figure 31 Tensile strength against hardwood content, and fitted cubic regression function

Exercise on Section 2

Exercise 3 *Residual plots and transformations*

Suppose a dataset is made up of measurements of two variables on each of a number of individuals, and we are in the usual regression situation of wishing to model the behaviour of one of the variables, the response variable, as a function of the other, the explanatory variable. Figure 32 is a repeat of Figure 20 of Unit 11. It shows four typical shapes of residual plots produced after fitting a line to such data by least squares.

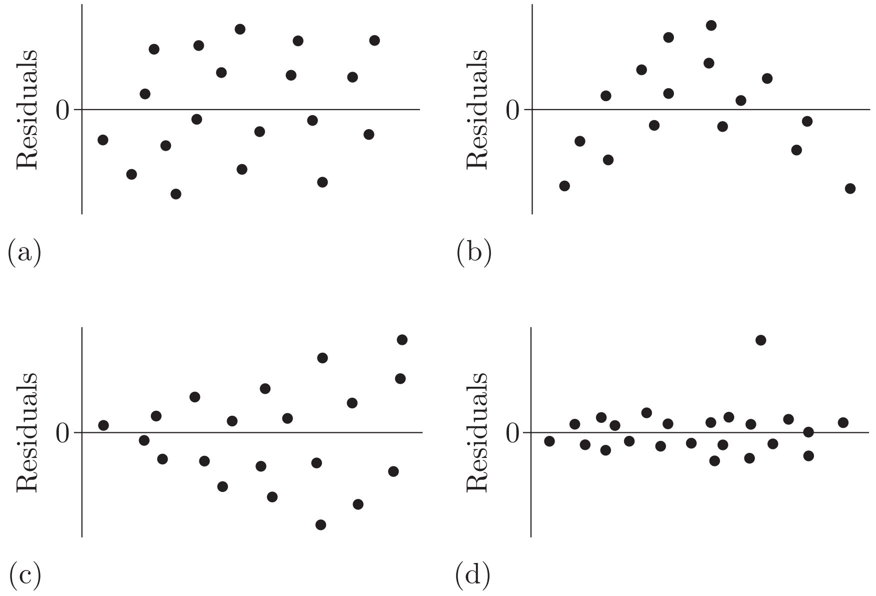


Figure 32 Four residual plots

For each panel of Figure 32 in turn, state whether or not you would transform the data in order to obtain a set-up better aligned with a standard linear regression model, and if you would, which variable you would choose to transform. Give reasons for your answers.

3 The modelling process

The beginning of most statistical investigations is usually a practical problem. For example:

- a medical researcher might want to know whether or not a treatment for cancer works
- an engineer might wish to estimate the tensile strength of a particular material
- a social scientist might seek to understand what factors influence school performance
- an economist might wish to predict future inflation rates.

In a statistical investigation, the problem is formulated in statistical terms, appropriate data are collected and analysed, and the conclusions are summarised in a statistical report. The journey from practical problem to statistical report is best thought of as a research process, which can be represented by the flow chart in Figure 33. Note that, in practice, the various stages of the statistical modelling process might arise in a slightly different order from that in Figure 33. For example, it is sometimes more convenient to check assumptions after the model has been fitted.

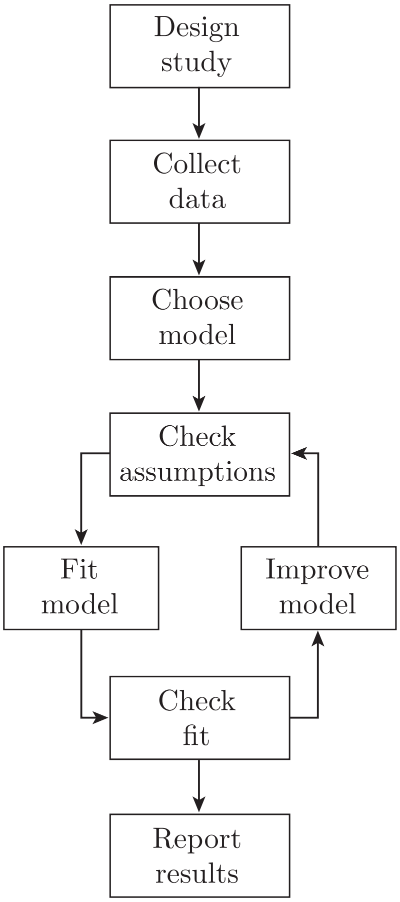


Figure 33 The statistical modelling process

Formulating the right questions, and designing studies to answer them, are important statistical issues, though they are not dealt with in this module. Typically, they involve collaborations with specialists in other disciplines – medical doctors, engineers, social scientists, economists, and so on – and usually require some knowledge of the particular application area. However, this unit focuses on the issues of statistical modelling and reporting, which are similar in all application areas. The starting point is therefore the problem or question under consideration, together with the data that have been collected to throw light on it. Thus, in terms of the flow chart in Figure 33, you will begin at the box marked 'Choose model'.

What is a model? Well, as you have seen, in general terms, it is a simplified representation of the process generating the data. A key component of a statistical model is the underlying distribution from which the data are sampled; but the model might also include other components – for example, transformations or regression relationships between variables. However, the terms 'distribution' and 'model' are used interchangeably in much of the remainder of this unit, as they have been throughout the module.

A suitable model to start with is one that reflects the most important attributes of the data. For example, if the data consist of measurements on a continuous variable, then it makes sense to choose a continuous distribution to represent the underlying variation. Also, the question to be answered will often suggest how the model will be used – for example, to calculate a confidence interval or to carry out a hypothesis test. Having chosen a model, you will need to check that it fits the data, and that any assumptions required are satisfied. If either is not the case, then you will need to alter the model in some way, or perhaps even try a completely different one. You will then need to repeat the process, improving the model at each stage until it is good enough for its purpose. Having chosen a model, the final stage of the modelling process is to report your results.

The modelling process

Statistical modelling is, to some extent, an art as much as a science, requiring common sense and judgement and, just occasionally, a little inspiration. It is as well to remember that statistical models are at best idealisations of reality: you should not expect to find a 'perfect' model. The real skill is in finding a model that is good enough for your purposes, and from which you can draw valid conclusions.

The remainder of this unit pans out like this. First, in the remainder of this section, we will briefly collect and review *some* of the tools that were put in your statistical toolbox earlier in the module (and hence can be considered as revision, to some extent). However, the focus here is not so much on the individual methods or techniques but on using them to choose the most appropriate method(s) of analysis for the data and questions of interest. The opportunity to put into practice the statistical skills you have acquired will then be provided in a single substantial example in Section 4, while appropriate ways to go about statistical report writing will be described and practised in Section 5.

More practical applications await you in Unit 13.

3.1 Choosing a model: getting started

In this subsection, some guidance is provided about how to approach the task of selecting a model. It is important to remember that there are no fixed rules for this, only general principles – and even they, on occasion, might reasonably be set aside.

Decisions, decisions . . .

It is recommended that, as part of the process of choosing a model, you explore the data using graphical methods. The graphical displays first discussed way back in Unit 1 are, therefore, very relevant at this stage of the statistical modelling process. Histograms and, latterly, scatterplots are the types of graph that have been used extensively 'to get a feel for the data' in the module to date; bar charts and boxplots have similar roles, but have been used a little less. Normal probability plots have also been used extensively in the module but not so much at this initial stage of the process; rather, they have been particularly used to contribute strongly to the 'Check fit' stage of the modelling process.

In conjunction with looking at the data using graphical methods, we also wish to emphasise the importance of the setting or context in which the data are obtained, and the type of data collected, in structuring your ideas. For the purposes of statistical modelling, the first major distinction to be drawn is whether the data should be modelled as being discrete or continuous, as discussed in Units 1 and 2. In many examples, the choice is reasonably clear. Thus even before seeing any data you may be able to narrow down your choice of model to one suited to discrete data or one suited to continuous data. These can then be refined further – or perhaps set aside – after looking at histograms, bar charts or other graphical displays. (The distinction between bar charts and histograms is also, you should recall, that one is appropriate to discrete data, the other to continuous data.)

The distinction between discrete and continuous data is fundamental but it is worth being reminded that the distinction is not always clear-cut. For example, on hearing that one's data consist of counts, a discrete model would be expected. However, arguably, this is premised on the assumption that the counts are low integers $0, 1, 2, 3$, and so on. But what if the counts are typically in the hundreds? The underlying distribution is still discrete, but the data might reasonably be modelled as continuous, as an error of one in hundreds could be considered to be negligible. Conversely, measurements on continuous variables are often rounded to some fixed number of decimal places, and so may be regarded as discrete. If the rounding can be ignored, it is reasonable to treat the rounded measurements as continuous. On the other hand, in some cases, the measurement might be very crude and the data may then best be regarded as discrete. These points are illustrated in Example 10.

Example 10 *Shipwrecks*

In the nineteenth century, there were no fewer than 177 shipwrecks recorded along the stretch of coast from Pevensey to Rye in East Sussex. (Source: Renno, D. (2002) *East Sussex Shipwrecks of the 19th Century (Pevensey-Hastings-Rye)*, Sussex, Book Guild.) Consider the problem of modelling the times between successive shipwrecks, the dates on which they occurred having been recorded. The time between successive shipwrecks is measured in days, so this could arguably be regarded as discrete. However, it could equally be argued that the rounding to the nearest day is immaterial, that many of the intervals between shipwrecks comprise large numbers of days and that, for all practical purposes, the data should be treated as continuous. In situations such as this, the data can be treated either as continuous or as discrete, and both approaches should lead to similar results.

Other important aspects of shipwrecks might be their size, the wrecked vessels ranging from rowing boats to large cargo ships, and the severity of the shipwreck in terms of lives lost and/or goods destroyed. Number of lives lost is clearly a discrete count. Goods destroyed is, in principle, a continuous variable, whether measured in tonnage or financial value. However, there are many reasons why, for such historical events, precise amounts of goods lost (and sometimes even of lives lost) are not known. It might therefore be more reasonable to 'discretise' the continuous variable 'amount of goods destroyed' to, say, one of a small number of approximate round numbers of pounds (or even to discretise lives lost, for some purposes, to a binary variable representing 'some' or 'none').

Example 10 illustrates the point that clear rules even about the apparently simple matter of deciding on a discrete or continuous model can be difficult to specify. When the error involved in treating a continuous variable as discrete, or vice versa, is negligible, then it may not matter which choice is made. In this case, the choice might reasonably be made on the grounds of convenience, and how well the proposed model fits the data.

We, perhaps, claimed an element of this in our treatment of the elephant mating data in Example 8.

3.2 The models at your disposal

In this module, you have met ten families of probability models. They are listed in alphabetical order in Table 1, along with the number of the unit and section or subsection thereof in which they were introduced and primarily discussed.

Table 1 Distributions and where to find them in M248

Distribution	Unit	S(ubs)ection
Bernoulli	3	1.1
binomial	3	2
chi-squared	10	2.2
continuous uniform	3	5.2
discrete uniform	3	5.1
exponential	5	2.2
geometric	3	3.1
normal	6	all
Poisson	3	4
t	8	4.2

Activity 8 *Which are discrete and which are continuous?*

For the ten distributions listed in Table 1, identify which are discrete distributions and which are continuous distributions.

Points made in Subsection 3.1 and the first part of this subsection are summarised below.

Choosing a model: discrete or continuous?

- If the random variable X is discrete, then you might choose a discrete distribution – for example, Bernoulli, binomial, discrete uniform, geometric or Poisson.
- If the random variable X is continuous, then it usually makes sense to choose a continuous distribution – for example, chi-squared, continuous uniform, exponential, normal or t.
- In some circumstances, it is appropriate to model a continuous variable as discrete or a discrete variable as continuous. This is typically the case when the error involved in doing so is negligible.

Having narrowed the field to either discrete models or continuous models, the next step is to choose which of the models in each of these categories is most likely to be suitable. Let us consider discrete distributions first.

3.2.1 Discrete models

The Bernoulli model can be regarded as a special case of the binomial model with $n = 1$. As mentioned when it was introduced in Unit 3, the Bernoulli distribution is the only possible distribution for data taking just two values. Thus the choice of discrete models available to you in other discrete situations is really between the binomial, discrete uniform, geometric and Poisson distributions. Choosing between these can be helped by prior understanding, knowledge or intuition about the process generating the random variable X which you wish to model. In particular, the four discrete distributions were introduced – all in Unit 3 – in specific settings; and each setting may be regarded as the standard one for this model. If your data were collected in such a setting, then it makes sense to try the corresponding model.

There are many other discrete distributions; these are just the ones you have met in this module.

Choosing a model: standard settings for discrete distributions

- If X may be regarded as the number of successes in some known number n of independent Bernoulli trials with constant probability p of success at each trial, then choose the binomial distribution $B(n,p)$.
- If X has a finite range and every outcome has the same probability, or at least is believed to be equally likely, then try the discrete uniform distribution.
- If X may be regarded as the number of trials up to and including the first success in a sequence of independent Bernoulli trials with constant success probability p, then choose the geometric distribution $G(p)$.
- If X may be regarded as a count of a number of events, then try the Poisson distribution. This applies in particular if events are believed to occur at random and X is the number of events that occur during intervals of fixed length.

These derivations of the models are not guaranteed to apply in practice – in particular, you will need to check the assumptions required in each case. On the other hand, the 'mechanism' underlying the data at hand might be unclear and, for example, a geometric distribution might still prove to be a better model for a certain set of counts than a Poisson distribution, even if the assumptions associated with waiting times in Bernoulli trials are not met. In such circumstances, the distribution must be chosen on *empirical* grounds; that is, your choice of distribution can be guided by the range and shape of the distribution.

Below is a reminder of the ranges and shapes of the binomial, discrete uniform, geometric and Poisson distributions.

Choosing a discrete model: range and shape

- Binomial, $B(n,p)$: finite range $\{0,1,\ldots,n\}$; one mode, which can take any value within the range (depending on the value of p); symmetric for $p=1/2$, left-skew for $p>1/2$, right-skew for $p<1/2$.
- Discrete uniform: finite range; constant p.m.f.; no mode.

- Geometric, $G(p)$: range $\{1, 2, 3, \ldots\}$, unbounded to the right; decreasing p.m.f. so that its mode is always at 1.
- Poisson(λ): range $\{0, 1, 2, \ldots\}$, unbounded to the right; one mode, which can take any value within the range (depending on the value of λ), including decreasing p.m.f. when $\lambda < 1$.

In some cases, no distribution fits all the requirements. Even then, all may not be lost. First, the models as listed often provide a convenient starting point. Second, it is important to remember that the purpose of statistical modelling is not to find a perfect model, but to find a 'good enough' model. Provided that the model does not fail in some key respect, it might still be useful.

The following quirky example and its associated activity illustrate some of the above considerations.

Example 11 *Reusing reusable envelopes*

In 1990, William Sutherland worked in a large organisation in which internal notes and memoranda were sent in reusable envelopes. This type of envelope is little used these days, since communication via email and other electronic methods has become more convenient; examples of such reusable envelopes are shown in the picture accompanying this example. Here's how they work. Each envelope has a number of spaces (windows) for the names of recipients; the type considered by Dr Sutherland had twelve such windows. New users cross out their own name, and write in the next window the name of the person they wish to contact. Dr Sutherland kept a count of how many names, including his own, were written on some of the envelopes he received. Notice that the first window is always filled. Also, the probability that a window is filled depends on the position of the window: those higher up the envelope are more likely to have been filled than those below (since people usually used up the windows in sequence).

Reusable envelopes of the type considered in Example 11

The purpose of the analysis is to describe the distribution of the number of used windows, and thus to obtain some idea of the age structure of envelopes in circulation.

Activity 9 *Models for numbers of used windows on envelopes*

Consider the setting described in Example 11. The data are counts of the number of used windows on envelopes, ranging from 1 to 12, so clearly they are discrete.

(a) For each of the four discrete distributions in the box above, consider the standard settings underlying each distribution and the range of each distribution to see if one of them might fit this problem.

(b) So far, the discussion of this example has not involved any data, just a rather abstract discussion of the setting. Figure 34 shows a bar chart of the numbers of used windows for a sample of 311 windowed envelopes.

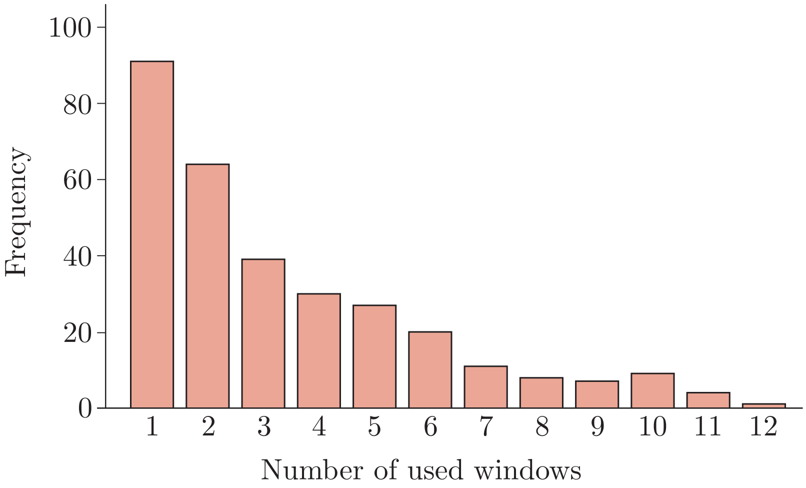


Figure 34 Numbers of used windows

(Source: Sutherland, W. (1990) 'The great pigeonhole in the sky', *New Scientist*, 9 June, vol. 1720, pp. 73–4)

Do you think a discrete uniform model is appropriate for these data? What model (or models) does the shape of the data suggest?

(c) The model preferred in the solution to part (b) can be checked using a chi-squared goodness-of-fit test. This was done after suitable combination of cells with expected frequencies less than 5, yielding a chi-squared test statistic of 13.646 on 9 degrees of freedom, and hence a p-value of 0.135. Does the model fit the data?

It seems that, from Activity 9, a geometric distribution provides a good model for the envelope windows data despite the facts that we failed to identify an underlying Bernoulli process and that the geometric distribution has an unbounded range, whereas the number of used windows can be no greater than 12. This illustrates the point that it is important to keep an open mind when modelling and concentrate on the important aspects of the data, rather than just focus on the limitations of the distribution: in this case, getting the shape right is probably more important than abiding by constraints about the range of the data. What is more, the fitted geometric model happens to give a value for $P(X > 12)$ of 0.016. This is indeed quite a small probability of obtaining values greater than 12. And, in this particular situation, a twist might be that, occasionally, if all the windows on an envelope were full, the envelope would still be re-used, with an extra window or windows drawn on by users. The geometric model might therefore be an even more appropriate model for these data than it appeared to be!

An over-full reused envelope!

3.2.2 Continuous models

Choosing an appropriate continuous distribution follows much the same process as choosing a discrete distribution, with one major exception: if none of the available distributions seems appropriate, then you can try transforming the data (as discussed in Section 1). For the time being, consider the standard continuous distributions you have met so far. These are the chi-squared, continuous uniform, exponential, normal and t-distributions. The chi-squared distribution and t-distribution were introduced specifically in the context of statistical tests. Nothing precludes you from using them for modelling, and they are increasingly used in this way in the modern world, but in this module only the continuous uniform, exponential and normal distributions are used for this purpose.

There are many other continuous distributions; these are just the ones you have met in this module.

As for the discrete distributions above, there are standard settings for these continuous distributions. These can help you to select a candidate distribution.

Choosing a model: standard settings for continuous distributions

- If X takes values between a and b, and each value in the interval $a < x < b$ is equally likely, or believed to be equally likely, then try the continuous uniform distribution $U(a, b)$.
- If events are thought to occur at random in time and X is the waiting time between successive events, then try the exponential distribution.
- The normal distribution is a good first choice when X is clustered around a central value, and is as likely to lie below as above this value. It is also likely to be suitable if you can perceive your data values as being means of other values.

Bringing transformations back into the mix, the box below summarises the ranges and shapes of the continuous distributions available to you. (Apart from the transformation item, the information is the same as in the box at the start of Section 1.)

Choosing a continuous model: range and shape

- Continuous uniform: finite range $a < x < b$; constant p.d.f.; no mode.
- Exponential distribution, $M(\lambda)$: range $0 < x < \infty$, unbounded to the right; decreasing p.d.f.
- Normal distribution, $N(\mu, \sigma^2)$: unbounded range $-\infty < x < \infty$ and is symmetric about a single mode that coincides with the mean; values far from the mean have low probability.
- For data on any range, a suitable transformation of the data *might* be modelled by a normal distribution.

The following activity illustrates some of the above considerations. It also serves to illustrate the important point that models are at best approximate representations of reality. The aim is to formulate a reasonable model, not a perfect one.

Activity 10 *Heredity and head shape*

How similar are the head shapes of these two famous first and second sons, Princes William (right) and Harry (left)?

In a study of 25 families where there were at least two sons, measurements were taken on the head length and head breadth of the first and second sons. Head size can be measured as length + breadth, head shape can be measured by the head shape index calculated as 100 × (breadth/length). One issue of interest is whether there is a difference between the head shapes, as just defined, of first and second sons.

Use some of the above considerations to identify suitable candidate models for the following variables on head size and shape.

(a) Head size, that is, length + breadth.

(b) Head shape, that is, 100 × (breadth/length).

(c) Head shape difference, measured as head shape of first son minus head shape of second son.

Before ending this subsection, it is worth emphasising that there are many more distributions than those covered in detail in this module, although this module does include some of the most important ones in statistics. Minitab provides several other distributions, which you might care to explore (this is entirely optional); and there are many others beside these. However, the approach to selecting an appropriate distribution is much the same whatever the collection of distributions to which you have access.

3.3 Dealing with outliers

You are already aware of the notion of statistical outliers; in this subsection, we seek to give a little more advice about how to deal with them in practice. Outliers are particularly disconcerting if they are found in very small datasets, as in Example 12.

Example 12 *Radiocarbon age determinations*

The data in Table 2 are a set of radiocarbon age determinations, in years, of eight samples from the same stratigraphic layer of a site at Lamoka Lake, New York, USA. The samples should all come from the same period, that of the earliest occupation of the site by ancient people of the Lamoka culture, and so should be, at least approximately, the same age.

Most determinations in Table 2 suggest an age of between 2400 and 2600 years. However, sample C-367 indicates an age of 3433 years, which is quite out of step with the other values: this sample is a clear outlier.

Table 2 Radiocarbon dating

Sample number	Radiocarbon age determination
C-288	2419
M-26	2485
C-367	3433
M-195	2575
M-911	2521
M-912	2451
Y-1279	2550
Y-1280	2540

(Source: Long, A. and Rippeteau, B. (1974) 'Testing contemporaneity and averaging radiocarbon dates', *American Antiquity*, vol. 39, no. 2, pp. 205–15)

If at all possible, when outliers are present, your first step should be to check that they are not the result of recording, coding or data entry errors. Such errors are very common. It is well worth repeating here that if you enter your own data, you should always check your computer data file against the original. Not all data entry errors will necessarily appear as outliers!

The study of outliers and how to treat them can be rather complex, so only a little general guidance will be given in this module. Broadly speaking, the treatment of outliers depends on how many appear in the data, what effect they have on the conclusions, and how far you are prepared to go in believing that you have been unlucky enough to obtain a few 'atypical' values, rather than believing that the distributional assumptions are not viable. This last point is important: the outliers might just reflect the fact that you have chosen the 'wrong' model. The effect of model choice on outliers is illustrated in Example 13.

We offer no specific advice about when a data point is sufficiently unusual to be classified as an outlier; as with much of statistics, this question is not clear-cut, its answer depending on context and purpose.

Activity 11 *Treatment duration*

The duration (in days) of treatment was recorded for a set of 86 long-term hospital patients. Treatment can last several months, so it makes sense to model these data as continuous; they are also positive. But which model is appropriate? Despite the positivity of the data, the normal model might be a good one, if there is a 'typical' treatment length around which the values cluster and, as suggested above, treatment durations tend to be long (corresponding to a normal model with negligible probability of negative durations). Alternatively, you could think of treatment durations as waiting times, which might suggest that an exponential model may be suitable.

A hospital main entrance: the way out as well as the way in!

Let us now consider the actual data. Figure 35 (overleaf) shows a frequency histogram of treatment times; the sample mean of the data is 122.3 days and the sample standard deviation is 146.7 days.

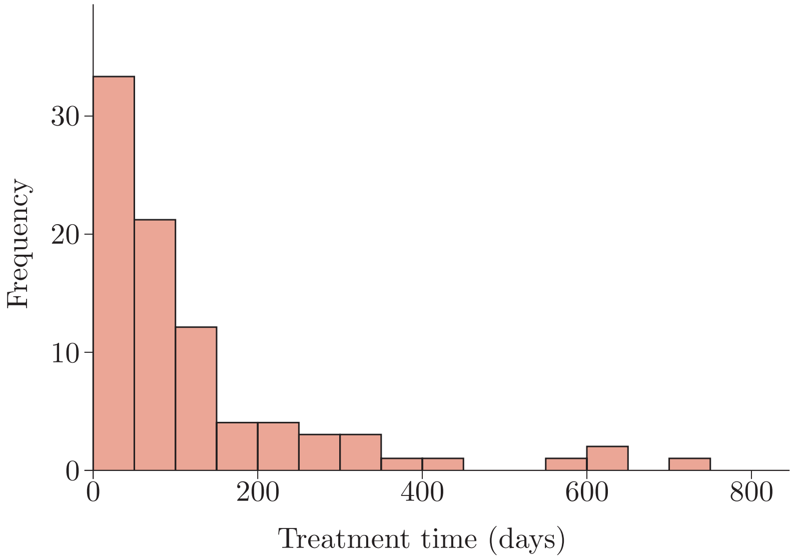


Figure 35 Treatment times for hospital patients

(Source: Copas, J.B. and Fryer, M.J. (1980) ‘Density estimation and suicide risks in psychiatric treatment’, *Journal of the Royal Statistical Society, Series A*, vol. 143, no. 2, pp. 167–76)

(a) On the basis of the histogram, do you think that an exponential model is likely to be a good model for these data?

(b) In Subsection 2.2 of Unit 5, it was noted that the mean and standard deviation of the exponential distribution are equal, which ‘provides a method for checking quickly, given data, whether an exponential model is worth considering’. On the basis of the given sample statistics, do you think that an exponential model is likely to be a good model for these data?

Example 13 *Treatment duration, continued*

In this example, we take consideration of the treatment duration data introduced in Activity 11 somewhat further. From the histogram in Figure 35 and by looking at the mean–standard deviation relationship, you may well have concluded that an exponential model might be a reasonable one for these data. Using the reciprocal of the sample mean (which is the maximum likelihood estimate) as an estimate of the exponential parameter λ leads to the fitted exponential distribution superimposed on top of the unit-area version of the histogram of Figure 35, which is shown in Figure 36.

Arguably, this figure suggests that the exponential distribution fits the main body of the data very well, but there are four data values – those making up the two little bumps to the right in the histogram – that are not in line with this distribution. (Further analysis using tools that you have not met in M248 makes this observation even clearer.) So, relative to the behaviour expected under an exponential model, it seems that there may be four outliers in this dataset. These outliers could be interpreted as relating to an atypical group of patients with unusually long treatment times.

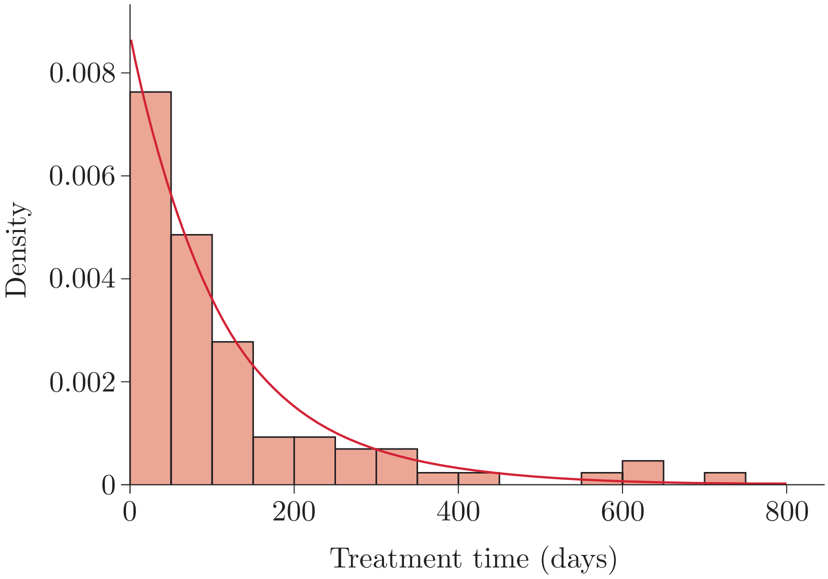


Figure 36 Unit-area histogram of treatment times with fitted exponential distribution

However, if the transformation $x^{1/4}$ is applied to the data, then the corresponding frequency histogram, which is shown in Figure 37(a), is roughly symmetric; and the normal probability plot in Figure 37(b) suggests that a normal model might be appropriate for the transformed data.

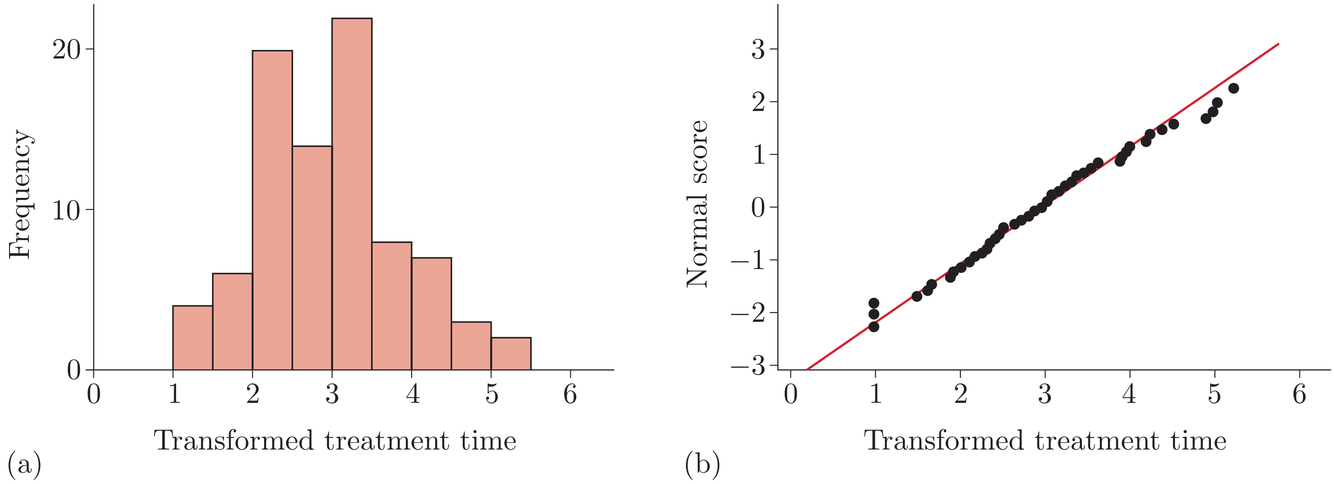


Figure 37 Transformed data: (a) histogram (b) normal probability plot

There is no strong suggestion of a problem at longer treatment times now! There are three ill-fitting points with treatment times of one day. Since times are counted in whole days, it is possible that some of the patients to which these times relate were treated for a slightly longer or slightly shorter time than one day. The poor fit in the lower tail might thus just be the result of crude rounding of the data.

So which model is 'correct'? Are the untransformed data exponentially distributed, with a group of four atypical patients with long treatment times? Or are the transformed data normally distributed? It is not possible to settle the issue without further information, in particular about the four possibly atypical patients, or a lot more data. As it stands, both models for analysis are equally reasonable. In any case, which model to use would depend on the purpose of the analysis.

An important consideration in dealing with outliers is the purpose of the analysis. If there are relatively few outliers, and the model you have selected appears to fit the rest of the data reasonably well, then a sensible procedure is to examine the sensitivity of your conclusions to the outliers. This is easily done by undertaking two analyses, one including the outliers and the other excluding them. If your conclusions do not differ substantially under the two procedures, then the outliers are not influential and should not be a major source of concern. In this case, you should report the presence of the outliers, and state that the conclusion reached is not sensitive to inclusion or exclusion of these outliers. On the other hand, if the outliers do have a big impact on the conclusions, then it can be appropriate to report your findings with the data analysed both ways.

Lamoka Lake

Activity 12 *Age of the Lamoka Lake site*

The radiocarbon data of Example 12 are to be used to provide an estimate of the age of the Lamoka Lake site. Provide such an estimate using all of the data points, and investigate its sensitivity to the outlier. How would you report your results?

There are no hard and fast rules to decide how many data values may be deleted in order to salvage a particular modelling assumption. In practice, it is best to remove only a very few values.

If in doubt, an alternative is to keep all the values and revert to a distribution-free method. Using ranks instead of data values loses information about how far apart the values are but, on the other hand, it removes sensitivity to abnormally large or abnormally small values.

If decisions about which method to use seem unduly vague, you should remember that there is not always a definitely right or wrong way of performing a statistical analysis. All you can do is use your common sense.

4 Modelling with Minitab

In this section, you will have the opportunity to tackle an exercise, or mini-project, in statistical modelling using Minitab. The aim is to give you practice in the skills of statistical modelling, pulling together a number of things that you have learned about in the module and been reminded about in this unit. The mini-project begins with some background, a scientific question and a description of a dataset relevant to the question. You will then progress through the various stages of exploratory analysis, model and method choice, model checking, and performing relevant statistical calculations.

Refer to Chapter 5 of Computer Book C for the work in this section.

5 Writing a statistical report

A statistical investigation usually begins with a practical problem and ends with the results being summarised in a statistical report. The stages involved were discussed briefly in Section 3 and represented in a flow chart in Figure 33. In this section, some general advice is given on how to write – and how not to write – a statistical report.

The statistical report is an account of what you did, why you did it, what you found, and what your results mean in terms of the original scientific question. The main challenge in writing a report is that it is aimed at two very different readerships. First, it should be sufficiently detailed to allow other statisticians to understand clearly what you did, and enable them to assess the validity of your conclusions. But it is equally important that it should provide a non-technical account of your investigation to non-statisticians who are interested primarily in the original question. These distinct aims are usually reconciled by writing a non-technical summary of your investigation to accompany the more technical report.

The report is also important for your own record.

It is also important to stress that the report should be succinct and, if possible, short. A long-winded, rambling document is of little use to anyone!

In Subsection 5.1, the structure of a typical statistical report is described, and in Subsection 5.2 an example is discussed in detail. You will then have the opportunity to assemble some statistical reports yourself.

5.1 The structure of a statistical report

The key to a good report is its structure. This makes for easy reading. For example, a non-specialist might ignore the more technical parts of the report dealing with statistical methods, and read only the summary and the discussion. Also, structure makes a report easier to write, as it helps to organise the material.

A possible structure for a statistical report is set out in the following box.

The structure of a statistical report

A statistical report comprises the following sections.

- Summary
- Introduction
- Methods
- Results
- Discussion

This structure is reasonably standard, though some authors might use different section headings – for example, *Abstract* instead of *Summary*, *Background* instead of *Introduction*, *Conclusions* instead of *Discussion*, and so on.

The *Summary* should be completely self-contained. It should state briefly the aim of the analysis, the method used, the key finding or findings, and the interpretation. It is usually written last, and should use largely non-technical language. The 'largely' in the previous sentence is a reflection of the fact that it is often simply not possible to provide an accurate summary of results without using some statistical terminology or referring to some statistical concepts. It is far better to give a slightly technical, but correct, summary than one apparently easily understood by all, but potentially misleading.

The *Introduction* should contain a brief description of the question or hypothesis to be investigated, the setting in which the data were collected, and the data available. Note that, in this module, the starting point is always a problem or question, and some data relevant to that problem or question.

The *Methods* section should include a description of the model, the procedures used to check the model, the statistical tests employed, the method used for calculating confidence intervals, and any other relevant techniques you have used, such as data transformations. The key guide to this section is to include enough detail to allow other statisticians to evaluate your method, and to repeat your investigation if they had the same data. You should not include all the blind alleys and dead ends you travelled (we all travel them) before settling on your preferred solution. However, if you found two equally plausible models that give appreciably different results, then you should include both.

A rather beautiful blind alley in Florence, Italy

The *Results* section should contain descriptive summaries of your data (for example, graphical and numerical summaries), evidence that your model is appropriate and, finally, the numerical results of statistical tests or confidence interval calculations. Sometimes, it might be appropriate to round your numerical results further when reporting them in the *Results* section. It is important to remember that this section, as all others, should be written in prose: a collection of numbers and graphs is not sufficient.

The *Discussion* should contain your own assessment of the statistical evidence relating to the original question or hypothesis. In particular, you should discuss any evidence of lack of fit of your model, any problems with the data (for example, outliers), or any other matter that might have a bearing on the interpretation of the results.

There is no set order in which to write the sections of a report but you should present the sections in the order just described. Many readers will not read all the sections – for example, many will read only the *Introduction* and *Discussion* – so it is important to structure your report so that they can find the sections they are interested in quickly. In some sense, the *Results* section forms the heart of the report. The *Methods* section is organised in such a way as to explain how you obtained your results, while the *Discussion* is your interpretation of the results. Some authors prefer to write the *Results* section first, followed by the *Methods* section. You should use whatever order you feel most comfortable with. In any case, you will probably find yourself going back over previous sections to make sure everything fits together in a coherent whole.

Finally, one important general rule: the shorter, the better. If you can describe something accurately in one sentence rather than two, then so much the better! (But, of course, two short sentences are better than one long rambling sentence.)

Activity 13 *Organising the report*

The following are a few notes from a statistical analysis that you wish to write up as a statistical report. It includes one hypothesis test that you might not have carried out yourself because we have not discussed the details of it in this module. However, this should be no barrier to carrying out the task required in this activity. Organise the material into an outline report under the headings *Introduction*, *Methods*, *Results*, *Discussion*.

> Two groups, continuous variable. Checked that sample variances are similar. Did two-sample t-test, $p = 0.16$. Normality seemed OK in each group (probability plot). 95% two-sample t-interval $(-3.92, 17.63)$. Conclude means could be the same for both groups. Sample sizes 24 and 32.

Note that you are not expected to write the report, just to reorganise the information in the sentences or parts of sentences under the four headings.

5.2 Writing the report

In this subsection, an analysis of the data on head shape that were introduced in Activity 10 is used to illustrate what might be included in a statistical report. Only the comparison of head shapes of first and second sons will be considered. Example 14 provides a brief account of the analysis of these data.

Example 14 *Head shapes of first and second sons*

Head measurements were taken for the first and second sons from 25 families. For each son, the head shape index was calculated as head breadth divided by head length (both in the same units), multiplied by 100. It is assumed that this index does not vary substantially over the age range of the data. The question is whether there is any difference in shape indices between first and second sons.

The data are paired, so it makes sense to look at the differences between the head shape indices of first and second sons. The mean difference is 0.19, so it appears that the head shape index is slightly greater for first sons than for second sons. In Activity 10, you may have suggested that a normal model might be appropriate for the differences. A histogram and a normal probability plot of the differences are shown in Figure 38.

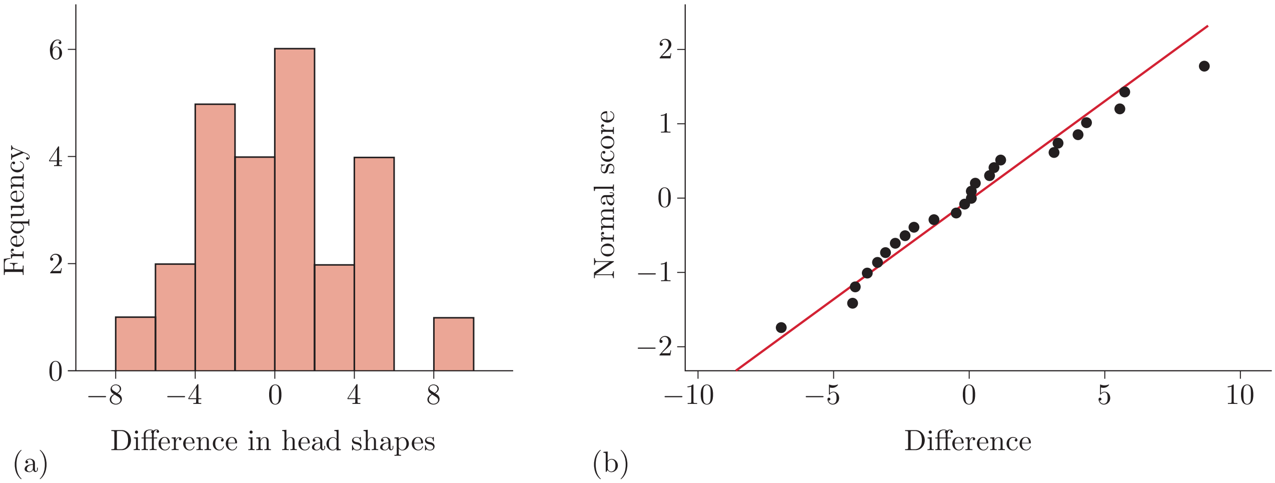


Figure 38 Head shape differences: (a) histogram; (b) normal probability plot

(Source: Frets, G.P. (1921) 'Heredity of head form in man', *Genetica*, vol. 3, no. 3, pp. 193–384)

The plots in Figure 38 both serve to confirm the reasonable validity of the normal model.

The next step is to calculate a t-interval for the mean difference: the 95% t-interval is $(-1.35, 1.72)$. Finally, a t-test of the null hypothesis of zero mean difference might be carried out. This gives $t = 0.25$ on 24 degrees of freedom; the p-value for the test is 0.803. There were no problems with outliers.

Example 15 *Writing the report*

The information in Example 14 can be organised into a statistical report as follows.

The *Introduction* states the problem and describes the data available, including the source of the data. You should not include extraneous material such as theories of heredity or genetics here (though, of course, the scientists who gave you the data might wish to do so in their report – but that is up to them). So the following (taken from the description of Activity 10 as well as Example 14) is adequate.

> *Introduction*
> In 25 families where there were at least two sons, measurements were taken on the head length and head breadth (both measured in mm) of the first and second sons. The head shape index is defined as 100 × (breadth/length). The issue addressed in this report is whether there is a difference between the head shapes of first and second sons. The data for this analysis were taken from Frets, G.P. (1921) 'Heredity of head form in man', *Genetica*, vol. 3, no. 3, pp. 193–384.

The next step is to write the *Methods* section. You should state the variables and describe the methods that you used to reach your conclusions, but not all the blind alleys that you might have explored in the process! For example, a normal model was chosen, so it should be mentioned that a probability plot was used to check the adequacy of the model. This information is required so that if other statisticians read your results, they will know that the model is valid. However, if you originally thought you might use, say, an exponential model, but dropped the idea once you looked at the data, this information should not go in the *Methods* section. Readers are not interested in your thought processes, or all the mistakes you might have made along the way, but simply want to know how you obtained your results, and whether your methods were appropriate.

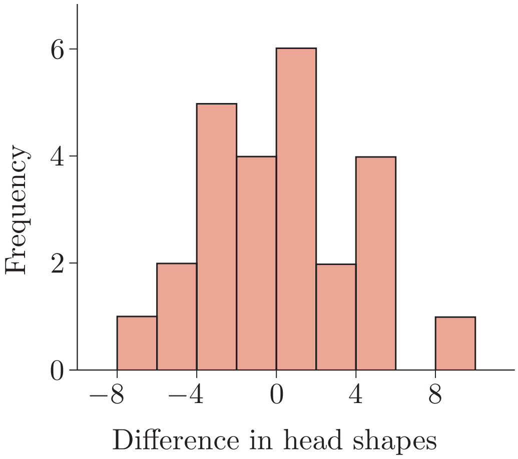
The *Methods* section is generally aimed at a statistical readership, and hence you can quote standard methods without describing them in any detail. For example, it is perfectly appropriate to say '95% t-intervals were calculated' without explaining what a confidence interval is or what the t-distribution is. Indeed, you should most definitely *not* describe what they are! Finally, it is often useful to include details of the software you used, usually the name and version. Here is a suggested *Methods* section for the report on head shapes.

> *Methods*
> As the data are paired, the analysis was based on differences between the head shape indices of first sons and second sons. A normal model was used for the differences; this assumption was checked using a normal probability plot. A 95% t-interval was calculated and a t-test was used to test the hypothesis that the mean difference is zero. All analyses were performed using Minitab Version 17.

Next comes the *Results* section. A general rule is to separate the results into descriptive summaries and analytic results. Descriptive summaries might include some relevant numerical summaries (for example, median and interquartile range) or graphs. The aim is to convey some feel for the data. However, you should beware of including too many descriptive summaries: the aim is to highlight aspects of the data that are relevant to the question of interest. For example, for the head shape data, you might include a boxplot or a histogram of the differences in head shape indices; one such graph is enough here.

The analytic results include those that directly address the original question set out in the *Introduction*. The original question relates to the difference between the head shapes of first and second sons. Thus you should report the mean difference and the 95% t-interval. In addition, you should report the result of the t-test. Finally, you need to provide some evidence that your methods are justified. In this case, a probability plot was used to test the normality assumption. It is not essential to show this plot. To save space, it is quite reasonable simply to state that you used this method to check the assumption.

> *Results*
> The distribution of the 25 differences between the head shape indices of first and second sons is shown in the histogram below. The data were approximately normally distributed, as confirmed by a probability plot (not shown).



> The mean difference in head shape indices was 0.19, with 95% t-interval $(-1.35, 1.72)$. A t-test of the hypothesis of zero mean difference gave $t = 0.25$ on 24 degrees of freedom, with p-value of 0.803.

The next section is the *Discussion* section, in which you give your interpretation of the results in the light of the original question. This is also the place where you should comment on the possible impact of any other factors (such as missing data or outliers) on the interpretation. In this example there are no such factors. The section can thus be suitably brief: there is no evidence of a difference. However, it is worth qualifying this conclusion by reminding the reader that the sample size was rather small.

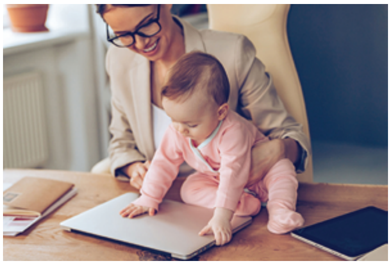
In general, it is important to write concisely and to the point.

> *Discussion*
> We conclude that there is little or no evidence against the hypothesis of no difference between the head shape indices of first and second sons. However, the sample size for this study was quite small, being only 25.

Finally, having assembled and re-read the report, you can now write the *Summary*. This states briefly the purpose of the analysis, the method used, the key finding and its interpretation. It should be largely non-technical.

> *Summary*
> The aim of this analysis was to compare head shapes of first and second sons, using a head shape index based on the ratio of head breadth to head length. Data on 25 pairs of first and second sons were obtained from a published source and analysed using a normal model. We found no significant difference between the head shapes of first and second sons.

This completes the report. The final step is to read through the report and check it.



Yes, mum, you've finished your report . . .

The sections of the report on head shapes of first and second sons that were written in Example 15 are assembled in the following box.

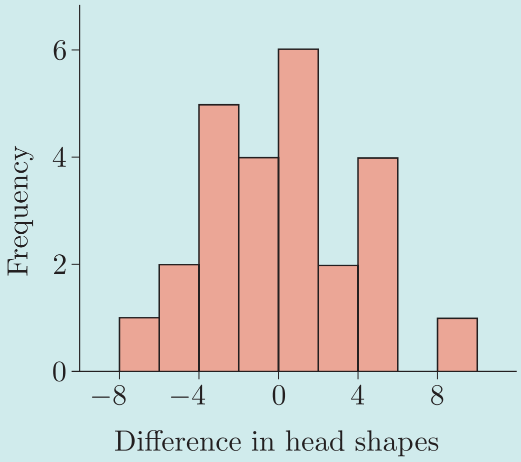
A complete statistical report

Summary
The aim of this analysis was to compare head shapes of first and second sons, using a head shape index based on the ratio of head breadth to head length. Data on 25 pairs of first and second sons were obtained from a published source and analysed using a normal model. We found no significant difference between the head shapes of first and second sons.

Introduction
In 25 families where there were at least two sons, measurements were taken on the head length and head breadth (both measured in mm) of the first and second sons. The head shape index is defined as $100 \times (\text{breadth}/\text{length})$. The issue addressed in this report is whether there is a difference between the head shapes of first and second sons. The data for this analysis were taken from Frets, G.P. (1921) 'Heredity of head form in man', *Genetica*, vol. 3, no. 3, pp. 193–384.

Methods
As the data are paired, the analysis was based on differences between the head shape indices of first sons and second sons. A normal model was used for the differences; this assumption was checked using a normal probability plot. A 95% t-interval was calculated and a t-test was used to test the hypothesis that the mean difference is zero. All analyses were performed using Minitab Version 17.

Results
The distribution of the 25 differences between the head shape indices of first and second sons is shown in the histogram below. The data were approximately normally distributed, as confirmed by a probability plot (not shown).



The mean difference in head shape indices was 0.19, with 95% t-interval $(-1.35, 1.72)$. A t-test of the hypothesis of zero mean difference gave $t = 0.25$ on 24 degrees of freedom, with p-value of 0.803.

Discussion
We conclude that there is little or no evidence against the hypothesis of no difference between the head shape indices of first and second sons. However, the sample size for this study was quite small, being only 25.

Activities 14 and 15 will give you some practice at writing short statistical reports.

Activity 14 *Used windows on reusable envelopes*

This activity is based on the data on used windows on reusable envelopes, which were discussed in Example 11 and analysed in Activity 9. (Source: Sutherland, W. (1990) 'The great pigeonhole in the sky', *New Scientist*, 9 June, vol. 1720, pp. 73–4.) The aims of a slightly extended version of the analysis in Activity 9 were as follows:

- to estimate the mean number of used windows
- to find a suitable model for the distribution of used windows.

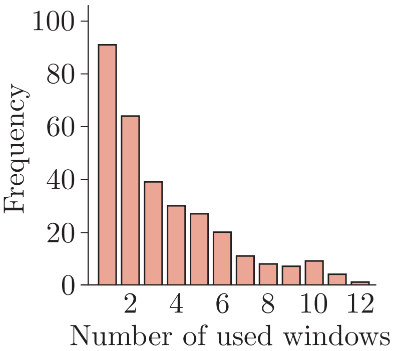


Figure 39 Numbers of used windows

A bar chart of the numbers of used windows is shown in Figure 39. (This is a repeat of Figure 34.) The mean number of used windows for this sample of 311 envelopes was 3.412, with large-sample 95% confidence interval for the mean $(3.122, 3.701)$.

In Activity 9, you may have concluded that the shape of the distribution was suggestive of a geometric distribution, even though the maximum possible number of used windows is 12. The goodness-of-fit of the geometric model was assessed using a chi-squared goodness-of-fit test; after suitable combination of cells with low expected frequencies, a chi-squared test statistic of 13.646 on 9 degrees of freedom was obtained, $p = 0.135$.

Write a short report of this analysis.

Activity 15 *The effect of caffeine on finger-tapping*

This activity is based on the data on the effect of the stimulant caffeine on alertness, as measured by speed of finger-tapping of student subjects, described and analysed in Activities 12, 16 and 18, and Example 15, of Unit 11. (Source: Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn, New York, John Wiley and Sons, p. 425.) To remind you, 30 male college students were trained in finger-tapping; they were then randomly divided into three groups of ten, and the students in each group received different doses of caffeine (0 mg, 100 mg and 200 mg). Two hours after treatment, each student was required to do finger-tapping, and the number of taps achieved per minute was recorded. For the purposes of this activity, the aims of the analysis of this dataset were as follows:

- to understand the relationship between caffeine dose and number of taps per minute by using a suitable model
- to predict the number of taps per minute that might be expected if a student had received a dose of 40 mg of caffeine.

A scatterplot of the number of taps per minute against caffeine dose is shown in Figure 40. (This is a repeat of Figure 19 of Unit 11.)

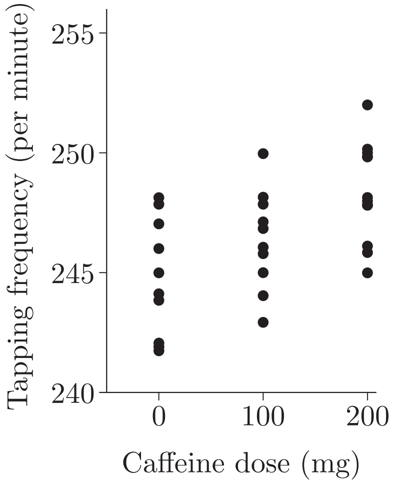


Figure 40 Tapping performance against caffeine dose

A regression line was fitted to the data using least squares; this line has the formula

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose}.$$

Examination of a residual plot and a normal probability plot of the residuals showed no evidence that the simple linear regression model was not suitable for these data. A hypothesis test of whether there is a regression relationship resulted in a p-value of 0.0013.

For a student taking a caffeine dose of 40 mg, the predicted number of taps per minute, using this regression model, is 245.45 with 95% prediction interval $(240.85, 250.05)$.

Write a short report of this analysis.

Exercise on Section 4

Exercise 4 *Fish traps*

In this activity, you are invited to write a report on a further analysis that was carried out of the fish traps data that were considered in Exercises 1 and 2 of Unit 8. (Source: David, F.N. (1971) *A First Course in Statistics*, 2nd edn, London, Griffin.) These data arose from an experiment in which $n = 100$ fish traps were set and the number of fish caught in each trap were counted. The aims of this further analysis were as follows:

- to describe the distribution of catches
- to estimate the mean number of fish caught per trap
- to estimate the proportion of traps with no catch.

Out of the 100 fish traps, 72 of the traps contained between 3 and 5 fish, and the maximum number of fish caught in any trap was 8. The distribution of catches is shown in the bar chart in Figure 41.

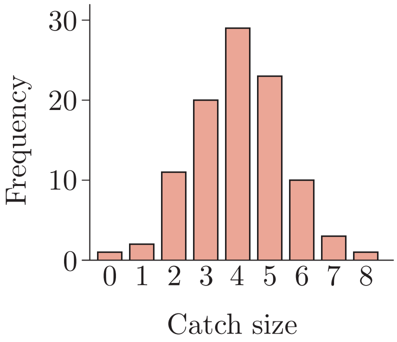


Figure 41 Fish trapped

The mean number of fish per trap was 4.04, with 95% z-interval $(3.76, 4.32)$. One trap failed to catch any fish. The proportion of traps with no catch was thus 0.01. The exact 95% confidence interval for this proportion – obtained using a method not covered in this module – is $(0.00025, 0.054)$.

Write a short report of this analysis.

Summary

This unit is in two parts.

The first part consists of an investigation of the role of transformations of the data in statistics.

In the one-sample case, transformations to normality of continuous variables have been considered and, for use with positive data, the ladder of powers was introduced. In the regression situation, you have seen that sometimes a general regression model can be simplified to a linear regression model by an appropriate transformation of the data: a transformation may be applied to the explanatory variable to linearise the relationship, or a transformation may be applied to the response variable to improve the behaviour of the random terms in the model.

In the second part of this unit, the methods that you have learned so far in the module have been integrated into a statistical modelling process.

Some basic principles for thinking about data and models, even before looking at the data, have been reviewed. Key issues include:

- whether the data are discrete or continuous
- whether the setting in which the data were collected conforms to any of the standard settings
- what is the likely range and shape of the distribution.

These basic principles can help to formulate a starting point for choosing a model, which can be revised in the light of the data. The handling of outliers has been discussed briefly.

You have undertaken extended analysis of a dataset using Minitab, starting from a scientific question, progressing through the various stages of exploratory analysis, model and method choice, model checking, and performing the relevant statistical calculations.

Finally, you have learned how to structure and write a statistical report. A convenient structure includes paragraphs entitled *Summary*, *Introduction*, *Methods*, *Results* and *Discussion*.

Learning outcomes

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

- choose a transformation of a single continuous variable to make its distribution more normal, if necessary
- use the ladder of powers to help choose a transformation in the case of positive random variables
- use a transformation of the explanatory variable to straighten out a non-linear relationship between the variables in a general regression model, so that a linear regression model can be fitted to the transformed data
- use a transformation of the response variable to improve the behaviour of the random terms in the general regression model, so that a linear regression model can be fitted to the transformed data
- fit quadratic and cubic functions of a single explanatory variable using multiple regression
- appreciate that statistical analysis is a process, beginning with a question or problem of interest, ending with a statistical report, and involving data exploration, model choice and model checking, in a cycle that may be repeated several times
- appreciate that the aim of statistical modelling is to draw valid and relevant inferences, not to find a perfect model
- use information about the setting of a problem and the type of data collected to set out an initial modelling framework
- choose appropriate statistical techniques to address a specific problem or question
- identify outliers and explore their influence
- combine the data manipulation, calculation, statistical and graphical facilities of Minitab to undertake a complete statistical analysis
- structure and write a statistical report; such a report comprises a non-technical summary, an introduction, a methods section, a results section, and a discussion.

Solutions to activities

Solution to Activity 1

A continuous uniform distribution is not a good model for these data because the p.d.f. is clearly not constant over whatever suitable range for the distribution of the whole dataset might have been chosen. In addition, the range of a continuous uniform distribution is finite; here, the lower limit, a, of the range could be 0 but what could one justifiably choose for the upper limit, b, the highest interspike time that could possibly occur? Also, the data being clearly right-skew indicates that a normal model is not appropriate either (even if the probability of a negative interspike interval were sufficiently small under a normal model).

Solution to Activity 2

The only available transformation of the four is $y = (2+x)^2$.

The transformation $y = \sqrt{x}$ is not available because it is not defined for the negative values in the range of x.

The transformation $y = x^4$ is defined over the required range but, like x^2 in Example 4, is neither increasing nor decreasing over the entire range. This is illustrated in Figure 42, where the graph of $y = x^4$ is shown for $-1 < x < 1$: $y = x^4$ is decreasing between -1 and 0, then increasing between 0 and 1. Mathematically, if $h(x) = x^4$, then $h'(x) = 4x^3$, so $h'(x) < 0$ for $-1 < x < 0$ and $h'(x) > 0$ for $0 < x < 1$.

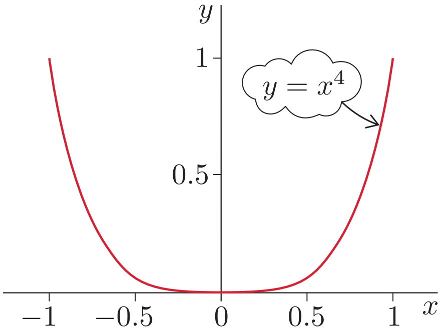


Figure 42 The graph of $y = x^4$ for x between -1 and 1

The transformation $y = (2+x)^2$ is both defined and increasing over the required range. To see the latter, observe that if $h(x) = (2+x)^2$, then, using the chain rule (Subsection 3.1 of Unit 7), $h'(x) = 2(2+x)$. This is a linear function taking the value 2 when $x = -1$ and 6 when $x = 1$, and so is positive for all $-1 < x < 1$. The transformation is shown in Figure 43.

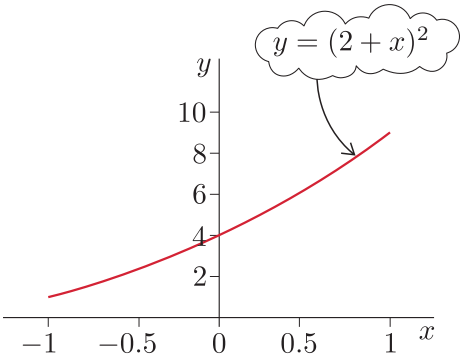


Figure 43 The graph of $y = (2 + x)^2$ for x between −1 and 1

The transformation $y = -\log x$ is not available because log, and hence − log, is not defined for the negative (or zero) values in the range of x.

Solution to Activity 3

(a) The data are, in this case, left-skew. To tackle left-skew, it is necessary to go up the ladder of powers, and it would be worth trying a transformation such as x^2 or x^3. (As it happens, the simulated nature of these data mean that there is a correct answer, that the transformation x^3 would lead to a normally distributed dataset. However, you cannot ascertain this without further exploration that you were not asked to take on.)

(b) The histogram already looks quite symmetric, so no transformation is required. (Because the data are simulated, this is known to be the correct answer, even if you *might* have doubted the normality assumption from the histogram!)

(c) The data are, in this case, right-skew. To tackle right-skew, it is necessary to go down the ladder of powers, and it would be worth trying a transformation such as $z^{\frac{1}{2}}$ or $\log z$. (In fact, in this case, the log transformation is the one that would lead to a normally distributed dataset.)

Solution to Activity 4

(a) Transformations from the ladder of powers with powers less than 1 (including log) would be expected to be the most appropriate when attempting to transform the interspike interval data to normality. This is because the data are right-skew.

(b) The original data are markedly right-skew. Both the log and the square root transformations have reduced the skew by ‘pulling in’ the values to the right of the mode and ‘stretching out’ those to the left. This ‘stretching out’ effect is more marked for the log transformation as can be seen from the fact that low values deviate from the line in the normal probability plot. However, it is rather difficult to decide which of the two transformed datasets is better fitted by the normal distribution. One *might* argue that the log transformation has overcompensated for the right-skew, introducing some points on the normal probability plot that are out of line with the others but ‘in the opposite direction’ at either extreme. Arguably, the points on the normal probability plot of the square-root-transformed data better follow a straight line (but a slight bend can still be perceived in the line of points). Has this transformation not quite done enough?

Solution to Activity 5

The first and third regression functions can be linearised by employing the transformations $x' = x^3$ and $x' = \log(x/(1-x))$, respectively. The other two regression functions cannot be linearised in this way because the potential transformations depend on further unknown parameters (λ in the case of the second regression function, both μ and γ in the case of the fourth regression function).

Solution to Activity 6

The prediction of the *square root* of the number of matings of an elephant of age 40 years is

$$\sqrt{y} = -0.812 + 0.0632 \times 40 = 1.716.$$

Therefore the predicted number of matings is

$$y = 1.716^2 = 2.945 \simeq 3.$$

Solution to Activity 7

(a) The regression function in the fitted model is not quite right. This is because the points in the residual plot display some structure. Inspection of the figure reveals a distinctly curved shape.

(b) For the cubic model, the points in the residual plot can be argued to show no clear pattern, so the random terms plausibly have a constant, zero mean and constant variance. Also, the residuals in the normal probability plot lie roughly along a straight line, so the assumption of normality of random terms seems plausible. The model assumptions seem adequate in this case. (As so often, there is a caveat. It could alternatively be argued that there is more spread in the residual values at low fitted values than at high, but there are not really enough data points to be sure.)

(c) In order to make a prediction, it is useful to rewrite the fitted regression model

$$y = 5.65 + 3.58\,x_1 + 0.654\,x_2 - 0.0552\,x_3$$

in terms of the hardwood content x itself, as

$$y = 5.65 + 3.58\,x + 0.654\,x^2 - 0.0552\,x^3.$$

For paper with $x = 10$, the predicted value of tensile strength is

$$\begin{aligned} y &= 5.65 + 3.58 \times 10 + 0.654 \times 100 - 0.0552 \times 1000 \\ &= 51.65 \text{ p.s.i.} \end{aligned}$$

Solution to Activity 8

Five of the distributions in Table 1 are discrete and five are continuous.

The discrete distributions are

Bernoulli, binomial, discrete uniform, geometric, Poisson.

The continuous distributions are

chi-squared, continuous uniform, exponential, normal, t.

Solution to Activity 9

(a) Start with the binomial model. To fit this in with its standard setting, we need to identify a relevant Bernoulli trial. The obvious choice is whether or not a window has been filled. However, as mentioned in Example 11, the probability that a window is filled depends on the position of the window. So, say, if we were to think of each window going down the envelope as a Bernoulli trial, the probability, p, of a 'success' (the window being filled) would change from window to window, and hence not be constant. Thus the standard setting for the binomial model does not seem appropriate. Also, the range of the binomial distribution includes zero, which is not possible in these data. Unless p is such that $P(X = 0)$ is very small under the binomial model, it seems that the binomial distribution is not likely to be a suitable model for these data. (Here, X is a random variable representing the number of used windows.)

The discrete uniform distribution can have the finite range $1, 2, \ldots, 12$. However, it is not clear at all from the setting that each outcome has the same probability; indeed, one would probably expect windows higher up the envelope to have higher probabilities of being filled than windows towards the bottom. However, the jury is out until we see some data!

Our inability to define what would constitute a constant-probability Bernoulli trial in this context rules out the standard setting for the geometric distribution. The range of the geometric distribution starts at 1, which is appropriate to these data, but continues beyond 12 ('to infinity'!), which is not. However, the range constraint might not be a problem if p is such that, under this model, $P(X > 12)$ is sufficiently small.

The standard setting for the Poisson model requires an 'event' that occurs at random, and some fixed interval of time within which such events occur. This doesn't really seem to fit the envelope situation at all. The range of the Poisson distribution includes both 0 and values beyond 12, so it has the difficulties in this respect of the binomial and geometric distributions combined! (However, a way forward if no simpler model proves useful might be to try modelling $X - 1$ using a Poisson distribution.)

From these considerations, there is no compelling reason to opt for any of the standard models, though some – the discrete uniform and geometric – seem more likely to be appropriate than others.

(b) It is clear that envelopes with few used windows are more frequent than envelopes with many used windows. So the uniform distribution is not appropriate.

The shape of the data appears to be consistent in general terms with either a Poisson model or a geometric model. Because its range starts at 1, a geometric model seems to be the more suitable.

(c) The p-value of 0.135, being greater than 0.1, means that there is little or no evidence against the null hypothesis that the geometric distribution provides a good model for these data.

Solution to Activity 10

All three variables are continuous, so a continuous model should be chosen in each case.

(a) The continuous uniform distribution is a *possible* model, but it requires all head sizes to be equally likely over some range of values of head size (and impossible otherwise), which doesn't seem especially plausible for this measurement. Head size is necessarily positive. So the exponential distribution is a possible model for these data. However, the shape of the exponential model does not seem appropriate, with its high probability of head sizes close to zero! Indeed, values of head length plus head breadth are likely to cluster around some typical value some distance from zero, so that a normal model is more likely to be appropriate. Note, however, that the normal model is not ideal since it theoretically allows negative values. If the data on head size are not very normally distributed – for example, they are skew – then a transformation to normality might be considered.

(b) Much the same considerations as in part (a) for head size apply to data on head shape also.

(c) Differences between head shapes can reasonably be expected to take both negative and positive values, perhaps clustered close to zero. A normal model again seems appropriate here, as a first choice; an exponential model is certainly not now appropriate, because of the negative values. A transformation to normality remains a possibility, the transformation being different from that used in part (a) because of the different ranges of the distributions of the data involved.

Solution to Activity 11

(a) The shape of the histogram suggests that the exponential model might, on the face of it, be a reasonable model for these data. (By the way, the normal model seems out of the question, owing to the substantial skewness of the histogram.)

(b) The sample mean and sample standard deviation do not differ greatly. This suggests that an exponential model might be a reasonable choice.

Solution to Activity 12

An estimate of the age of the site based on all eight observations is given by the sample mean, which is approximately 2621.8 years. However, this mean is rather unsatisfactory, as it lies above seven of the eight data points. This is because it is greatly influenced by the value for sample number C-367, which is 3433 years. When this point is omitted, the mean of the remaining seven points is approximately 2505.9 years. Sample number C-367 clearly has a big influence on the mean. You might therefore report the calculations both including and excluding sample C-367, and perhaps suggest further investigation of the outlier. Additionally, you might note that the median of all eight observations is 2530.5 years, which is broadly in line with the sample mean of the data without the outlier. (And the resistant nature of the sample median is reflected in it taking the similar value of 2521 years when the outlier is omitted.)

Solution to Activity 13

Reorganising the material should produce something like the following.

Introduction
Compare means of a continuous variable in two groups given samples of sizes 24 and 32.

Methods
Check normality in each group using probability plots.
Check assumption of equality of variances.
Calculate 95% two-sample t-interval.
Perform two-sample t-test.

Results
Normal model reasonable.
95% t-interval was $(-3.92, 17.63)$.
Two-sample t-test gave $p = 0.16$.

Discussion
Little or no evidence that the means are different in the two groups.

Solution to Activity 14

Here is a possible report.

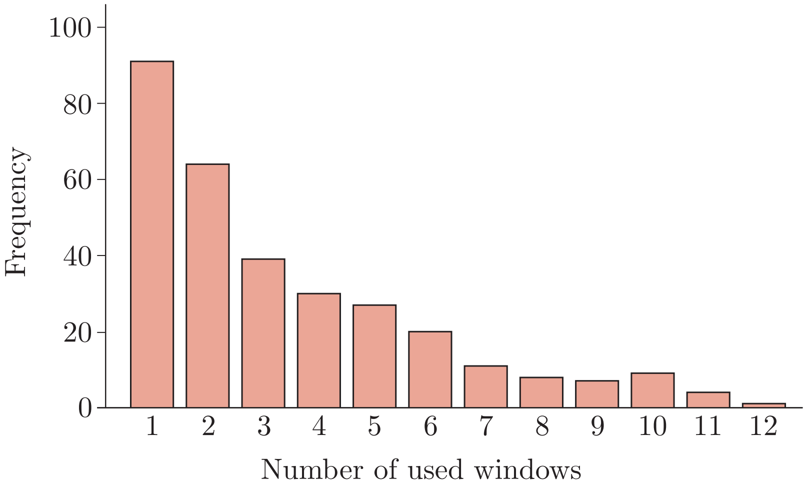
Summary
The main aim of this analysis is to develop a model for the distribution of the number of windows used on a sample of reusable envelopes. Data on a sample of 311 envelopes were obtained from a published source. A geometric model was found to provide a good fit to the data.

Introduction
The numbers of used windows on a sample of 311 reusable envelopes used for internal circulation of documents within an office were counted. The aim of the analysis is to describe the distribution of the number of used windows, estimate the mean number of used windows, and obtain a model for the distribution of used windows. The data for this analysis were

obtained from Sutherland, W. (1990) 'The great pigeonhole in the sky', *New Scientist*, 9 June, vol. 1720, pp. 73–4.

Methods
A 95% z-interval for the mean number of used windows was calculated. The fit of the geometric model was assessed using the chi-squared goodness-of-fit test. Most calculations were performed using Minitab Version 17.

Results
The numbers of used windows on the 311 envelopes were distributed as shown in the bar chart below. The mean number of used windows was approximately 3.41, with approximate 95% confidence interval $(3.12, 3.70)$.



A geometric model with $p = 1/3.412 \simeq 0.29$ provided a good fit to the data. The chi-squared test statistic was 13.646 on 9 degrees of freedom, $p = 0.135$.

Discussion
The average number of used windows on the envelopes is about 3.41. A geometric distribution provided a very good fit to the data. However, an approximate aspect of the model is that it does not allow for the fact that the number of windows on each envelope is restricted to 12.

Solution to Activity 15

Here is a possible report.

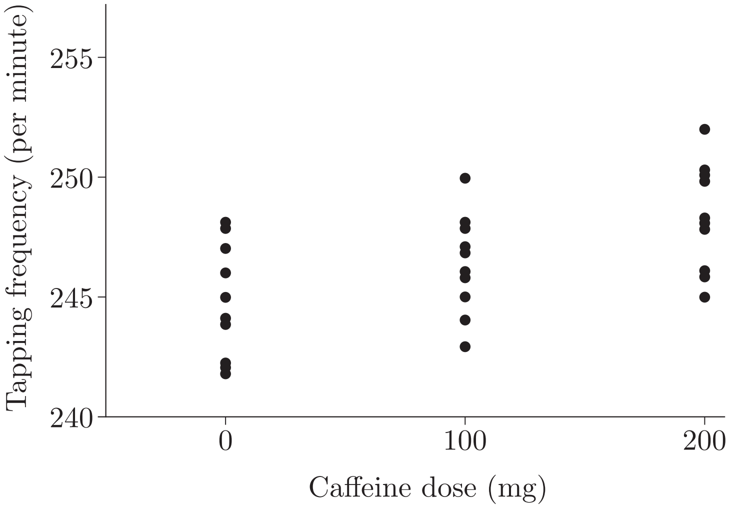
Summary
The main aim of this analysis is to develop a model for the effect of the stimulant caffeine on the alertness of subjects, as measured by their rate of finger-tapping. Data from an experiment on a sample of male subjects trained in finger-tapping were obtained from a published source. A linear regression model was found to provide a good fit to the data, allowing for interpretation and prediction of the effect of caffeine on finger-tapping.

Introduction
The tapping rates per minute of a trained sample of 30 male subjects,

each given one of three doses of caffeine (0 mg, 100 mg and 200 mg) were recorded. The aim of the analysis is to develop a model for the effect of caffeine on finger-tapping, and to use the model to interpret and predict this effect. The data for this analysis were obtained from Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn, New York, John Wiley and Sons, p. 425.

Methods
A linear regression model was fitted to the data using least squares. The fit of the regression model was assessed using a residual plot and a normal probability plot of the residuals. A hypothesis test of whether there is any regression relationship was performed. Point and interval predictions of finger-tapping rates for someone taking 40 mg of caffeine were obtained. Most calculations were performed using Minitab Version 17.

Results
A scatterplot of the number of taps per minute against caffeine dose is shown below.



The regression line fitted to these data is

$$\text{taps} = 244.75 + 0.0175 \times \text{caffeine dose}.$$

Examination of a residual plot and a normal probability plot of the residuals showed no evidence that the simple linear regression model was not suitable for these data. A hypothesis test of whether there is a regression relationship resulted in a p-value of 0.0013; there is strong evidence of the existence of a relationship.

The model suggests that, on average, the number of taps per minute increases by 0.0175 for each extra mg of caffeine taken.

For a student taking a caffeine dose of 40 mg, the predicted number of taps per minute, using this model, is about 245.45 with 95% prediction interval $(240.85, 250.05)$.

Discussion
There is strong evidence of a linear regression relationship between

caffeine dose and number of taps per minute, a small increasing effect being observed. This relationship has been used for prediction. It is unclear how far a linear relationship would continue to hold for caffeine doses greater than those (up to 200 mg) used in the experiment.

Solutions to exercises

Solution to Exercise 1

Any member of the ladder of powers other than log can be written as $h(x) = x^p$ where p takes values $\ldots, -2, -1, -1/2, 1/2, 1, 2, 3, 4, \ldots$. The derivative of $h(x)$ is

$$h'(x) = px^{p-1}.$$

For $x > 0$, $x^{p-1} > 0$ for all values of p on the ladder. It follows that $px^{p-1} > 0$ for all ladder values of $p > 0$ and $px^{p-1} < 0$ for all ladder values of $p < 0$. It then follows that $h(x) = x^p$ is an increasing function of positive x for all ladder values of $p > 0$ (including $p = 1/2, 1, 2, 3, 4$) and that $h(x) = x^p$ is a decreasing function of positive x for all ladder values of $p < 0$ (including $p = -2, -1, -1/2$).

Solution to Exercise 2

(a) While a normal model is not an indefensible one for these data, it does seem from the histogram that there is a certain amount of left-skew.

(b) Transformations from the ladder of powers with powers greater than 1 would be expected to be the most appropriate when attempting to transform the glass fibre strength data to normality. This is because the data are left-skew.

(c) The left-skew in the data is visible in Figure 16(a) as a bend in the line of points which ‘faces the other way’ in comparison with the bend due to right-skew in the normal probability plot of Figure 13, for example. All three transformations have reduced the left-skew by ‘pulling in’ the values to the left of the mode and ‘stretching out’ those to the right. It appears that the square transformation (Figure 16(b)) may not have done enough to remove the bend in Figure 16(a), while the fourth power transformation (Figure 16(d)) seems to have transformed the data ‘too far’: while the central dots in the latter plot follow a reasonable straight line, there are now outlying points at both extremes. Of the candidates offered, the cube transformation (Figure 16(c)) has, arguably, set the best balance in achieving something approaching normality. That said, one might still reject normality as a suitable model even for the cubes of the data.

Solution to Exercise 3

Figure 32(a) is a residual plot which is of the type that you might expect to obtain when the assumptions are justified. There is therefore no need to employ any transformations.

Figure 32(b) displays a strong pattern indicating a systematic discrepancy from the assumed mean of the model. The relationship between the response and explanatory variables appears to be non-linear. Transforming the explanatory variable might therefore provide a remedy. (Indeed, the quadratic/cubic nature of the main trend in the residual plot might indicate the need for a treatment like that of the kraft paper in Subsection 2.3 using multiple regression, involving certain transformations of the explanatory variable.)

In Figure 32(c), the pattern is indicative of a variance that is not constant, but increasing as the fitted value increases. Transformation of the response variable is an approach open to you to try to accommodate such behaviour.

Finally, the residual plot in Figure 32(d) is well-behaved except for a single outlier. Transformations are not, typically, the answer here (but see Subsection 3.3 for a little more on dealing with outliers).

Solution to Exercise 4

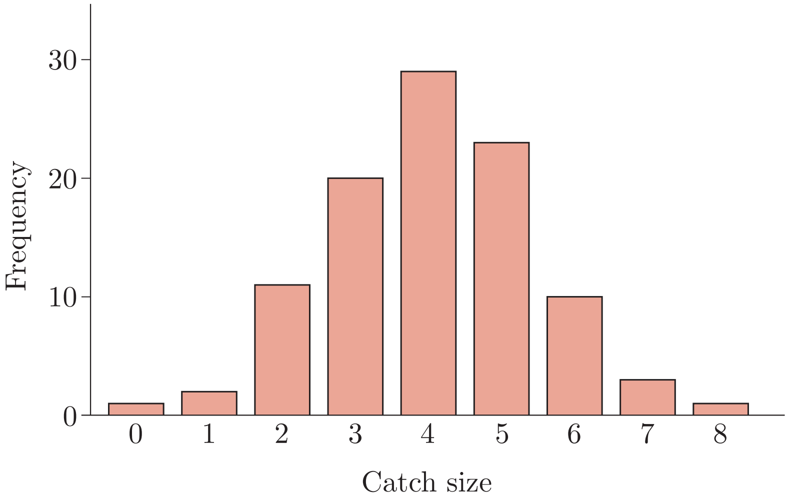
Here is a possible report.

Summary
The observed distribution of catch sizes from 100 fish traps is described and some aspects of the population distribution of catch sizes are estimated using data obtained from a published source.

Introduction
The numbers of fish caught in 100 traps were counted. The aims of this analysis are to describe the observed distribution of fish catches, and to estimate the population mean catch per trap and the population proportion of traps with zero catch. The data for this analysis were obtained from David, F.N. (1971) *A First Course in Statistics*, 2nd edn, London, Griffin.

Methods
The observed data were graphed. An approximate 95% confidence interval for the mean catch per trap was calculated using large-sample methods. A 95% confidence interval for the proportion of traps with zero catch was calculated using exact methods. All calculations were performed using Minitab Version 17.

Results
A bar chart of the numbers of catches is shown below.



Of the 100 traps, 72 contained between 3 and 5 fish. The minimum catch was 0, the maximum was 8. The average catch per trap was 4.04 fish, with approximate 95% confidence interval for the population mean $(3.76, 4.32)$. Out of the 100 traps, only one produced a zero catch. The proportion of traps with no catch is thus 0.01, with exact 95% confidence interval for the population proportion $(0.00025, 0.054)$.

Discussion

The distribution of fish catches was approximately symmetric with mean about 4. Only one out of one hundred traps had zero catch.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 81: © Katarzyna Bialasiewicz / Dreamstime.com

Page 82: Courtesy of Richard Mooney, Duke University

Page 85: © 2017 MSP Communications, Inc

Page 96: © LUZI et al. (2004) www.sswm.info This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 97: © Fernando Gregory Milan / www.123rf.com

Page 99: © NSIL/Dick Roberts, Visuals Unlimited / Science Photo Library

Page 101: © Sicnag This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 103: © Duncan Noakes / www.123rf.com

Page 106: © Mikhail Olykaynen / www.123rf.com

Page 111 top: © hxdbzxy / www.123rf.com

Page 111 bottom: © John Sommer / iStock / Getty

Page 112: © The Keasbury-Gordon Photograph Archive / Alamy Stock Photo

Page 115: Taken From: https://www.amazon.co.uk/New-Guardian-Intertac-Envelopes-Resealable/dp/B000I2BZI6

Page 118: © Guy Corbishley / Alamy Stock Photo

Page 119: © Roger This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 122: © Courtesy Lamoka Waneta Lakes Association

Page 124: © alkanc / www.123rf.com

Page 127: © Andrei Zaripov / www.123rf.com

Page 129: © Andrei Zaripov / www.123rf.com

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.