

7 Basic Normal theory and the Central Limit Theorem

The module now changes direction. So far, we have mostly looked at statistical methods which are ‘data-driven’, i.e. tend to use data to answer questions in very direct ways. The remainder of the module will deal mostly with methods which are much more based on modelling data and with analysing these models in order to answer questions.

This section introduces some fundamental distributional results, including the Central Limit Theorem (the result which justifies the prominence given to the normal distribution in statistical theory).

7.1 Basic properties of normal distributions

7.1.1 Linear transformation of a normal r.v.

Let X be a random variable with mean μ and variance σ^2 . If Y is defined as $Y = a + bX$ then

$$E(Y) = \quad \text{and} \quad \text{var}(Y) = \quad .$$

Further, if X is normally distributed then so is Y . Thus

$$X \sim N(\mu, \sigma^2) \Rightarrow a + bX \sim N(a + b\mu, b^2\sigma^2). \quad (7.1)$$

Prove this important result. (Hint: Express $P(a + bX < c)$ as an integral.)

Now suppose that $X \sim N(\mu, \sigma^2)$, and consider the linear transformation

$$Z = \frac{X - \mu}{\sigma}.$$

What is the distribution of Z ?

This result is useful for working out probabilities associated with any normal distribution (if \mathbf{R} is not available) — just transform to the standard normal distribution and use the published tables for probabilities associated with $N(0, 1)$. As an example, calculate the probability that a random variable $X \sim N(3, 4)$ takes a value between 4 and 5. You may like to use Table 5 of the K & Y Tables or the following table.

| | | | | | | |
|------------|---------|---------|---------|---------|---------|---------|
| z | 0.0 | 0.5 | 1.0 | 2.0 | 2.5 | 3.0 |
| $P(Z > z)$ | 0.50000 | 0.30854 | 0.15866 | 0.02275 | 0.00621 | 0.00135 |

7.1.2 Sums of independent normal r.v's

Sums of normal random variables occur quite frequently in statistical theory. If these r.v's are *independent* then their sum has exactly the distribution that you would expect. It will be shown in Honours that the sum of two *independent* normal r.v's is also normal. Thus

$$X_1, X_2 \text{ independent with } X_i \sim N(\mu_i, \sigma_i^2) \quad i = 1, 2 \Rightarrow X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \quad (7.2)$$

A useful extension of (7.2) is that if X_1, \dots, X_n are independent r.v's with

$$X_i \sim N(\mu_i, \sigma_i^2) \quad i = 1, \dots, n,$$

and a_1, a_2, \dots, a_n are constants then

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right). \quad (7.3)$$

Use (7.3) to obtain the distribution of the mean \bar{X} of n independent and identically distributed random variables X_1, \dots, X_n from $N(\mu, \sigma^2)$. (This is a particularly important result, so make sure that the answer is clearly laid out.)

These results will be used later, when we look at modelling data as observations of normally distributed r.v's. Before that, we consider what is so special about normal distributions.

7.2 The Central Limit Theorem (CLT)

In the previous sub-section, we noted that the mean of n independent identically normally distributed random variables is itself normally distributed. More surprisingly, the central limit theorem states that the mean of n i.i.d. r.v's from almost *any distribution* is approximately normally distributed for large enough n . This theorem is a major reason for paying so much attention to the theory of normally distributed random variables.

Central Limit Theorem: Let X_1, \dots, X_n be independent identically distributed random variables from any distribution having mean μ and variance σ^2 . Then

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \rightsquigarrow N(0, 1) \quad n \rightarrow \infty, \quad (7.4)$$

where ' \rightsquigarrow ' means 'is distributed approximately as'.

Statement (7.4) is equivalent to

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1) \quad n \rightarrow \infty. \quad (7.5)$$

(The notation ' \rightsquigarrow ' is not quite standard.)

There are more general versions of the Central Limit Theorem which do not even require the assumption that the random variables concerned are identically distributed. (An example can be found in De Groot.)

Example: A bridge can hold at most 400 vehicles if they are bumper-to-bumper and stationary. The mean weight of vehicles using the bridge is 2.5 tonnes with a standard deviation of 2.0 tonnes. What is the probability that the maximum design load of 1100 tonnes will be exceeded in a traffic jam?

7.3 Approximating other distributions by normal distributions

The Central Limit Theorem provides the justification for approximating several other distributions by a normal distribution in certain circumstances. In this section, two examples are considered.

The **binomial** probability function

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

becomes hard to evaluate for large n (because the factorials in the binomial coefficient ‘explode’). However, if $X \sim \text{bin}(n, p)$ then X can be written as a sum of n independent binomial random variables: $X = X_1 + \dots + X_n$, where $X_i \sim \text{bin}(1, p)$. (These are also called *Bernoulli* r.v.’s.) Each X_i has mean p and variance $p(1-p)$. Thus the CLT implies that

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\sim}{\sim} N(0, 1) \quad n \rightarrow \infty, \quad (7.6)$$

which you may prefer to remember as

$$X \overset{\sim}{\sim} N(np, np(1-p)) \quad n \rightarrow \infty.$$

Note that the CLT does not tell us how large n should be for this approximation to hold. The usual rough guide is that the approximation is good enough for most purposes when $\max(np, n(1-p)) > 5$.

The **Poisson** probability function

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \quad x = 0, 1, \dots$$

is also awkward to evaluate for high values of μ . However, we can exploit the useful general result (which will be proved in Homours) that

$$X_1, X_2 \text{ independent with } X_i \sim \mathcal{P}(\mu_i) \quad i = 1, 2 \Rightarrow X_1 + X_2 \sim \mathcal{P}(\mu_1 + \mu_2). \quad (7.7)$$

Using (7.7), we can see that if $X \sim \mathcal{P}(\mu)$ then $X = X_1 + \dots + X_n$, where X_1, \dots, X_n are independent and $X_i \sim \mathcal{P}(\mu/n)$. Since the variance of a Poisson distribution is equal to its mean (Exercise!), each X_i has mean and variance μ/n . Then the CLT gives

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

so that

$$X \approx N(\mu, \mu). \quad (7.8)$$

This approximation is reasonable for large μ .

The above approximations to binomial and Poisson may seem a bit odd, in that discrete p.f.'s are approximated by continuous p.d.f.'s, and the probability of getting any particular value from a continuous distribution is zero! However, a more careful use of the above approximations employs the following *continuity correction*. We approximate a discrete random variable X (here binomial or Poisson) by a continuous (normal) random variable X^* and we use the approximation

$$P(X = x) \simeq P(x - 0.5 < X^* < x + 0.5) \quad \text{for integer } x.$$

Find (an approximation to) the probability that a Poisson distributed r.v. with mean 25 takes a value in the range 26 to 30 (inclusive). [You may like to know that if $Z \sim N(0, 1)$ then $P(Z > 0.1) = 0.46017$ and $P(Z > 1.1) = 0.13567$.]

8 Practical Applications of Normal Distributions

The practical reasons for paying special attention to the normal distributions are:

- (i) The Central Limit Theorem shows that sums of independent random variables tend towards normality even if the distributions of the r.v.'s themselves are non-normal.
- (ii) Applicability: There are many data sets for which a normal distribution seems to provide a good model: e.g. errors in measurements in the physical sciences, heights of people, IQ scores. Often this may be because the things being measured are made up of many small additive random effects, so that the measurement itself could be viewed as a sum of a large number of independent random variables.
- (iii) Convenience: Normal distributions are often easy distributions with which to work mathematically (even though there is no expression in 'closed form' for the c.d.f.).
- (v) Robustness: Procedures based on the assumption of normality are often insensitive to small violations of the assumption.
- (vi) Transformation: Data which are not from normal distributions can often be transformed to approximate normality.

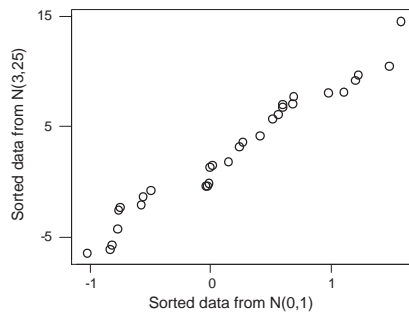
In this section, normal distributions will be used to model the population from which some sample of data has been obtained. Simple hypothesis tests and confidence intervals will be presented, in order to demonstrate how a mathematical model of variability can be used to develop tests and interval estimates. The concept of the power of a test will also be formalised.

8.1 Testing for normality: normal scores

Before using the normal distribution as a model of data, we need some way of checking whether or not the model is plausible, i.e. whether or not a set of data could plausibly have come from a normal distribution. There are rigorous tests for this, but in this module we shall concentrate on a simple graphical method for checking normality. Note that no test will *prove* normality. The best that we can do is to fail to reject the idea that the data come from a normal distribution. Whenever possible, you should test that an assumption of normality is reasonable.

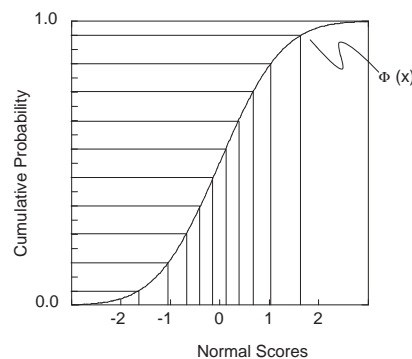
Normal scores plots are based on a simple idea, which can be understood by following an intuitive argument. Suppose that you take two samples x_1, \dots, x_n and z_1, \dots, z_n of size n from a standard normal distribution. If you sorted both samples into ascending order, you would expect that the smallest z would be close to the smallest x , that the next smallest z would be close to the next smallest x , and so on. Thus if you plotted the sorted z values against the sorted x values then you would expect to see something very close to a straight line.

Any normal r.v. is just a linear transformation of a standard normal r.v., so that if we take a random sample of observations y_1, \dots, y_n from any normal distribution and plot the sorted y values against the sorted x values (from $N(0,1)$) then we would get something close to a straight line plot. On the other hand, if y_1, \dots, y_n come from a non-normal distribution then we would not get a straight line. As an example, here is a plot of a random sample from $N(3, 25)$ plotted against a sample of the same size from $N(0, 1)$:



Thus one way of checking the normality of a sample would be to simulate a second sample (of the same size) from $N(0, 1)$ (or any normal distribution) and to produce a plot like that above, which would then be checked visually for linearity. The problem with this is that plotting one random sample against another gives a very variable plot, whose closeness to a straight line is hard to judge. For this reason it is better to produce a sort of idealised average sample from $N(0, 1)$, known as *normal scores*.

To see how to generate an idealised sample, consider a sample of 10 from a standard normal distribution. ‘On average’, we would expect 1 data point to lie below the 10% quantile of the c.d.f., 2 to lie below the 20% quantile, 3 to lie below the 30% quantile, and so on, until we get to 10 points below the 100% quantile. Thus to get an idealised data set, we choose data points so that this is exactly true, as shown in the following diagram (where Φ denotes the c.d.f. of a standard normal distribution):



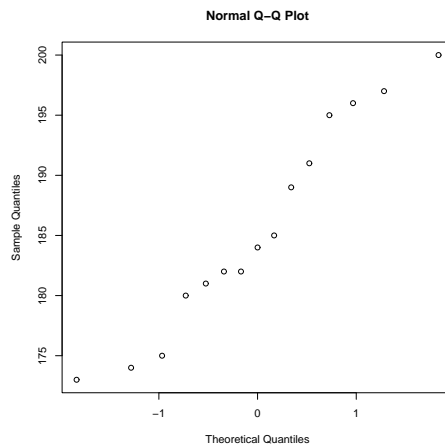
Note that 10 points have been spaced evenly along the (vertical) cumulative probability axis and that these values have then been fed into the inverse c.d.f. for the standard normal distribution to get corresponding *normal scores*. The normal scores are then used as an idealised normal sample to plot against a (sorted) set of data points, whose normality we would like to check. This procedure can be applied to samples of any size n . The formula used for normal scores as calculated in the diagram is

$$s_i = \Phi^{-1} \left(\frac{i - 0.5}{n} \right) \quad i = 1, \dots, n, \quad (8.1)$$

where Φ^{-1} denotes the inverse c.d.f. of a standard normal distribution. (The importance of the values $(i - 0.5)/n$ is that they are n evenly spaced numbers between 0 and 1.)

Example: Early in the 20th century, a Colonel L.A. Waddell collected 32 skulls from Tibet. He collected 2 groups: 17 from graves in Sikkim and 15 from a battlefield near Lhasa. Here are maximum skull length measurements (in mm) for the Lhasa group: 182, 180, 191, 184, 181, 173, 189, 175, 196, 200, 185, 174, 195, 197, 182. Before doing anything with these data that involves assuming normality of distribution, it is wise to check that they could plausibly be normal.

Here is a normal scores plot of the above data.



This plot was produced in R by the command

```
qqnorm(skull.lhasa)
```

(The data were in `skull.lhasa`.) R has given this plot the heading ‘Normal Q-Q Plot’ (short for ‘Normal quantile-quantile plot’) because the *sample quantiles* (i.e. observations sorted into increasing numerical order) are plotted vertically against the *theoretical quantiles* (i.e. normal scores) s_1, \dots, s_n along the horizontal axis.

Aside

R calculates the normal scores using the formula

$$\Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right) \quad i = 1, \dots, n, \quad (8.2)$$

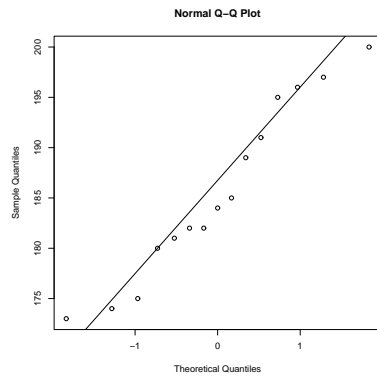
which is slightly better than (8.1). Formula (8.2) attempts to account for the fact that in any interval, a normally distributed r.v. is always slightly more likely to be in the half of the interval closer to the mean than in the other half.

To help us check linearity of a normal scores plot, we can add a straight line. The R command `qqline` adds the straight line which passes through the first and third quartiles (of the observations and of the normal scores). If the data lie near this line then this indicates that normality is acceptable. [Further, very crude estimates of the population mean and standard deviation can be obtained using the line – the population mean is very approximately the point on the vertical axis corresponding to 0 on the horizontal (normal scores) axis, while the standard deviation σ is very approximately half the distance between the points on the vertical axis corresponding to ± 1 on the horizontal (normal scores) axis.] For the skull data, the R commands

```
qqnorm(skull.lhasa)
qqline(skull.lhasa)
```

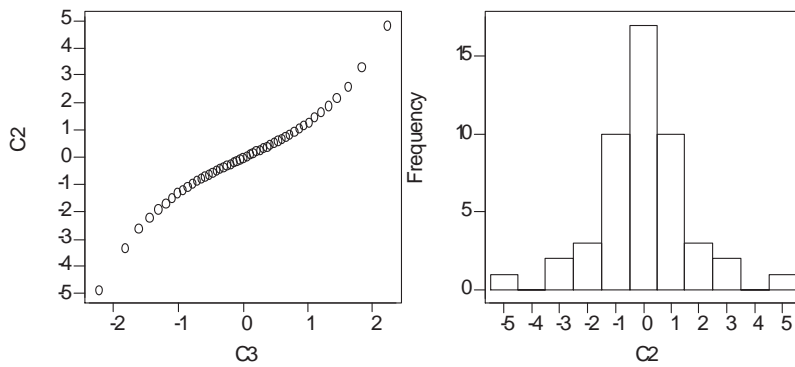
produce the following plot.

Normal scores plots require some judgement in order to decide what is ‘close enough’ to a straight line. You can acquire this judgement by simulating some samples from normal distributions and checking their normal scores plots. It is also possible to produce tests of normality using the correlation coefficient of data and normal scores as a test statistic to test the null hypothesis of normality against the alternative of non-normality. The magnitude of the sample correlation coefficient will be high for normal data, lower for non-normal data. (You could obtain the distribution by a computer-intensive method.)

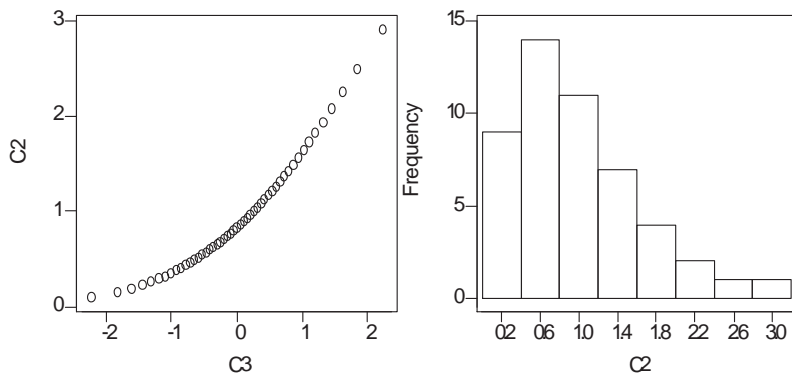


8.1.1 Interpretation of normal scores plots

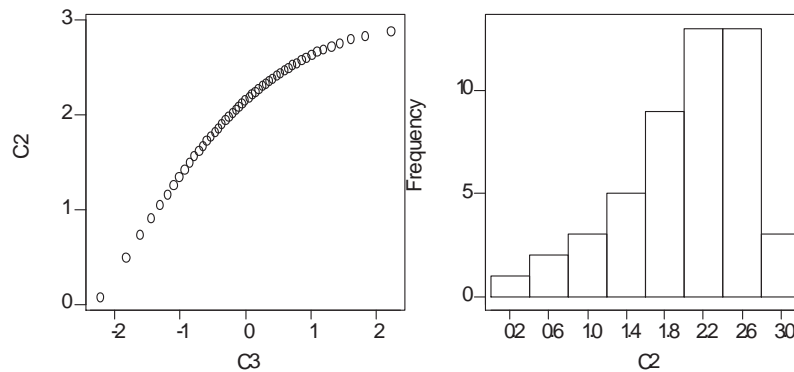
Here are some examples of the shapes obtained when a sample is not from a normal distribution. Normal scores plots are shown on the left and histograms on the right. First, here is the normal scores plot of a sample from a distribution with more probability in the tails and centre of the distribution and less in the 'shoulders' than a normal distribution.



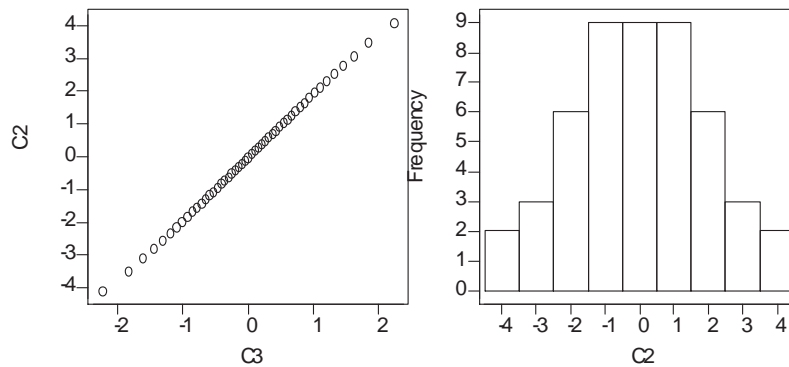
The following is for observations from a positively skewed distribution.



Here is an example for observations from a negatively skewed distribution.



Here is the equivalent plot for a normal sample.



In conclusion: if you are using the normal distribution as a statistical model for a set of data then check its validity using a normal scores plot.

8.2 Using a normal distribution as a model when the variance σ^2 is known

This section will deal with the situation in which you wish to learn something about a population on the basis of a random sample from that population, and you can assume that the population may be modelled using a normal distribution of known variance. The methods developed for this extremely simple and rather contrived case are not going to be very widely applicable, but the usefulness of the methods is not the point. This section is about how a model of random variation can be applied to infer things from data. It is the process of doing this and the principles which are employed (rather than the specific simple methods derived) that are of wide applicability.

In this context 'population' means the set of all possible units which could have been selected for the sample of data being analysed. For example, if you wish to investigate IQ of students at St Andrews, you might measure the IQ's of a random sample of students. The 'population' in this case is just all students in St Andrews, or (more specifically) the IQ measurements of all students in St Andrews. Sometimes the 'population' is rather more abstract. A physicist investigating the speed of light might measure the time taken for a light pulse to travel some fixed distance. The population of interest in this case is the population of all times that might possibly have been measured, given experimental error (the population from which you would be sampling if you measured the time taken again and again and again).

Usually it is impractical to work with the whole population (even in the cases where it is finite), so instead we test hypotheses about the population on the basis of a sample drawn from it. In the IQ example we could ask what the average IQ of St Andrews students is, or test the hypothesis that it takes

some value. In the speed of light case, information about the true speed of light is wanted on the basis of the variable measurements that can be obtained.

8.2.1 Hypothesis testing: parametric approach

Remember the general method for testing a hypothesis about the population from which a sample of data has been obtained.

1. Define a null hypothesis H_0 and an alternative hypothesis H_1 with which to compare it.
2. Choose a test statistic which will distinguish between the two hypotheses by taking ‘extreme’ values if H_1 is true and more moderate values otherwise. (For example, it might tend to take small values if H_0 is true and large values if H_1 is true.)
3. Work out what the distribution of the test statistic would be *if the null hypothesis were true*.
4. Hence determine the probability of obtaining a test statistic at least as ‘extreme’ as the one observed if the null hypothesis is true. This probability is known as the *p-value* of the data under the test.
5. A very low p-value suggests that the null hypothesis is false.

This approach applies to all the hypothesis testing met in this module. A *statistic* is any function of one or more random variables. The sample mean and sample variance are examples of statistics.

An alternative approach to hypothesis testing specifies in advance a range of ‘extreme’ values (the *critical region*) of the test statistic. If the test statistic falls in the critical region then the null hypothesis is rejected. The probability of rejecting the null hypothesis if the null hypothesis is true is called the *significance level* and often denoted by α . Note that this approach to hypothesis testing is equivalent to rejecting the null hypothesis if the p-value is less than the significance level.

Consider a set of independent observations x_1, \dots, x_n from a population which can be modelled by a normal distribution of known variance σ^2 . Using this set of observations, it is possible to test hypotheses about the population mean μ , e.g. to test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0,$$

where μ_0 is some specified number. A suitable test statistic is the mean \bar{x} of the observations, which will tend to be close to μ_0 if the null hypothesis is true but further away from μ_0 otherwise. Under the null hypothesis, x_1, \dots, x_n can be modelled as observations of r.v.’s X_1, \dots, X_n with $X_i \sim N(\mu_0, \sigma^2)$. Thus, under the null hypothesis,

$$\bar{X} \sim \quad ,$$

which can be converted into a standard normal r.v. Z :

Thus, if we had chosen to use a significance level of 5%, we would reject the null hypothesis if the observed value of Z lay outside the middle 95% of the standard normal distribution. The appropriate range of values can be looked up in tables. Alternatively, we can calculate what proportion of values are at least as improbable as the observed value under the null hypothesis, and quote this as the p-value.

In the case of the test just described, extreme values on either side of the mean are of interest, since the alternative hypothesis does not distinguish between them. However, if the alternative had been $H_1 : \mu > \mu_0$ then only values corresponding to $\bar{x} > \mu_0$ would have offered support for the alternative hypothesis, and so values of $\bar{x} < \mu_0$ would not count as ‘extreme’.

There are two types of error that can be made when hypothesis testing:

- (i) Rejecting H_0 when it is true is a *type I error*;
- (ii) Accepting H_0 when it is false is a *type II error*.

An easy way to remember which error is which is to learn ‘1, 2, ra, ra, ho, ho!’ (for I, II, reject, accept, H_0). If you conduct a test at the 5% level, what is the probability of a type I error? (Generalise this.)

Note that the null and alternative hypotheses are not treated equally. Firm evidence is required to reject the null hypothesis in favour of the alternative. In much statistical analysis, the null hypothesis is in some way simpler than the alternative, and statistics amounts to being sceptical about employing complicated explanations when a simple one will do.

Although the circumstance of knowing the population variance is rare, we can use the sample variance in place of the population variance for very large samples, so the simple one-sample normal test considered above (often known as the *z-test*) is useful in such cases. This test can be implemented in R as follows. Suppose that (we believe that) the data in `pig` come from a normal distribution with unknown mean μ and known variance 1.44, and that we wish to test whether or not $\mu = 3.5$. We could use

```
z<-(mean(pig)-3.5)/sqrt(1.44/length(pig)) # calculates z
2*(1-pnorm(z))                          # p-value for 2-sided alternative mu not 3.5
[1] 0.06907158
1-pnorm(z)                               # p-value for 1-sided alternative mu > 3.5
[1] 0.03453579
> pnorm(z)                               # p-value for 1-sided alternative mu < 3.5
[1] 0.9654642
```

Thus, at the 5% significance level, we would reject $H_0 : \mu = 3.5$ against the 2-sided alternative $H_1 : \mu \neq 3.5$ but reject H_0 against the 1-sided alternative $H_1 : \mu > 3.5$ (and accept H_0 against the 1-sided alternative $H_1 : \mu < 3.5$).

8.2.2 The power of a test

Definition: The *power* of a test is the probability of rejecting the null hypothesis when it is false. This is a very important concept. Learn it. This probability will depend on how wrong the null hypothesis is. Consider a sample of size 30 from a standard normal population. If the (incorrect) null hypothesis is $H_0 : \mu = 0.001$ then the probability of rejecting H_0 is small. However, if the (incorrect) null hypothesis is $H_0 : \mu = 100$ then the probability of rejecting H_0 is large. Because power depends on what is actually correct, it is expressed as a function of the ‘true’ values of the parameters under H_1 , in this case in terms of the mean μ . Calculation of power is best demonstrated by example.

Consider testing

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0$$

using a sample of size n from a normally distributed population with variance σ^2 . We can use the test statistic \bar{x} , the sample mean, and under the null hypothesis compare

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

with values which would be expected from $N(0, 1)$. For what range of z values would H_0 be rejected at the 5% level?

It is now straightforward to work out the probability that the test would reject the null hypothesis if the true population mean were μ .

$$\begin{aligned}
 \pi(\mu) &= P(\text{Reject } H_0 | \mu) \\
 &= P(Z > z_{\alpha/2} | \mu) \\
 &= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} | \mu\right) \\
 &= P(\bar{X} > \mu_0 + z_{\alpha/2} \sigma/\sqrt{n} | \mu)
 \end{aligned}$$

Given that the population mean is μ , we have $\bar{X} \sim N(\mu, \sigma^2/n)$. Use this result to express $\pi(\mu)$ in terms of the probability of a standard normal r.v. exceeding some appropriate value.

Finally, write $\pi(\mu)$ in terms of the c.d.f. Φ of the standard normal distribution.

8.2.3 Confidence Intervals

Confidence intervals combine some of the ideas of hypothesis testing and power. To remind you of the notion, here is an example. Consider a sample of size n from a normal population of known variance σ^2 and unknown population mean μ . On the basis of this sample, what range of hypothesised population means μ would be accepted at the 5% level*? This range is a 95% *confidence interval* for μ .

In general, a $100(1 - \alpha)\%$ *confidence interval* is the range of values of μ which would be accepted at level α against a 2-sided alternative. An $x\%$ confidence interval for a parameter is an interval having probability $x\%$ of including the true value of the parameter, in the sense that $x\%$ of intervals calculated in the same way for similar samples will include the true value of the parameter. This slightly clumsy definition is intended to emphasise the fact that the parameter being estimated is a *fixed quantity*, not a random variable; it is the interval that is random.

Make sure that you understand the last two paragraphs well. Confidence intervals are important, so it is worth being very clear about the concept. You have already calculated confidence intervals by computer-intensive methods. Now we shall see some examples of how intervals can be obtained analytically.

Returning to the sample of n observations from a normal population with variance σ^2 , let us construct a 95% confidence interval for the population mean μ . First consider

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

*It is conventional to assume a two-sided alternative.

where μ is a hypothesised population mean. We would accept (against a 2-sided alternative) any value for μ such that

$$-1.96 < Z < 1.96,$$

i.e. any value of μ such that

$$-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96$$

would be accepted. Rearranging this gives a 95% confidence interval for μ .

Example: The wavelengths of light pulses from a semiconductor laser are approximately normally distributed, with variance calculated by theory to be 100 nm^2 . The mean wavelength for individual lasers varies. Measurements of 100 pulses from a laser give an average wavelength of 598 nm. Find a 95% confidence interval for the mean wavelength of the laser.

There is an important connection between confidence intervals and (2-tailed) hypothesis tests: the hypothesis $H_0 : \theta = \theta_0$ is accepted at significance level α against $H_1 : \theta \neq \theta_0$ if and only if θ_0 lies in the $100(1 - \alpha)\%$ confidence interval for θ . For example, if I have a 95% C.I. for the mean HIV prevalence of (0.3%, 0.6%) then I can immediately reject the hypothesis that the mean rate is 0.2% at the 5% level, since 0.2% is not in the range of values that I would have accepted at the 5% level, i.e. it is not in the 95% confidence interval.

Perhaps more important is the fact that a confidence interval tells you more than a hypothesis test. If someone tells you that a study has found no evidence for an increase in burglaries in St Andrews in the last 10 years, you might be impressed with the efficiency of the local constabulary and the honesty of the population. However, such a conclusion would probably be based on failure to reject the hypothesis that the change in burglary rate was zero. If it turned out that the 95% confidence interval for the change in burglary rate was (-10%, +160%) then you would be less impressed. The confidence interval has actually told you something about the sort of change which could have been detected by the study (i.e. something about power).

9 Distributions derived from normal distributions

The statistical methods derived in the previous section are of limited use, since one rarely knows the variance of the population from which observations have come, and frequently there is not enough data in a sample to estimate σ^2 without having to worry about the uncertainty in the estimate. Methods are required that deal with the situation in which both population mean and variance are unknown, and sample sizes may be small. In order to develop the theory of such methods, we need to introduce some distributions related to normal distributions. In particular, we need to consider the χ^2 , t and F distributions.

9.1 χ^2 distributions

Definition: If Z_1, \dots, Z_n are independent r.v.'s distributed as $N(0, 1)$ then the random variable

$$U = \sum_{i=1}^n Z_i^2$$

has a χ^2 (*chi-squared*) distribution with n degrees of freedom, written as $U \sim \chi_n^2$.

An important property of χ^2 -distributed r.v.'s is immediately apparent from the definition, namely

$$U \sim \chi_n^2 \text{ and } V \sim \chi_m^2 \Rightarrow U + V \sim \chi_{m+n}^2.$$

Quantiles [†] of χ^2 distributions with various degrees of freedom are given in Table 7 of K&Y and can be found in R by using `qchisq`, e.g.

```
> qchisq(0.95,4)
[1] 9.487729
```

finds the *upper* 5% quantile (i.e. the *lower* 95% quantile) of the χ_4^2 distribution to be 9.487729.

Note that χ^2 -distributed r.v.'s are positive and that the shape of the χ_n^2 distribution depends on n .

The χ^2 distributions play an important role in testing 'goodness-of-fit' of statistical models (see later in this module), and are used for comparing sophisticated models of data using generalised linear models. (Details are given in the Honours module on *Linear Models and Data Analysis*.) For the moment, use a suitable χ^2 distribution to solve the following simple problem.

Example: Suppose that X , Y and Z are coordinates in 3-dimensional space which are independently distributed as $N(0, 1)$ (where all measurements are in cm). What is the (approximate) probability that the point (X, Y, Z) lies more than 3 cm from the origin?

[†]Recall that an *upper quantile* (or *critical point*) is a value above which some specified proportion of the area of a p.d.f. lies.

9.1.1 Mean and variance of χ^2 r.v's

Consider $V \sim \chi_n^2$. Then $V = \sum_{i=1}^n Z_i^2$, where Z_1, \dots, Z_n are independent with $Z_i \sim N(0, 1)$. Thus the mean of V is

$$\begin{aligned} E(V) &= E\left(\sum_{i=1}^n Z_i^2\right) \\ &= \sum_{i=1}^n \text{var}(Z_i) \\ &= n. \end{aligned}$$

Now use a similar approach to obtain the variance of a χ_n^2 -distribution. [Hint: Use the fact (which will be derived in Honours) that, if $Z \sim N(0, 1)$ then $E(Z^4) = 3$.]

Thus

$$V \sim \chi_n^2 \Rightarrow E(V) = n, \quad \text{var}(V) = 2n.$$

9.2 The F distributions

Definition: If U and V are independent r.v's with $U \sim \chi_n^2$ and $V \sim \chi_k^2$ then

$$\frac{U/n}{V/k}$$

has the F distribution with n and k degrees of freedom, usually denoted by $F_{n,k}$. Quantiles (alias critical points) of F distributions are given in K&Y Table 9, where columns relate to upper degrees of freedom (n) and rows to lower degrees of freedom (k). They can be obtained in R by using `qf`, e.g.

```
> qf(0.95, 2, 8)
[1] 4.45897
```

finds the *upper* 5% quantile (i.e. the *lower* 95% quantile) of the $F_{2,8}$ distribution to be 4.45897.

It is important to note that $F_{n,k}$ is not the same as $F_{k,n}$. It follows from the definition that an F -distributed r.v. is necessarily positive (with probability 1). Note also that $F_{n,\infty}$ is the distribution of U/n , where $U \sim \chi_n^2$.

The F distributions are used for testing for equality of variances, e.g. to check the assumption required for a 2-sample t -test and, more importantly, in analysis of variance when fairly complicated models of data are compared (see the sections on regression and ANOVA, later in this module).

An important property of F distributions is

$$W \sim F_{n,k} \Rightarrow \frac{1}{W} \sim F_{k,n}. \quad (9.1)$$

(This comes directly from the definition.) Most statistical tables exploit (9.1) by giving only the upper quantiles of F distributions. Usually these are all that we need. If we need lower quantiles, we can get

them as follows. First note that we shall use $F_{n,k;\alpha}$ to denote the upper α quantile of the $F_{n,k}$ distribution. Then, by the definition of an upper quantile, $P(W > F_{n,k;1-\alpha}) = 1 - \alpha$, so that

$$P(W < F_{n,k;1-\alpha}) = \alpha = P(1/W \geq 1/F_{n,k;1-\alpha}) = P(1/W > F_{k,n;\alpha}),$$

i.e.

$$F_{n,k;1-\alpha} = 1/F_{k,n;\alpha}. \quad (9.2)$$

Given that $F_{3,2;0.025} = 39.17$, find $F_{2,3;0.975}$.

Later, we shall use F distributions to test equality of variances in samples from normal distributions. They will also feature heavily in the ANOVA section of the module.

9.3 The t distributions

Definition: Let Z and Y be independent r.v.'s with $Z \sim N(0, 1)$ and $Y \sim \chi_n^2$. Then

$$T = \frac{Z}{\sqrt{Y/n}} \quad (9.3)$$

has a t distribution with n degrees of freedom, which is written as $T \sim t_n$. The mean and variance of T are $E(T) = 0$ and $\text{var}(T) = n(n-2)$ for $n > 2$.

The shape of the p.d.f. of t_n depends on n . The p.d.f. of t_n looks like a normal p.d.f., but more of the probability is in the centre and tails of the distribution relative to the 'shoulders'. As $n \rightarrow \infty$, the t_n distribution approaches the $N(0, 1)$ distribution. In practice, the approximation $t_n \overset{\cdot}{\sim} N(0, 1)$ can be used for $n \geq 30$.

We shall use $t_{n;\alpha}$ to denote the upper α -quantile of the t distribution with n degrees of freedom. Quantiles of the t distributions with various degrees of freedom can be found in Table 8 of K & Y or obtained in R by using `qt`, e.g.

```
> qt(0.95,8)
[1] 1.859548
```

finds $t_{8;0.05}$, the upper 5% quantile (i.e. the lower 95% quantile) of the t_8 distribution, to be 1.859548. The main use of t distributions is for dealing with samples from normal distributions with unknown mean and variance.

10 Using t distributions

When we wish to model samples as having been taken from a population with a normal distribution of unknown mean and variance, it turns out to be appropriate to use t distributions. In order to understand why, we need the key result on the joint distribution of the sample mean \bar{x} and the sample variance s^2 of samples from normal distributions.

10.1 The independence and distributions of \bar{x} and s^2

The following result is the key to much of the theory of normal distributions.

Theorem

Let \bar{X} and s^2 be the sample mean and sample variance of random samples of size n from the $N(\mu, \sigma^2)$ distribution. Then

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1), \quad (10.1)$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2, \quad (10.2)$$

$$\bar{X} \text{ and } s^2 \text{ are independent.} \quad (10.3)$$

The proof of this theorem will have to wait until Honours. Note that property (10.3) is quite remarkable. It holds only for normal distributions.

Now combine the quantities on the left hand sides of (10.1)–(10.2) as follows:

$$\begin{aligned} T &= \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \\ &= \frac{\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2} / (n-1)}}. \end{aligned} \quad (10.4)$$

The quantity on the top of (10.4) is a standard normal r.v., while the quantity on the bottom has the form $\sqrt{Y/(n-1)}$, where $Y \sim \chi_{n-1}^2$. Hence, from the definition (9.3), $T \sim t_{n-1}$, i.e.

$$\frac{(\bar{X} - \mu)}{\sqrt{s^2/n}} \sim t_{n-1}. \quad (10.5)$$

This result will be very useful. The quantity T depends on the population mean μ , but not on the (unknown) population variance σ^2 . Hence we can use the distributional result (10.5) to test hypotheses about the mean of normal populations without prior knowledge of the population variance.

10.2 One-sample t -tests and confidence intervals

A set of 39 observations on pulse rates (in heart beats per minute) of Indigenous Peruvians had sample mean 70.31 and sample variance 90.219. A normal scores plot shows no major departures from normality, so it is reasonable to assume that the data are from a normal population. The question of interest is whether or not this group could plausibly be a random sample from a population with mean pulse rate of 75. Test at the 1% level

$$H_0 : \mu = 75 \text{ vs. } H_1 : \mu \neq 75.$$

What test statistic is suitable, and what will its distribution be under the null hypothesis?

Find the range of values for the test statistic which would cause you to reject H_0 at the 1% level. ($t_{38;0.005} \approx t_{40;0.005} = 2.7045$.) Do you reject H_0 ?

Now find the range of hypothesised values of μ for which you would have accepted H_0 , i.e. find a 99% confidence interval for μ .

Finally, write out the general formula for a $100(1 - \alpha)\%$ confidence interval for the mean based on a sample of n points.

This test can be implemented in R as follows. Suppose that the data are in `pulse`. We could use

```
> t<-(mean(pulse)-75)/sqrt(var(pulse)/length(pulse)) # calculates t
> t
[1] -3.083593
> 2*pt(t,length(pulse)-1) # p-value for 2-sided alternative mu not 75
[1] 0.003799049
```

(Since $t < 0$ here, we obtained the p -value by doubling the c.d.f. of t .)

We could have constructed a 95% confidence interval by

```
> cil<-mean(pulse)+qt(0.025,length(pulse)-1)*sqrt(var(pulse)/length(pulse))
> cil
[1] 67.23099
> ciu<-mean(pulse)+qt(0.975,length(pulse)-1)*sqrt(var(pulse)/length(pulse))
> ciu
[1] 73.38901
```

Thus a 95% confidence interval for μ is (67.23, 73.39). As this does not contain 75, we reject H_0 at the 5% significance level.

10.3 Paired t -tests

There are many situations in which two measurements are made on each of several units. In general, the two measurements made on each unit are *dependent*, so we should not treat them as independent. Furthermore, the distribution of measurements across units may be very non-standard. In this situation, it is sensible to work with the differences between the two measurements on each unit. For example, the following table gives corneal thickness in microns for both eyes of patients who have glaucoma in one eye. (The cornea is the transparent membrane right at the front of the eye. The lens is further back.)

| | | | | | | | | |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|
| Glaucoma | 488 | 478 | 480 | 426 | 440 | 410 | 458 | 460 |
| Healthy | 484 | 478 | 492 | 444 | 436 | 398 | 464 | 476 |

Because the corneal thicknesses are likely to be similar in the two eyes of any patient, it is inappropriate to model the two observations made on the same patient as independent random variables. Take the differences between the thicknesses of the two corneas and test whether or not there is evidence that the corneas differ in thickness between the good eye and the diseased eye. (You may like to know that $\sum d_i = -32$ and $\sum d_i^2 = 936$, where d_i denotes the i th difference. $t_{7;0.15} = 1.1192$.)

Note the assumptions that you have made: (i) that the differences in corneal thickness can be modelled as a random sample from a normal distribution, (ii) that the differences can be modelled as observations of *independent* random variables. Make sure that you are clear about the question which is being asked. The medical question is whether or not this sample of patients provides evidence that the corneal thickness is related to glaucoma in general. Obviously, there is a difference in the thicknesses in the sample. The statistical question is whether or not the difference is large enough that we can be confident that the difference applies in general (i.e. that other samples would show the same pattern). We want to know whether the effect is real or just the result of chance variability.

In medicine, questions like this turn up all the time. For example, people are always finding links between disease and some aspect of lifestyle. It is important to work out which of these may be real and which just the result of chance variability.

10.4 Two-sample t -tests

Two-sample t -tests deal with cases in which we have two sets of data. Here, we shall consider only situations in which can be modelled as being random samples from 2 different normal populations, of unknown mean and unknown variance. For simplicity, we shall assume that the (unknown) variances of the two populations are the same. (Fairly complicated modifications have to be made if this cannot be assumed.)

Consider two sets of independent random variables X_1, \dots, X_n and Y_1, \dots, Y_m , where

$$X_i \sim N(\mu_X, \sigma^2) \quad \text{and} \quad Y_j \sim N(\mu_Y, \sigma^2).$$

The distributions of the sample means of the two samples are

$$\bar{X} \sim N(\mu_X, \sigma^2/n) \quad \text{and} \quad \bar{Y} \sim N(\mu_Y, \sigma^2/m),$$

so from standard results (derived earlier) on sums of normal r.v's,

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right).$$

Now find a transformation of the difference in sample means which will have a standard normal distribution.

If we knew σ^2 then we could use this result to test hypotheses about the difference in means between the two groups, but usually σ^2 is unknown. To deal with this problem, first find the distribution (using the distribution of the sample variance and additivity of χ^2 r.v's) of

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{1}{\sigma^2} \sum_{j=1}^m (Y_j - \bar{Y})^2$$

as

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{1}{\sigma^2} \sum_{j=1}^m (Y_j - \bar{Y})^2 \sim \quad .$$

Note that the above quantity can be written as

$$\frac{(n+m-2)s^2}{\sigma^2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{\sigma^2},$$

where s_X^2 and s_Y^2 are the sample variances for the 2 groups, and

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

The quantity s^2 is sometimes called the *pooled sample variance*. It is an unbiased estimate of σ^2 . Use the two distributional results that you have obtained so far to obtain a test statistic which has a t distribution under H_0 .

This result allows us to test hypotheses about the difference in means between the populations from which the 2 groups of data have come.

As a quick example, consider 2 small samples from 2 normal populations:

| | | | | | |
|-----|----|----|----|----|----|
| x | 11 | 10 | 14 | 12 | 13 |
| y | 8 | 3 | 4 | 9 | |

Test the hypothesis that the two population means are equal against the alternative that they are not. (Note that $(n - 1)s_X^2 = 10$ and $(m - 1)s_Y^2 = 26$, and that $t_{7,0.025} = 2.3646$.)

The 2-sample t-test can be implemented in R as follows. Suppose that the data are in `x` and `y`. We could use

```
> ssq<-((length(x)-1)*var(x) + (length(y)-1)*var(y))/(length(x)+length(y)-2) # calculates
> # 'pooled' sample variance
> t<-(mean(x)-mean(y))/sqrt(ssq*(1/length(x)+1/length(y))) # calculates t
> t
[1] 3.944053
> 2*(1-pt(t,length(x)+length(y)-2)) # p-value for 2-sided alternative
[1] 0.005574311
```

10.5 Testing equality of variances

To apply the distributional result derived in the last section requires that the two samples being compared come from populations with equal variance. This can be tested using an F distribution. Consider two samples X_1, \dots, X_n and Y_1, \dots, Y_m , each from a normal distribution. (As usual, assume that the observations are of independent r.v.'s.) If σ_X^2 and σ_Y^2 denote the population variances of X_1, \dots, X_n and Y_1, \dots, Y_m , respectively, then we know that

$$\frac{1}{\sigma_X^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2 \quad \text{and} \quad \frac{1}{\sigma_Y^2} \sum_{j=1}^m (Y_j - \bar{Y})^2 \sim \chi_{m-1}^2.$$

Now consider testing

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{vs.} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

As usual, we would like a test statistic which will distinguish between the two hypotheses, and the distribution of which can be worked out assuming that H_0 is true. Under the null hypothesis, the population variances are equal, to σ^2 say, so consider

$$\frac{\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)}{\frac{1}{\sigma^2} \sum_{j=1}^m (Y_j - \bar{Y})^2 / (m-1)}.$$

From the definition of the F distributions, it follows that this quantity will be distributed as $F_{n-1, m-1}$ and that the two factors involving σ^2 cancel to give a statistic which involves only quantities calculable from the samples. However, if the alternative hypothesis were true then the population variances would not be equal and so would not cancel, so that the test statistic would be abnormally small or large for an observation of an F -distributed r.v. Thus, if the null hypothesis is true we have

$$\frac{s_X^2}{s_Y^2} \sim F_{n-1, m-1} \quad \text{or equivalently} \quad \frac{s_Y^2}{s_X^2} \sim F_{m-1, n-1}.$$

To test the null hypothesis, the above two quantities can be compared with the upper quantiles of their respective F distributions. Note that it is enough to compare the larger of s_X^2/s_Y^2 and s_Y^2/s_X^2 with the upper quantiles of the appropriate F distribution.

As a quick example, consider two random samples, one of size 11 and the other of size 16 from two normal populations. The sample variance of the first is 20 and the sample variance of the second is 30. At the 5% level, is there evidence to reject the hypothesis that the two populations have the same variance? ($F_{15, 10; 0.025} = 3.522$.)

11 Maximum Likelihood Estimation

In previous sections, mathematical descriptions of random variation were used to model the procedure of taking a random sample from a population, and to develop methods for making inferences about populations from samples. Those inferences involved finding confidence intervals for (or testing hypotheses about) parameters of the model used to describe the population. This section will look at a method for estimating the parameters of a model using data. The aim is to find the most likely value for a parameter, given some data which relate to it.

As an example, some unknown proportion of the population supports the Conservative party. This proportion can be viewed as a parameter p_c , which we would like to estimate. The obvious way to estimate it is to take a random sample from the population and to determine what proportion \hat{p}_c vote Conservative. This proportion can be used as an estimate of the population parameter p_c . In this example, the model for the population is extremely simple ('A proportion p_c of the population vote Conservative') and it is obvious how to obtain a suitable parameter estimate. In other cases it is not so obvious how to estimate parameters, and a general method is needed. Sticking with the political example, we might believe that the probability of being a Conservative voter depends on income in a linear way, so that $P(\text{Votes Conservative}) = \alpha + \beta \times \text{income}$. Again, we would take a random sample of voters and collect data on income and voting intentions for each. How could the parameters α and β be estimated?

A general method for using data to estimate model parameters is the method of maximum likelihood. It is a simple idea, most easily grasped by example. Consider rolling a loaded die, which has probability p of coming up with a 6. You roll it 10 times and get 4 sixes. The probability of this happening is

$$P(4 \text{ sixes in } 10 \text{ rolls}) = \binom{10}{4} p^4 (1-p)^6.$$

What is the value of p which will make this probability as high as possible?

This estimate is known as the *maximum likelihood estimate* of p . The idea is that the most likely value for the parameter is the one which makes the data appear most probable. Maximum likelihood estimation is the process of finding the parameters which make a set of data look as probable as possible.

In greater generality, maximum likelihood estimation works as follows. Consider a set of observations x_1, \dots, x_n which are modelled as observations of independent discrete random variables with probability function $f(x; \theta)$ which depends on some (vector of) parameters θ . According to the model, the probability of obtaining the observed data is proportional to the product of the p.f.'s for each observation, i.e.

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta).$$

We seek the parameters of the model that make the data look most probable, so we seek to maximise $L(\theta; x_1, \dots, x_n)$ w.r.t. θ . When $L(\theta; x_1, \dots, x_n)$ is considered as a function of the parameters in this way, it is known as the *likelihood* of the parameters (rather than the probability of the data). Note that the logarithm of a function is maximised at the same set of parameters as the function itself. Very often it is easier to maximise the *log-likelihood*

$$l(\theta; x_1, \dots, x_n) = \log L(\theta; x_1, \dots, x_n),$$

rather than the likelihood $L(\theta; x_1, \dots, x_n)$. Note that

$$l(\theta; x_1, \dots, x_n) = \sum_{i=1}^n \log f(x_i; \theta).$$

(Recall that $\log ab = \log a + \log b$.)

Example: Suppose that you have 4 observations x_1, x_2, x_3, x_4 on independent Poisson distributed r.v's, each with p.f.

$$f(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x \geq 0.$$

First form the likelihood of λ and then the corresponding log-likelihood.

Now maximise the likelihood w.r.t. λ .

The expression that you have just derived is known as an *estimator* if you consider it as a *function* (of x_1, \dots, x_n). The value of this function which is obtained by evaluating it on observation values x_1, \dots, x_n is an *estimate*. Suppose that the observations in this case are 1, 3, 8 and 2. What is the maximum likelihood estimate?

Note that, in general, you should check that you have obtained a *maximum* likelihood estimator and not a *minimum* likelihood estimator.

A more complicated example: Suppose that you have a series of measurements y_1, \dots, y_n of radioactive emission counts from samples of caesium of masses x_1, \dots, x_n , respectively. You wish to model the counts as Poisson random variables, where each Y_i has mean αx_i . Obtain the maximum likelihood estimator of α (the radioactivity per unit mass).

11.1 Likelihood for continuous distributions

The examples met so far have dealt with the really simple case of a single parameter estimated using discrete data. Maximum likelihood estimation works just as well for continuous random variables. The only difference is that we form the likelihood from the product of the p.d.f.'s of the random variables used to model the data. If x_1, \dots, x_n are observations of independent continuous r.v.'s with p.d.f.'s $f(x_i; \boldsymbol{\theta})$ which depend on some parameters $\boldsymbol{\theta}$ then the likelihood function is just

$$L(\boldsymbol{\theta}; x_1, \dots, x_n) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta}).$$

This is still a likelihood, but can no longer be interpreted as a probability of getting the observed data, given $\boldsymbol{\theta}$, only as a probability density (probability per unit interval [or per unit area or per unit volume]) of getting the observed data. This makes no difference to the actual calculations. We still maximise the likelihood w.r.t. the parameters, and it is still easier to use the log-likelihood in most cases.

Consider an example involving two parameters. Suppose that we have some observations x_1, \dots, x_n , which we wish to model as observations of i.i.d. r.v.'s from a normal distribution with mean μ and variance σ^2 , where the two parameters are unknown and so have to be estimated from data. First the likelihood function is formed by multiplying together the p.d.f.'s for the n r.v.'s, evaluated at the observed values to give

$$\begin{aligned} L(\mu, \sigma^2; x_1, \dots, x_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}. \end{aligned}$$

Again, it is convenient to work with the log-likelihood

$$l(\mu, \sigma^2; x_1, \dots, x_n) = \log L(\mu, \sigma^2; x_1, \dots, x_n) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

To maximise this, find its partial derivatives w.r.t. σ^2 and μ , and set these equal to zero. In this particular case, it is easiest to start with μ .

Now find the partial derivative of the log-likelihood w.r.t. σ^2 (or, if you prefer, w.r.t. σ), and obtain the m.l.e. by setting this to zero and substituting for μ .

Note that the maximum likelihood estimator $\hat{\sigma}^2$ of the variance σ^2 is **not** the sample variance s^2 (which is the usual estimator of σ^2).

Note: In general, maximum likelihood estimates are biased (although often the bias is fairly small). However, m.l.e.'s do have the advantage of being consistent.

11.2 Invariance of m.l.e's

The invariance property of maximum likelihood estimators:

If $\hat{\theta}$ denotes the maximum likelihood estimator of θ and g is any function of θ then the maximum likelihood estimator of $g(\theta)$ is $g(\hat{\theta})$.

Example: Suppose that x_1, \dots, x_k are observations on independent binomial r.v's, each with n trials and unknown probability p . The likelihood of p is

$$L(p; x_1, \dots, x_k) = \prod_{i=1}^k \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}.$$

Find the maximum likelihood estimator of p .

Using the invariance property, deduce the maximum likelihood estimators of the mean and variance of the $\text{bin}(n, p)$ distribution.

12 Regression

Many statistical investigations are concerned with relationships between two (or more) variables, e.g. height and weight. In many important cases, one variable (known as the *explanatory variable*^{*}) can be measured without error (or with negligible error), whereas the other variable (known as the *response variable*[†]) is random. In general, the distribution of the response variable (Y , say) when the explanatory variable (x , say) takes a given value depends on the value of x . Thus we can think of x as ‘explaining’ the corresponding observed value y of Y . Alternatively, we can imagine that if x varies then y ‘responds’. The relationship between the response variable and the explanatory variable is known (for historical reasons) as *regression*.

The typical regression problem is concerned with finding the regression relationship between the response variable Y and the explanatory variable x , on the basis of n pairs of observations $(x_1, y_1), \dots, (x_n, y_n)$ on (x, Y) .

12.1 Linear regression

The simplest form of regression is *linear regression*, in which the response variable Y is related to the explanatory variable x by

$$E(Y) = \alpha + \beta x, \quad (12.1)$$

where α and β are (unknown) parameters. Note that (12.1) says that $E(Y)$ depends *linearly* on x . The line

$$y = \alpha + \beta x, \quad (12.2)$$

is often called the (*population*) *regression line*, and α and β are called the *regression parameters*. It is useful to write Y_i for the random variable Y associated with the value x_i of x , for $i = 1, \dots, n$. Then (12.1) gives

$$E(Y_i) = \alpha + \beta x_i, \quad i = 1, \dots, n. \quad (12.3)$$

This can be re-written as

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n, \quad (12.4)$$

where $\epsilon_1, \dots, \epsilon_n$ are independent random variables with mean zero. Because the parameters α and β (and the error term) enter the model (12.4) in a linear way, this model for Y_1, \dots, Y_n is an example of a *linear model*. Such models are considered in detail in the Honours module *Generalized Linear Models and Data Analysis*.

12.1.1 Least-squares estimation

We need to estimate the values of the parameters α and β , i.e. to fit the model to data. A sensible way of doing this is by *least squares*. Consider the vertical distances

$$e_i = y_i - (\alpha + \beta x_i), \quad i = 1, \dots, n, \quad (12.5)$$

between the observed values y_1, \dots, y_n and the corresponding values $\alpha + \beta x_1, \dots, \alpha + \beta x_n$ given by the model. An intuitively appealing way of measuring the difference between the data and the model is by the sum of squares

$$\begin{aligned} S(\alpha, \beta) &= \sum_{i=1}^n \{y_i - (\alpha + \beta x_i)\}^2 \\ &= \sum_{i=1}^n e_i^2. \end{aligned} \quad (12.6)$$

^{*}Explanatory variables are also known as ‘covariates’, ‘predictor variables’, or even ‘independent variables’ (but it is best to avoid this last phrase).

[†]Response variables are sometimes known as ‘dependent variables’.

The *method of least squares* estimates α and β by the values $\hat{\alpha}$ and $\hat{\beta}$ of α and β which minimise $S(\alpha, \beta)$.

Obtain expressions for the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$. (Hint: First differentiate $S(\alpha, \beta)$ w.r.t. α , to get an expression for α in terms of β . Then substitute this into the expression which results from differentiating $S(\alpha, \beta)$ w.r.t. β .)

The least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are often expressed as

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \quad (12.7)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad (12.8)$$

where

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (12.9)$$

$$S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (12.10)$$

$$S_{YY} = \sum_{i=1}^n (y_i - \bar{y})^2. \quad (12.11)$$

(Although S_{YY} is not needed in (12.7) and (12.8), it will be used later.)

Note that estimation by least squares requires no assumptions about the distributions of Y_1, \dots, Y_n . Since Y_1, \dots, Y_n are random variables, the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ are also random variables. We can calculate their expectations as follows, using (12.8)–(12.10) and (12.3), and noting that x_1, \dots, x_n are *fixed*.

$$\begin{aligned}
 E[\hat{\beta}] &= E\left[\frac{1}{S_{XX}} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})\right] \\
 &= \frac{1}{S_{XX}} \sum_{i=1}^n (x_i - \bar{x}) E(Y_i - \bar{Y}) \\
 &= \frac{1}{S_{XX}} \sum_{i=1}^n (x_i - \bar{x}) \beta (x_i - \bar{x}) \\
 &= \beta.
 \end{aligned} \tag{12.12}$$

Similarly,

$$\begin{aligned}
 E(\hat{\alpha}) &= E(\bar{Y} - \hat{\beta}\bar{x}) \\
 &= (\alpha + \beta\bar{x}) - \beta\bar{x} \\
 &= \alpha.
 \end{aligned} \tag{12.13}$$

Thus $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β , respectively.

12.1.2 The normal linear regression model

In general, we need to do more than just find (point) estimates of the parameters α and β ; we need to test hypotheses about them and to construct confidence intervals for them. In order to do this, we need to make assumptions about the distributions of Y_1, \dots, Y_n . We shall suppose that Y_1, \dots, Y_n are independent, normally distributed with the same variance, and satisfy (12.3), i.e.

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, \dots, n \quad \text{with } Y_1, \dots, Y_n \text{ independent.} \tag{12.14}$$

Write down the likelihood function of α and β .

$$L(\alpha, \beta; (x_1, y_1), \dots, (x_n, y_n)) =$$

Hence obtain the log-likelihood function

$$l(\alpha, \beta; (x_1, y_1), \dots, (x_n, y_n)) =$$

Note two things about this log-likelihood:

- (i) it can be maximised w.r.t. α and β when σ^2 is unknown,
- (ii) maximising the log-likelihood is equivalent to minimising $S(\alpha, \beta)$.

Property (ii) means that, for the normal linear regression model (12.14) the maximum likelihood estimates are equal to the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$.

Since (from (12.8)–(12.10)) $\hat{\alpha}$ and $\hat{\beta}$ are linear combinations of the normal random variables Y_1, \dots, Y_n , $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed. It follows from this, together with (12.12)–(12.13) and calculation of $\text{var}(\hat{\beta})$ and $\text{var}(\hat{\alpha})$ that

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{XX}}\right) \quad (12.15)$$

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\right). \quad (12.16)$$

Since the variance σ^2 in (12.14) is usually unknown, it must be estimated from the data. The quantity s^2 is defined by

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (12.17)$$

where

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \quad (12.18)$$

is the *fitted value* corresponding to x_i . Calculation shows that s^2 is an unbiased estimator of σ^2 . Further calculation (postponed to Honours) shows that

$$(n-2) \frac{s^2}{\sigma^2} \sim \chi_{n-2}^2 \quad (12.19)$$

and

$$s^2 \text{ is independent of } \hat{\alpha} \text{ and } \hat{\beta} \quad (12.20)$$

(which are reminiscent of (10.2) and (10.3), respectively). [**Warning:** $\hat{\alpha}$ and $\hat{\beta}$ are *not* independent!]

Putting (12.15)–(12.16) and (12.19) together, we obtain

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{s^2}{S_{XX}}}} \sim t_{n-2} \quad (12.21)$$

$$\frac{\hat{\alpha} - \alpha}{\sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}} \sim t_{n-2}, \quad (12.22)$$

which are reminiscent of (10.5).

Results (12.21)–(12.22) are the basis of inference on α and β . In particular, they enable us to (use R to) calculate confidence intervals for α and β . For example, 95% confidence intervals for α and β are

$$\hat{\beta} \pm t_{n-2;0.025} \sqrt{\frac{s^2}{S_{XX}}} \quad (12.23)$$

$$\hat{\alpha} \pm t_{n-2;0.025} \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}, \quad (12.24)$$

respectively.

Good news: You do not need to memorise the expressions under the square root signs on the right hand sides of (12.23) and (12.24). The quantities

$$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{s^2}{S_{XX}}} \quad (12.25)$$

$$\text{s.e.}(\hat{\alpha}) = \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)} \quad (12.26)$$

are provided by R. These quantities are the *standard errors* (i.e. the estimated standard deviations) of $\hat{\beta}$ and $\hat{\alpha}$, respectively.

12.2 Regression using R

Linear regression can be carried out in R using the `lm` (linear modelling) command, which is designed for fitting quite general linear models. Recall that the linear regression model (12.4) is a linear model, in that the parameters α and β (and the error term) enter the model in a linear way.

The use of `lm` can be illustrated by the following example (which has been chosen partly because it is simple enough that the calculations could be done ‘by hand’ instead).

Example

The following measurements give the concentration of chlorine (in parts per million) in a swimming pool at various times after chlorination treatment.

| | | | | | | |
|-------------------|-----|-----|-----|-----|-----|-----|
| time (hours) | 2 | 4 | 6 | 8 | 10 | 12 |
| chlorine (p.p.m.) | 1.8 | 1.5 | 1.4 | 1.1 | 1.1 | 0.9 |

Consider the following R session.

```
> pool
      [,1] [,2] [,3] [,4] [,5] [,6]
time    2.0  4.0  6.0  8.0 10.0 12.0
chlorine 1.8  1.5  1.4  1.1  1.1  0.9
> swim<-lm(chlorine~time)
> swim
```

```
Call:
lm(formula = chlorine ~ time)
```

```
Coefficients:
(Intercept)      time
  1.90000     -0.08571
```

First the data (in `pool`) are inspected. Then the object `swim` is defined as the linear model in which `chlorine` is regressed on `time`, i.e. the linear regression equation

$$\text{chlorine} = \alpha + \beta \text{time}$$

is fitted to the data by least squares. Then `swim` contains the parameter estimates $\hat{\alpha} = 1.9$ and $\hat{\beta} = -0.0857$.

More detail of this regression is given by applying the `summary` command to `swim`, as follows.

```
> summary(swim)

Call:
lm(formula = chlorine ~ time)

Residuals:
    1     2     3     4     5     6
0.07143 -0.05714  0.01429 -0.11429  0.05714  0.02857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.900000   0.074642  25.455 1.41e-05 ***
time        -0.085714   0.009583  -8.944 0.000864 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08018 on 4 degrees of freedom
Multiple R-Squared:  0.9524, Adjusted R-squared:  0.9405
F-statistic:   80 on 1 and 4 DF,  p-value: 0.0008642
```

The *residuals* r_1, \dots, r_n are the differences between the data y_1, \dots, y_n and the fitted values \hat{y}_i given by (12.18), i.e.

$$r_i = y_i - \hat{y}_i, \quad i = 1, \dots, n. \quad (12.27)$$

The residuals have already appeared in the definition (12.17) of s^2 . As we shall see later, they play an important role in assessing how well the model fits the data.

The columns of the section of the above output which is labelled ‘**Coefficients:**’ provide

- (i) a name for the parameter ((**Intercept**) for α ; **time** for β);
- (ii) the estimates ($\hat{\alpha}$ and $\hat{\beta}$) of α and β ;
- (iii) the standard errors of $\hat{\alpha}$ and $\hat{\beta}$;
- (iv) the values of the t -statistics given by the left hand sides of (12.21) and (12.22);
- (v) the p -values of these t -statistics in tests of $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$ and of $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, respectively (Note that both of these alternative hypotheses are 2-sided.);
- (vi) a ‘star rating’ of these p -values.

The **Residual standard error** is s , where s^2 is defined in (12.17).

The quantity **Multiple R-Squared** is the squared (sample) correlation coefficient r^2 between x and Y . It is defined by

$$\begin{aligned} r^2 &= \frac{S_{XY}^2}{S_{XX}S_{YY}} \\ &= \frac{\left\{ \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right\}^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}. \end{aligned} \quad (12.28)$$

It can be shown that

$$r^2 \leq 1. \quad (12.29)$$

An important interpretation of r^2 is as the proportion of the variation in y_1, \dots, y_n which is explained by the model. The closer r^2 is to 1, the better the model explains the variation in the data.

In R, the fitted values $\hat{y}_1, \dots, \hat{y}_n$ and the residuals r_1, \dots, r_n can be obtained using `residuals` and `fitted`, respectively, e.g. for the chlorine data

```
> swimfit<-fitted(swim)
> swimfit
      1      2      3      4      5      6
1.7285714 1.5571429 1.3857143 1.2142857 1.0428571 0.8714286
> swimres<-residuals(swim)
> swimres
      1      2      3      4      5      6
0.07142857 -0.05714286 0.01428571 -0.11428571 0.05714286 0.02857143
```

obtains and displays the fitted values and then the residuals.

Example:

In the chlorine example, we can use R to obtain 95% confidence intervals for α and β based on (12.23)–(12.24) as follows.

```
> summary(swim)
```

Call:

```
lm(formula = chlorine ~ time)
```

Residuals:

```
      1      2      3      4      5      6
0.07143 -0.05714 0.01429 -0.11429 0.05714 0.02857
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.900000    0.074642  25.455 1.41e-05 ***
time         -0.085714    0.009583  -8.944 0.000864 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.08018 on 4 degrees of freedom

Multiple R-Squared: 0.9524, Adjusted R-squared: 0.9405

F-statistic: 80 on 1 and 4 DF, p-value: 0.0008642

```
> alphalimits <-c(1.900000+qt(0.025,4)*0.074642,1.900000-qt(0.025,4)*0.074642)
```

```
> alphalimits
```

```
[1] 1.692761 2.107239
```

```
> betalimits <-c(-0.085714+qt(0.025,4)*0.009583,-0.085714-qt(0.025,4)*0.009583)
```

```
> betalimits
```

```
[1] -0.11232067 -0.05910733
```

giving 95% confidence intervals of (1.693, 2.107) for α and (−0.1123, −0.0591) for β .

12.2.1 Regression through the origin

Sometimes we may wish to assume that the regression line passes through the origin, i.e. that $\alpha = 0$ in (12.3). The way to do this in the R `lm` command is to use `-1` to show that the constant term (i.e. the coefficient of 1) should be removed. Here is an illustration using the chlorine data (where it would not be sensible to fit a line through the origin!).

```
> swimnoc<-lm(chlorine~time-1)
> swimnoc
```

Call:

```
lm(formula = chlorine ~ time - 1)
```

Coefficients:

```
time
0.1335
```

(Note that, in this case, insisting that $\alpha = 0$ has changed $\hat{\beta}$ considerably.)

12.3 Confidence intervals and prediction intervals

One of the main purposes in fitting a regression line (of Y on x , say) to data $(x_1, y_1), \dots, (x_n, y_n)$ is to be able to say something about the values Y_0 of the response variable corresponding to any given value x_0 of the predictor variable. Note that x_0 need not be one of x_1, \dots, x_n . You may find it helpful to think of y_1, \dots, y_n as observations taken in the *past*, which we use (in the present) to fit the sample regression line

$$y = \hat{\alpha} + \hat{\beta}x, \quad (12.30)$$

whereas Y_0 is the random variable of *future* observations of Y with $x = x_0$.

There are 2 types of interval which are of interest:

- (i) confidence intervals for $E(Y_0)$,
- (ii) prediction intervals for Y_0 .

(i) **Confidence intervals for $E(Y_0)$:**

The regression equation (12.1) gives

$$E(Y_0) = \alpha + \beta x_0,$$

so that $E(Y_0)$ is *fixed* and depends on the unknown parameters α and β . It is sensible to estimate $E(Y_0)$ by

$$\hat{E}(Y_0) = \hat{\alpha} + \hat{\beta}x_0.$$

From the distributions of $\hat{\alpha}$ and $\hat{\beta}$ given in (12.15)–(12.16) (together with careful consideration of the covariance of $\hat{\alpha}$ and $\hat{\beta}$), it can be shown that

$$\frac{\hat{E}(Y_0) - E(Y_0)}{\sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}} \sim t_{n-2},$$

so that, e.g. a 95% confidence interval for $E(Y_0)$ is

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)}.$$

You do not need to memorise this. However, you should note its key feature:

confidence intervals for $E(Y_0)$ become wider as x_0 moves away from \bar{x} .

Thus estimates of $E(Y_0)$ become less reliable, the further x_0 is from \bar{x} . This is one reason why it is unwise to extrapolate Y outside the range of x_1, \dots, x_n .

(ii) **Prediction intervals for Y_0 :**

Whereas a confidence interval for $E(Y_0)$ considered above is a (random) interval which we would like to contain the (fixed) mean of Y_0 , a prediction interval for Y_0 is a (random) interval which we would like to contain the (random) value of a (future) observation y_0 of Y_0 . A prediction interval for Y_0 is centred on the fitted value $\hat{\alpha} + \hat{\beta}x_0$ corresponding to x_0 . Consider the difference

$$Y_0 - (\hat{\alpha} + \hat{\beta}x_0)$$

between Y_0 and the fitted value. Since the ('future') random variable Y_0 is independent of the ('past') random variables Y_1, \dots, Y_n which are used to fit the model, Y_0 is independent of $\hat{\alpha} + \hat{\beta}x_0$, and so

$$\begin{aligned} \text{var}\left(Y_0 - [\hat{\alpha} + \hat{\beta}x_0]\right) &= \text{var}(Y_0) + \text{var}(\hat{\alpha} + \hat{\beta}x_0) & (12.31) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right). \end{aligned}$$

It follows (after some calculation) that a 95% prediction interval for Y_0 is

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2;0.025} \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)}.$$

You do not need to memorise this. However, you should note its key features:

- (a) prediction intervals for Y_0 become wider as x_0 moves away from \bar{x} ,
- (b) prediction intervals for Y_0 are wider than the corresponding confidence intervals for $E(Y_0)$.

The intuitive explanation of (b) is that prediction intervals take into account the variability of the 'future' observation Y_0 , whereas confidence intervals are concerned only with the mean value $E(Y_0)$. This is made rigorous in (12.31).

Confidence intervals and prediction intervals can be produced in R using the `predict.lm` command. For example, the following portion of an R session used the chlorine data to give such intervals for the chlorine residual at the times (2,4,6,8,10,12 hours after chlorination) in the data set. (Recall that `swim` contains the results of the linear regression.)

```

> predict.lm(swim,interval=c("confidence"))
      fit      lwr      upr
1 1.7285714 1.5674575 1.889685
2 1.5571429 1.4361854 1.678100
3 1.3857143 1.2910190 1.480410
4 1.2142857 1.1195904 1.308981
5 1.0428571 0.9218997 1.163815
6 0.8714286 0.7103147 1.032542
> predict.lm(swim,interval=c("prediction"))
      fit      lwr      upr
1 1.7285714 1.4537746 2.003368
2 1.5571429 1.3037927 1.810493
3 1.3857143 1.1437994 1.627629
4 1.2142857 0.9723709 1.456201
5 1.0428571 0.7895070 1.296207
6 0.8714286 0.5966318 1.146225

```

To obtain confidence intervals and prediction intervals at other times (e.g. 9 and 11 hours after chlorination), we first put these times into a dataframe (e.g. `newswim`).

```

> newswim<-data.frame(time=c(9,11))
> predict.lm(swim,newswim,interval=c("confidence"))
      fit      lwr      upr
1 1.1285714 1.023258 1.233885
2 0.9571429 0.817192 1.097094
> newswim<-data.frame(time=c(9,11))
> predict.lm(swim,newswim,interval=c("prediction"))
      fit      lwr      upr
1 1.1285714 0.8823061 1.374837
2 0.9571429 0.6941945 1.220091

```

12.4 Checking the assumptions

The linear regression model (12.14) is

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, \dots, n \quad \text{with } Y_1, \dots, Y_n \text{ independent.}$$

Thus when fitting the linear regression model (12.14), we are making the assumptions

- (i) Y_1, \dots, Y_n are independent,
- (ii) Y_1, \dots, Y_n are normally distributed,
- (iii) $E(Y_i) = \alpha + \beta x_i$, i.e. the mean of Y_i is a linear function of x_i ,
- (iv) Y_1, \dots, Y_n have the *same* variance.

It is important to check that these assumptions hold.

Before fitting the linear regression model, it is sensible to plot the data.

Example (Chemical data):

The following measurements give the masses of desired product produced (from 100 g of reagent) by a certain chemical reaction at various temperatures.

| | | | | | | |
|-------------------|----|----|----|----|----|----|
| mass (g) | 40 | 32 | 44 | 36 | 59 | 61 |
| temperature (° C) | 0 | 10 | 20 | 30 | 40 | 50 |

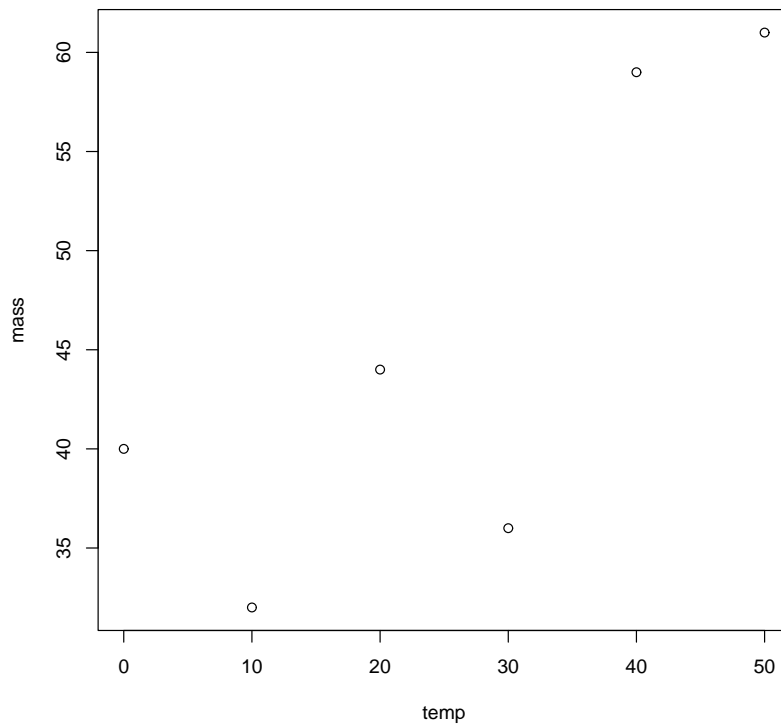
These measurements have been placed in the object `chemdata` in R.

```
> chemdata
      [,1] [,2] [,3] [,4] [,5] [,6]
temp    0  10  20  30  40  50
mass   40  32  44  36  59  61
```

The R command

```
> plot(temp,mass,xlab="temp",ylab="mass")
```

gives



Since the plot is roughly linear, it is sensible to fit the linear regression model to these data.

Once the model has been fitted to the data, it is sensible to check that assumptions (i)–(iv) above hold. The key to such checking lies in the residuals r_1, \dots, r_n , which were defined in (12.27) by

$$r_i = y_i - \hat{y}_i, \quad i = 1, \dots, n,$$

and measure the discrepancy between the model and the data.

Various residual plots should be examined to check the model assumptions of independence and equal variance. Applying the R command `plot` to the object created by `lm` produces 4 plots. The first 3 are particularly useful. They are

- (i) a plot (labelled ‘Residuals vs Fitted’ in R) of residuals r_1, \dots, r_n against fitted values $\hat{y}_1, \dots, \hat{y}_n$;
- (ii) a Q-Q plot (labelled ‘Normal Q-Q plot’ in R) of the standardised residuals r_1^*, \dots, r_n^* ;
- (iii) a plot (labelled ‘Scale-Location plot’ in R) of $\sqrt{|r_1^*|}, \dots, \sqrt{|r_n^*|}$ against fitted values $\hat{y}_1, \dots, \hat{y}_n$.

Here

$$r_i^* = \frac{r_i}{\sqrt{\widehat{\text{var}}(r_i)}}$$

is the i th *standardised residual*, which is obtained by scaling r_i to have variance 1. If the assumptions hold then

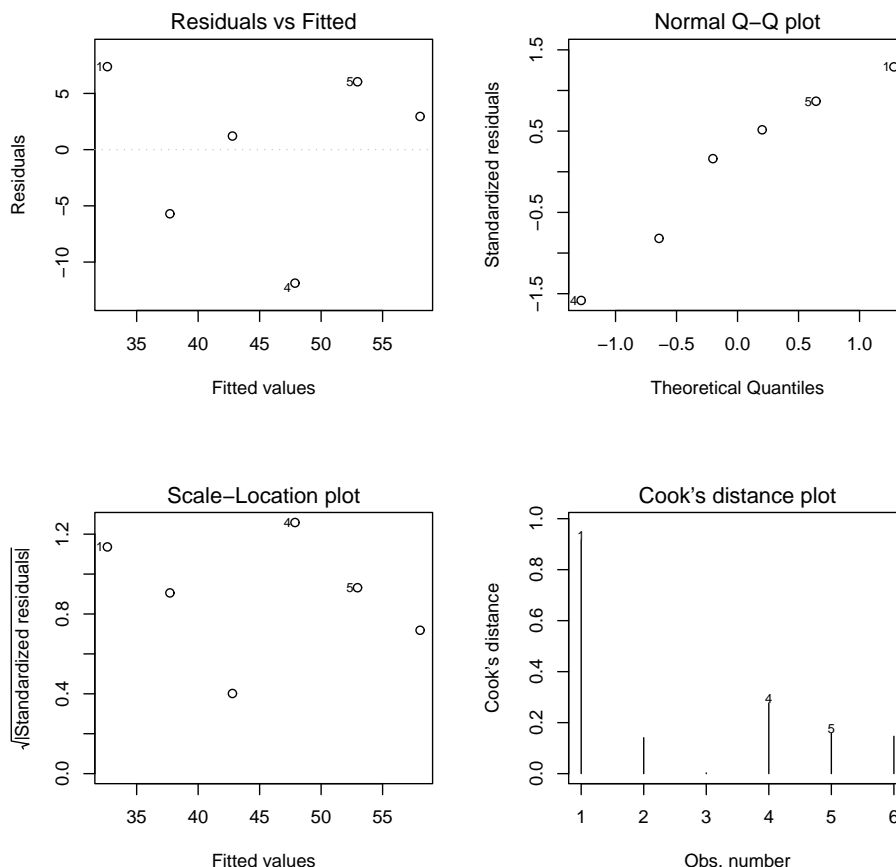
- (i) the plot of residuals against fitted values should show no obvious systematic pattern — the residuals should be well scattered above and below zero, with vertical spread (indicating variance) which does not depend much on the fitted value;
- (ii) the Q-Q plot of the standardised residuals should be approximately linear (reflecting the fact that r_1, \dots, r_n should be *approximately* like a random sample from some normal distribution with mean zero);
- (iii) the plot of $\sqrt{|r_1^*|}, \dots, \sqrt{|r_n^*|}$ against fitted values $\hat{y}_1, \dots, \hat{y}_n$ should show no trend.

Example (Chemical data):

The R commands

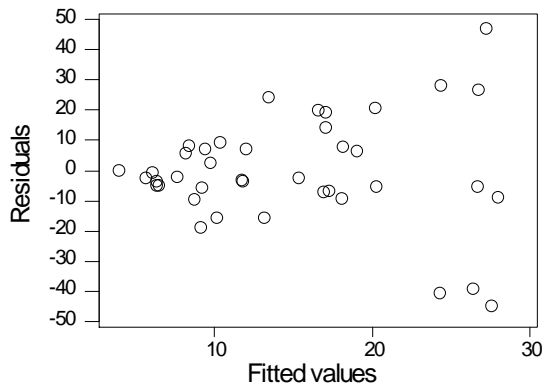
```
> chem<-lm(mass~temp)
> par(mfrow=c(2,2))
> plot(chem)
```

give

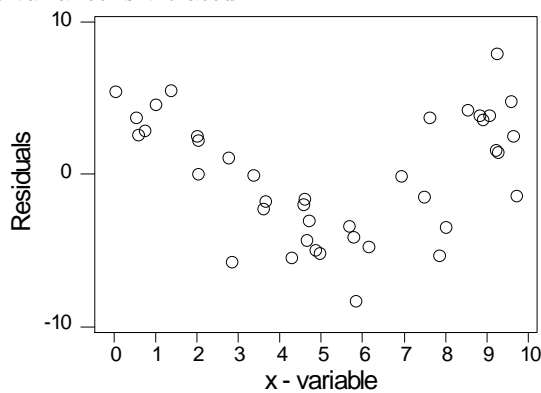


None of these plots gives grounds for doubting that the assumptions of the linear regression model hold.

Here are two examples of problematic plots:



This shows a clear relationship between variance and fitted values, indicating that the assumption (numbered (iv) above) of constant variance is violated.



This plot of residuals against the explanatory variable x shows a clear systematic pattern. This indicates that the assumptions of linearity and of constant variance (numbered (iii) and (iv) above) are not valid for this data set. The plot is reminiscent of the graph of a quadratic function. This suggests that the data come from a quadratic model

$$E(Y_i) = \alpha + \beta_1 x_i + \beta_2 x_i^2, \quad i = 1, \dots, n.$$

12.4.1 Transformation

If residual plots indicate that the linear regression model is not appropriate then sometimes transformation of the response variable Y or the explanatory variable x can linearise the relationship between Y and x , so that the linear regression model can be fitted to the transformed data. Three common transformations which it is worth trying are square root ($t \mapsto \sqrt{t}$ for $t \geq 0$), exponential ($t \mapsto \exp t$) and logarithm ($t \mapsto \log t$ for $t > 0$).

Example (Log transform):

If

$$y = Ax^\beta$$

then taking logarithms gives

$$\log y = \alpha + \beta \log x,$$

where $\alpha = \log A$, so that $\log y$ is a linear function of $\log x$.

Example:

If

$$y = \frac{\alpha}{1 + \beta x}$$

then we can re-write this as

$$\frac{\alpha}{y} = 1 + \beta x,$$

so that

$$\frac{1}{y} = \frac{1}{\alpha} + \frac{\beta}{\alpha}x,$$

which is of the linear form (12.1).

The choice of transformation is a matter of experience and experiment. The scatter plot of $(x_1, y_1), \dots, (x_n, y_n)$ and the residual plots may suggest a suitable transformation.

Note that transformation of the response variable has an effect on the validity of the assumption of normality, e.g. if Y is normally distributed then $\log Y$ is not.) As a *very rough* guide,

- (i) if the residual plots indicate normality with constant variance then transform the explanatory variable,
- (ii) if the residual plots do not indicate normality with constant variance then transform the response variable.

12.5 Multiple regression

So far, we have considered the relationship between the response variable and a *single* explanatory variable. In many cases, we wish to relate the response variable (Y , say) to several explanatory variables $(x_1, \dots, x_k$, say – note that the subscript now indicates which of the explanatory variables we are considering, not the value taken by a single explanatory variable). This can be done by straightforward generalisations of (12.1) and (12.14).

The appropriate generalisation of the linear regression model (12.1) is the multiple linear regression model

$$E(Y) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k. \quad (12.32)$$

The least-squares estimates $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$ of $\alpha, \beta_1, \dots, \beta_k$ are defined as the values which minimise

$$S(\alpha, \beta_1, \dots, \beta_k) = \sum_{i=1}^n \{y_i - (\alpha + \beta_1 x_{1i} + \dots + \beta_k x_{ki})\}^2,$$

where y_i is the response at the value (x_{1i}, \dots, x_{ki}) of the explanatory variables (x_1, \dots, x_k) . The estimates $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$ can be found by matrix algebra. (Details are given in the Honours module *Generalized Linear Models and Data Analysis*.)

The appropriate generalization of (12.14) is

$$Y_i \sim N(\alpha + \beta_1 x_{1i} + \dots + \beta_k x_{ki}, \sigma^2), \quad i = 1, \dots, n \quad \text{with } Y_1, \dots, Y_n \text{ independent.} \quad (12.33)$$

If (12.33) holds then the maximum likelihood estimates of $\alpha, \beta_1, \dots, \beta_k$ are the least-squares estimates $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$.

The statistic s^2 , defined by

$$s^2 = \frac{1}{n - (k + 1)} S(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k),$$

is an unbiased estimator of σ^2 . In **R**, $s = \sqrt{s^2}$ is called the *residual standard error*.

The **R** command `lm` can be used for multiple linear regression. The various predictor variables are joined by `+`, as shown in the following example

Example (Peruvian blood pressure data):

In a study of the effects of altitude on blood pressure, various measurements were made on 39 indigenous Peruvian men who had been born in the Andes and had migrated to parts of Peru at low altitude. The variables which are relevant here are

- (i) `systolic` = systolic blood pressure,

- (ii) `age` = age (in years),
- (iii) `years` = number of years since migration to low altitude,
- (iv) `weight` = weight.

The investigators had a feeling that systolic blood pressure might be related to age, the fraction of his lifetime for which such a man had been at low altitude, weight. Accordingly, the following R session began by calculating `fraction` as `years/age` and then regressing `systolic` on the 2 explanatory variables `fraction` and `weight`.

```
> fraction<-years/age
> peru<-lm(systolic~fraction+weight)
```

The summary is

```
> summary(peru)
```

Call:

```
lm(formula = systolic ~ fraction + weight)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|---------|--------|--------|---------|
| -18.4330 | -7.3070 | 0.8963 | 5.7275 | 23.9819 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-----------------------|----------|------------|---------|--------------|
| (Intercept) | 60.8959 | 14.2809 | 4.264 | 0.000138 *** |
| <code>fraction</code> | -26.7672 | 7.2178 | -3.708 | 0.000699 *** |
| <code>weight</code> | 1.2169 | 0.2337 | 5.207 | 7.97e-06 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.777 on 36 degrees of freedom

Multiple R-Squared: 0.4731, Adjusted R-squared: 0.4438

F-statistic: 16.16 on 2 and 36 DF, p-value: 9.795e-06

The highly significant p -values for `fraction` and `weight` indicate that there is very strong evidence against the null hypotheses that the regression coefficients of `fraction` and `weight` are zero, i.e. we can conclude that systolic blood pressure definitely depends on fraction of life lived at low altitude and on weight.

12.5.1 Polynomial regression

Multiple regression can be used to fit polynomial regression models of the form

$$Y_i \sim N(\alpha + \beta_1 x_i + \dots + \beta_k x_i^k, \sigma^2), \quad i = 1, \dots, n \quad \text{with } Y_1, \dots, Y_n \text{ independent.} \quad (12.34)$$

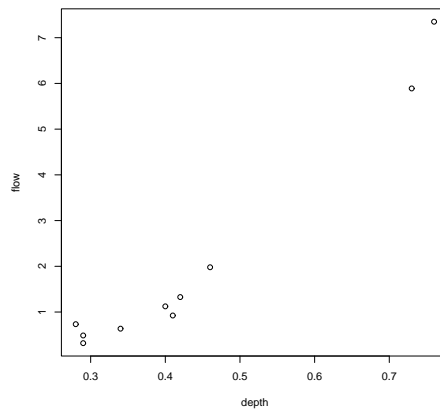
Model (12.34) is just the multiple regression model (12.32) with the powers x, x^2, \dots, x^k of x as the explanatory variables.

Example (Stream data):

Hydrologists were interested in the way in which the rate of flow in a stream varies with the depth at which the flow is measured. The following R session is a partial analysis of a data set consisting of 10 readings on flow rate (`flow`) and depth (`depth`).

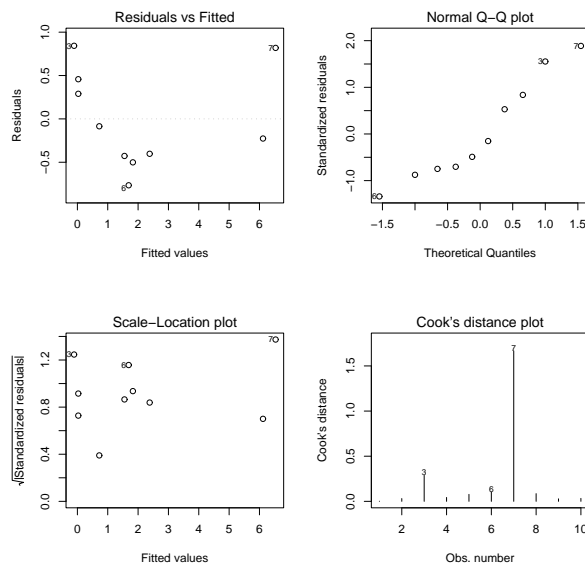
```
> plot(depth, flow, xlab="depth", ylab="flow")
```

produced the scatter plot on the next page



This looks fairly linear, so a linear regression model of **flow** on **depth** was fitted and residual plots were produced.

```
> stream<-lm(flow~depth)
> par(mfrow=c(2,2))
> plot(stream)
```



The residual plot suggests that a quadratic of the form

$$E(Y) = \alpha + \beta_1 x + \beta_2 x^2 \tag{12.35}$$

(where Y denotes **flow** and x denotes **depth**) would be more appropriate than the simple linear model

$$E(Y) = \alpha + \beta x.$$

The following piece of R code fits the quadratic model (12.35).

```
> depthsq<-depth^2
> stream2<-lm(flow~depth+depthsq)
```

```

> summary(stream2)

Call:
lm(formula = flow ~ depth + depthsq)

Residuals:
    Min       1Q   Median       3Q      Max
-0.406145 -0.163666 -0.002649  0.198973  0.327658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.683      1.059   1.589  0.1561
depth        -10.861      4.517  -2.404  0.0472 *
depthsq       23.535      4.274   5.506  0.0009 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2794 on 7 degrees of freedom
Multiple R-Squared:  0.99, Adjusted R-squared:  0.9871
F-statistic: 346.5 on 2 and 7 DF,  p-value: 1e-07

```

Note that the p -value of 0.1561 corresponding to the intercept is not significant at the 15% level, indicating that we have no reason to doubt the hypothesis $H_0 : \alpha = 0$ in (12.35). It is therefore sensible to fit the model

$$E(Y) = \beta_1 x + \beta_2 x^2, \quad (12.36)$$

which can be done by

```

> stream3<-lm(flow~depth+depthsq-1)
> summary(stream3)

Call:
lm(formula = flow ~ depth + depthsq - 1)

Residuals:
    Min       1Q   Median       3Q      Max
-0.39946 -0.08639 -0.03278  0.14220  0.45582

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
depth        -3.7492      0.6613  -5.669 0.000471 ***
depthsq     16.9382      1.1071  15.299 3.31e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 8 degrees of freedom
Multiple R-Squared:  0.9924, Adjusted R-squared:  0.9905
F-statistic:  522 on 2 and 8 DF,  p-value: 3.343e-09

```

12.5.2 Comparison of models

A common problem in multiple regression is that of testing that certain specified regression coefficients are zero (representing the intuitive idea that the corresponding explanatory variables ‘have no effect’). For example, in the above example on the stream data we tested the null hypothesis $H_0 : \alpha = 0$ in (12.35) against a 2-sided alternative.

In the above example, R calculated the p -value of 0.1561 by using a t -test. It is tempting to think that, in general, when testing that several regression coefficients are zero, we should use a sequence of t -tests, to test that each of these coefficients in turn is zero. However, this makes the calculation of p -values very complicated. It is better to use ANOVA (ANalysis Of VAriance). ANOVA is a general way of testing hypotheses which are formulated as nested linear models, using a test statistic based on the difference in the goodness-of-fit of the alternative models. ‘Nested’ means that the model corresponding to the null hypothesis is a special case of the model corresponding to the alternative hypothesis, being obtained from it by placing (linear) restrictions on the parameters.

In the context of multiple regression, we have a multiple regression model

$$Y_i \sim N(\beta_1 + \beta_2 x_{2i} + \dots + \beta_{p_1} x_{p_1 i}, \sigma^2), \quad i = 1, \dots, n \quad \text{with } Y_1, \dots, Y_n \text{ independent,} \quad (12.37)$$

and we are interested in a submodel in which $p_1 - p_0$ of the regression coefficients $\beta_1, \dots, \beta_{p_1}$ are zero. Thus we wish to test the null hypothesis

$$H_0 : \text{the specified regression coefficients are zero}$$

against the alternative hypothesis

$$H_1 : \text{there is no restriction on the specified regression coefficients.}$$

The intuitive idea is to see whether or not the *full model* (12.37) gives a *significantly* better fit to the data than the submodel (specified by H_0) does. Of course, the full model always fits the data a little more closely than the submodel, since it has more parameters. However, if H_0 is true then the difference in fit between the models should be quite small, while a big difference in fit would tend to suggest that H_0 is false.

The goodness-of-fit of either model to the data is measured by the *residual sum of squares* (rss), which is the sum of squares of differences between the model and data, e.g. for the full model, the residual sum of squares is

$$\text{rss}_1 = \sum_{i=1}^n \left\{ y_i - \left(\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_{p_1} x_{p_1 i} \right) \right\}^2. \quad (12.38)$$

The residual sum of squares rss_0 for the submodel is defined similarly, but using only the least squares estimates (under H_0) of the regression coefficients used in the submodel.

Calculations similar to those used to obtain the one-sample t -test (10.5) show that, under either model,

$$\frac{\text{rss}_1}{\sigma^2} \sim \chi_{n-p_1}^2,$$

but only under H_0 do we have

$$\begin{aligned} \frac{\text{rss}_0}{\sigma^2} &\sim \chi_{n-p_0}^2, \\ \frac{\text{rss}_0 - \text{rss}_1}{\sigma^2} &\sim \chi_{p_1-p_0}^2, \end{aligned}$$

$$\text{rss}_0 - \text{rss}_1 \text{ is independent of } \text{rss}_1.$$

Hence

$$\frac{(\text{rss}_0 - \text{rss}_1)/(p_1 - p_0)}{\text{rss}_1/(n - p_1)} \sim F_{p_1-p_0, n-p_1} \quad \text{under } H_0. \quad (12.39)$$

If H_0 is false then this statistic will tend to be too large for consistency with $F_{p_1-p_0, n-p_1}$. This gives a test of H_0 against H_1 .

Example (Stream data revisited):

Here $n = 10$. The full model is (12.35), i.e.

$$E(Y) = \alpha + \beta_1 x + \beta_2 x^2$$

(where Y denotes flow and x denotes depth). Thus $p_1 = 3$.

(i) If we take the null hypothesis to be

$$H_0 : \alpha = 0$$

then the submodel is (12.36). In this case $p_0 = 2$ and the value of the statistic (12.39) is $2.524921 = 1.589^2$, where 1.589 is the t -value given for (Intercept) in `stream2`. In the output for `stream2`, 1.589 is compared with t_7 ; according to (12.39), 1.589^2 is compared with $F_{1,7} = t_7^2$. As the p -value is 0.1561, we accept H_0 .

(ii) If we take the null hypothesis to be

$$H_0 : \beta_1 = \beta_2 = 0$$

then the submodel is

$$E(Y) = \alpha,$$

so that the mean of Y does not depend on x . In this case $p_0 = 1$. The last line of `summary(stream2)` is

`F-statistic: 346.5 on 2 and 7 DF, p-value: 1e-07`

meaning that the value of the statistic (12.39) is 346.5. This is to be compared with $F_{2,7}$. As the p -value is so small, we reject H_0 .

Remark The name *Analysis of Variance* comes from the fact that (12.39) is based on the decomposition

$$\frac{\text{rss}_0}{\sigma^2} = \frac{\text{rss}_1}{\sigma^2} + \frac{\text{rss}_0 - \text{rss}_1}{\sigma^2}, \quad (12.40)$$

which splits up the (scaled) variability rss_0/σ^2 of the data about the submodel into the sum of the (scaled) variability rss_1/σ^2 of the data about the full model and the (scaled) difference $(\text{rss}_0 - \text{rss}_1)/\sigma^2$ between the two models. From the algebraic/geometric point of view, (12.40) can be regarded as an example of Pythagoras's Theorem. Under H_0 , the distributions of the terms in (12.40) are the corresponding terms in the decomposition

$$\chi_{n-p_0}^2 \sim \chi_{n-p_1}^2 + \chi_{p_1-p_0}^2. \quad (12.41)$$

13 Analysis of variance

13.1 One-way ANOVA

A common problem is that of comparing several populations. The usual technique for doing this is one-way analysis of variance (which is a special example of the ANOVA introduced in section 12.5.2).

Consider k distributions (or populations) with means μ_1, \dots, μ_k , and suppose that we wish to test

$$H_0 : \mu_1 = \dots = \mu_k$$

against

$$H_1 : \mu_1, \dots, \mu_k \text{ are not all equal.}$$

Note

The alternative hypothesis is *not* ' $H_1 : \mu_1, \dots, \mu_k$ are all different'. Thus, if $k = 3$ and $\mu_1 = \mu_2 \neq \mu_3$ then H_1 would hold.

Suppose that we have random samples of sizes n_1, \dots, n_k , respectively, from the k distributions. Let y_{ij} denote the j th observation on the i th distribution, for $i = 1, \dots, k$ and $j = 1, \dots, n_i$. We think of y_{ij} as an observation on a random variable Y_{ij} .

We shall assume that

$$Y_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, \dots, k, \quad j = 1, \dots, n_i \quad \text{with the } Y_{ij} \text{ independent.} \quad (13.1)$$

With its assumptions of independent normal distributions with the same variance, this model is reminiscent of the simple linear regression model (12.14). We now show that it is a special case of the multiple regression model (12.32).

Define indicator variables x_1, \dots, x_k by

$$x_i = \begin{cases} 1 & \text{if the observation is from } i\text{th distribution,} \\ 0 & \text{otherwise.} \end{cases}$$

Then the condition

$$E(Y_{ij}) = \mu_i$$

on the mean of Y_{ij} can be written as

$$E(Y_{ij}) = \mu_1 x_1 + \dots + \mu_k x_k, \quad (13.2)$$

which is the multiple regression model (12.32) with no constant term α . We can now apply the general results of Section 12.5.2. The full model (13.2) has k parameters, i.e. $p_1 = k$. The submodel given by H_0 is

$$E(Y_{ij}) = \mu, \quad (13.3)$$

and so has one parameter, i.e. $p_0 = 1$. There are n observations, with

$$n = n_1 + \dots + n_k.$$

Then (12.39) gives

$$\frac{(\text{rss}_0 - \text{rss}_1)/(k-1)}{\text{rss}_1/(n-k)} \sim F_{k-1, n-k} \quad \text{under } H_0, \quad (13.4)$$

where rss_0 and rss_1 denote the residual sums of squares under the null model (13.3) and the alternative (full) model (13.2), respectively. If H_0 is false then the above statistic will tend to be too large for consistency with $F_{k-1, n-k}$. Thus large values of the statistic lead to rejection of H_0 .

In one-way ANOVA, the residual sums of squares rss_0 and rss_1 have nice expressions, which we now derive. Let \bar{y} and \bar{y}_i (for $i = 1, \dots, k$) denote the overall sample mean (of all n observations) and the sample mean of the i th sample, respectively.

(i) The *total sum of squares* is

$$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

and represents the variability of all the observations;

(ii) The *between sum of squares* is

$$SS_B = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y})^2$$

and represents the variability between the k sample means;

(iii) The *within sum of squares* is

$$SS_W = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

and represents the variability of all the observations within the k groups.

Calculation shows that

$$SS_T = SS_B + SS_W, \quad (13.5)$$

which is the basic decomposition in one-way ANOVA.

Further calculation shows that the maximum likelihood estimates (which are also the least squares estimates) $\hat{\mu}_1, \dots, \hat{\mu}_k$ and $\hat{\mu}$ of μ_1, \dots, μ_k and μ in (13.2) and (13.3), respectively, are

$$\begin{aligned} \hat{\mu}_i &= \bar{y}_{i.} & i = 1, \dots, k \\ \hat{\mu} &= \bar{y} \end{aligned}$$

and that

$$\begin{aligned} \text{rss}_0 &= SS_T \\ \text{rss}_1 &= SS_W \\ \text{rss}_0 - \text{rss}_1 &= SS_B. \end{aligned}$$

Thus the left hand side of (13.4) can be expressed in terms of SS_B and SS_W . It is useful to express the numerator and denominator of the left hand side of (13.4) as *mean squares*.

(i) The *between mean square* is

$$MS_B = \frac{SS_B}{k-1};$$

(ii) The *within mean square* is

$$MS_W = \frac{SS_W}{n-k}.$$

Then the left hand side of (13.4) is

$$F = \frac{MS_B}{MS_W} \quad (13.6)$$

and

$$F \sim F_{k-1, n-k} \quad \text{under } H_0. \quad (13.7)$$

Many statistical packages (including R) set out the sums of squares, mean squares, etc. in an *ANOVA table*.

| Source | d.f. | SS | MS | F | p |
|---------|---------|--------|--------|-----|-----|
| Between | $k - 1$ | SS_B | MS_B | F | p |
| Within | $n - k$ | SS_W | MS_W | | |
| Total | $n - 1$ | SS_T | | | |

where F is given by (13.6) and p denotes the p -value of F .

An important property of the within mean square MS_W is that it is an unbiased estimator of the variance σ^2 in (13.1). The *residual mean square* is

$$s^2 = MS_W, \tag{13.8}$$

and s is often called the *residual standard error*.

13.1.1 One-way ANOVA in R

One-way ANOVA can be carried out in R, using the `anova(lm())` command. In order to use this, we have first to put the data into a dataframe and then to indicate within the `anova(lm())` command that the ‘explanatory variable’ is a *factor*, i.e. a discrete variable which just tells R which group (i.e. distribution) each observation comes from.

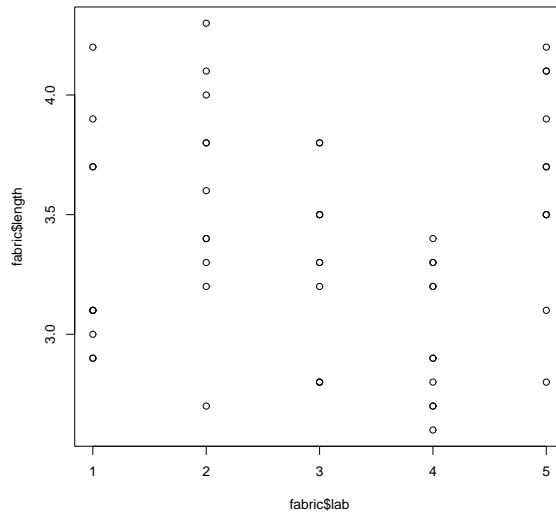
Example (Fabric data):

A standard measurement of the flammability of fabric is given by the length of the burnt portion of a piece of the fabric which has been held over a flame for a given time. An investigation to see whether or not there was a difference between the measurements obtained by 5 laboratories obtained the following data.

| | Laboratory | | | | |
|-----|------------|-----|-----|-----|---|
| | 1 | 2 | 3 | 4 | 5 |
| 2.9 | 2.7 | 3.3 | 3.3 | 4.1 | |
| 3.1 | 3.4 | 3.3 | 3.2 | 4.1 | |
| 3.1 | 3.6 | 3.5 | 3.4 | 3.7 | |
| 3.7 | 3.2 | 3.5 | 2.7 | 4.2 | |
| 3.1 | 4.0 | 2.8 | 2.7 | 3.1 | |
| 4.2 | 4.1 | 2.8 | 3.3 | 3.5 | |
| 3.7 | 3.8 | 3.2 | 2.9 | 2.8 | |
| 3.9 | 3.8 | 2.8 | 3.2 | 3.5 | |
| 3.1 | 4.3 | 3.8 | 2.9 | 3.7 | |
| 3.0 | 3.4 | 3.5 | 2.6 | 3.5 | |
| 2.9 | 3.3 | 3.8 | 2.8 | 3.9 | |

The following R code entered the data into the dataframe `fabric`, with `lab` as a vector to indicate the laboratory used and with `length` as a vector of the values of the response (which we can think of as obtained by reading the columns of the above table one after another), and then plotted `length` against `lab`.

```
> lab1<-c(2.9,3.1,3.1,3.7,3.1,4.2,3.7,3.9,3.1,3.0,2.9)
> lab2<-c(2.7,3.4,3.6,3.2,4.0,4.1,3.8,3.8,4.3,3.4,3.3)
> lab3<-c(3.3,3.3,3.5,3.5,2.8,2.8,3.2,2.8,3.8,3.5,3.8)
> lab4<-c(3.3,3.2,3.4,2.7,2.7,3.3,2.9,3.2,2.9,2.6,2.8)
> lab5<-c(4.1,4.1,3.7,4.2,3.1,3.5,2.8,3.5,3.7,3.5,3.9)
> fabric<-data.frame(lab=c(rep(1,11),rep(2,11),rep(3,11),rep(4,11),rep(5,11)),
+ length=c(lab1,lab2,lab3,lab4,lab5))
> plot(fabric$lab,fabric$length)
```



The ANOVA table is obtained by

```
> anova(lm(length~as.factor(lab), data=fabric))
Analysis of Variance Table

Response: length
      Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(lab)  4  2.9865  0.7466  4.5346 0.003337 **
Residuals     50  8.2327  0.1647
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p -value is so small, we can reject the null hypothesis

$$\mu_1 = \dots = \mu_5,$$

where μ_i denotes the mean length of burnt fabric in measurements from laboratory i ($i = 1, \dots, 5$).

It is important to check the assumptions of the ANOVA model (13.1). In particular, it is important to check

- (i) that the observations in each group come from a normal distribution,
- (ii) that the variances are equal.

Such checks can be carried out by

- (i) forming a normal probability plot of the *residuals*

$$y_{ij} = y_{ij} - \bar{y}_{i.},$$

- (ii) inspecting the sample variances.

Example (Fabric data – checking assumptions):

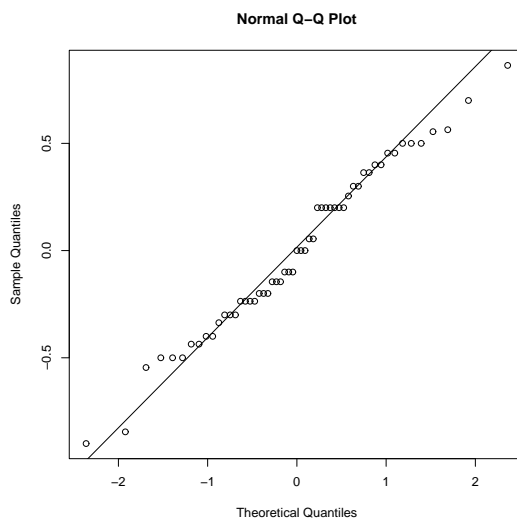
- (i) The R code

```

> resfab1<-lab1-mean(lab1)
> resfab2<-lab2-mean(lab2)
> resfab3<-lab3-mean(lab3)
> resfab4<-lab4-mean(lab4)
> resfab5<-lab5-mean(lab5)
> resfab<-c(resfab1,resfab2,resfab3,resfab4,resfab5)
> qqnorm(resfab)
> qqline(resfab)

```

produces the plot



which gives us no cause to doubt normality.

(ii) The R output

```

> var(lab1)
[1] 0.2045455
> var(lab2)
[1] 0.212
> var(lab3)
[1] 0.138
> var(lab4)
[1] 0.082
> var(lab5)
[1] 0.1867273

```

gives us no cause to doubt equality of the variances.

13.1.2 Least Significant Differences

If the null hypothesis is rejected then the one-way ANOVA has told us only that μ_1, \dots, μ_k are not all equal; it has not indicated which differences between groups are most important.

To do the latter, it is usual to compare the groups in pairs, using t -tests. If $\mu_i = \mu_j$ then we have

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim N(0, 1),$$

s^2 ($= r_{SS1}/(n - k) = MS_W$, as in (13.8)) is independent of $\bar{Y}_i - \bar{Y}_j$, and $s^2(n - k)/\sigma^2 \sim \chi_{n-k}^2$, where $s^2 = MS_W$, so that

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim t_{n-k}.$$

Note the differences between this and the 2-sample t -statistic: here,

- (i) s^2 is obtained from all the data (not just groups i and j),
- (ii) the t -statistic has $n - k$ degrees of freedom (rather than $n_i + n_j - 2$).

If the above t -statistic is significant (i.e. has large absolute value) then we consider that $\mu_i \neq \mu_j$. The significant pairwise comparisons should be used *only as a guide* to which means differ from which, rather than as rigorous tests.

If $n_1 = \dots = n_k$ (i.e. the samples are of equal size) then it is worth calculating out the smallest difference in sample means that would lead to rejection of the null hypothesis that two groups have equal population means. This quantity is known as the *least significant difference (LSD)*. It is then easy to look for the pairs of groups with sample means differing by more than the LSD. If there are k groups, each with m observations (so that $n = mk$) then the LSD for significance level α is

$$t_{n-k;\alpha/2} \sqrt{\frac{2s^2}{m}}.$$

Example (Fabric data – least significant difference):

The R output

```
> lsd<-qt(0.975,50)*sqrt(2*0.1647/11)
> lsd
[1] 0.3475762
> mean(lab5)
[1] 3.645455
> mean(lab2)
[1] 3.6
> mean(lab1)
[1] 3.336364
> mean(lab3)
[1] 3.3
> mean(lab4)
[1] 3
```

suggests that $\mu_4 < \mu_2$ and $\mu_4 < \mu_5$, but does not suggest any other differences between the μ_i .

13.2 Two-way ANOVA

Suppose that each observation belongs to more than one group. For example if each observation belongs to exactly two groups (of two different kinds), then we have a two-way ANOVA. Consider a reading comprehension test given to pupils of ages 9, 10 and 11 from 4 schools (A, B, C and D), yielding the following scores:

| | age 9 | age 10 | age 11 |
|----------|-------|--------|--------|
| School A | 71 | 92 | 89 |
| School B | 44 | 51 | 85 |
| School C | 50 | 64 | 72 |
| School D | 67 | 81 | 86 |

Now if y_i is the score, which as usual we treat as an observation on an independent random variable $Y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i \equiv E(Y_i)$, then there are several models that might be of interest. The most complicated is:

$$\mu_i = \alpha_j + \gamma_k \text{ if } y_i \text{ is from school } j \text{ and age group } k$$

Let's refer to this as model 3. This model states that the expected comprehension score is the sum of a school effect and an age effect. If you think about this model, it has one slight complication, best seen by example. Consider adding 10 to $\alpha_A, \alpha_B, \alpha_C$ and α_D and at the same time subtracting 10 from γ_1, γ_2 and γ_3 . This process would change all the parameters of the model, but would leave all the modelled μ_i s unchanged. Since it is by fitting $\boldsymbol{\mu}$ to \mathbf{y} that the parameters are estimated, this creates a problem: the model as written does not have a unique set of best fit parameters. This problem is easily eliminated: just set one of the parameters to zero (e.g. set $\gamma_3 = 0$), and the problem goes away, while the resulting reduced model can still fit the data exactly as well as the original model.

There are two reductions of this model that might be considered:

$$\mu_i = \alpha_j \text{ if } y_i \text{ is from school } j$$

(model 1, say) and

$$\mu_i = \gamma_k \text{ if } y_i \text{ is from age group } k$$

(model 2, say).

Finally there is the simplest model (model 0):

$$\mu_i = \alpha \text{ for all } i$$

The analysis of variance for comparing these models may be expressed:

| Source | d.f. | SS | MS | F | p |
|--------------------|-----------------|--------|--------|-------|-------|
| Between schools | $J - 1$ | SS_S | MS_S | F_S | p_S |
| Between age groups | $K - 1$ | SS_A | MS_A | F_A | p_A |
| Within | $n - J - K + 1$ | SS_W | MS_W | | |
| Total | $n - 1$ | SS_T | | | |

where $J = 4$ is the number of schools, $K = 3$ is the number of age groups, $F_S = \frac{MS_S}{MS_W}$, $F_A = \frac{MS_A}{MS_W}$, and p_S and p_A denote the corresponding p -values. SS_S is the between sum of squares for the schools, and SS_A is the between sum of squares for the age groups.

13.2.1 Two-way ANOVA in R

The above data can be analysed in R using the following code.

```
> score<-c(71,92,89,44,51,85,50,64,72,67,81,86)
> school=c(rep(1,3),rep(2,3),rep(3,3),rep(4,3))
> age=rep(c(1,2,3),4)
> data.frame(school,age,score)
  school age score
1      1  1    71
2      1  2    92
3      1  3    89
4      2  1    44
5      2  2    51
6      2  3    85
7      3  1    50
8      3  2    64
9      3  3    72
10     4  1    67
11     4  2    81
12     4  3    86
```

The ANOVA table is obtained by

```
> anova(lm(score~as.factor(school)+as.factor(age)))
Analysis of Variance Table

Response: score
          Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(school)  3 1260.00  420.00  6.2069 0.02861 *
as.factor(age)     2 1256.00  628.00  9.2808 0.01458 *
Residuals         6  406.00   67.67
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13.2.2 LSD's for two-way ANOVA

If some effects turn out to be significant in a two-way ANOVA, then it is of some interest to know where the major differences lie. For example if school has an effect, it is a good idea to figure out which schools contribute most to this result. By similar reasoning to that employed in the one-way ANOVA case, Least Significant Differences can be obtained. For example if school was significant in the example, then the LSD (5% level) for a difference in mean score between schools is:

$$t_6(0.025) \times \sqrt{2s^2/3}$$

where s^2 is the residual mean square, MS_W . Similarly the Least Significant Difference in means between age groups would be:

$$t_6(0.025) \times \sqrt{2s^2/4}.$$

So, once you have found that a factor is significant, you can use LSD's to see which pairs of levels of that factor appear to differ. For example, is the difference in mean score between School A and School B greater than the LSD? If so, we have evidence that mean reading comprehension differs between those two schools.

14 Analysing count data

Much of the modelling and analysis that has been covered so far has concerned continuous data. This final section will look briefly at the analysis of discrete data. In particular, we shall look at using binomial distributions as models for data which are *counts* (i.e. numbers $0, 1, 2, \dots$) and consider various tests on these distributions.

14.1 Binomial data

Recall (from Section 2.1.1) that the binomial distribution $\text{bin}(n, p)$ has probability function

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n,$$

where n is a positive integer and $0 \leq p \leq 1$.

The standard example of a random variable X with the $\text{bin}(n, p)$ distribution is the number of successes in n independent trials, each of which has two possible outcomes ('success' and 'failure') with probability p of success.

Many sorts of data can be modelled by a binomial distribution. Examples include

- (i) The total mark X obtained by a student in a multiple choice exam by pure guesswork, e.g. in a 20-question exam where there are 4 alternative answers for each question, $X \sim \text{bin}(20, 0.25)$.
- (ii) The number X of girls in a family of four children. In this case $p = 0.48$ (approximately), so that $X \sim \text{bin}(4, 0.48)$.
- (iii) The number X of bad pixels in a liquid crystal display. Each of the n pixels has a very small probability p of being faulty, and faulty pixels occur independently, so that $X \sim \text{bin}(n, p)$. Typical values of the parameters are $n = 480,000$ and $p = 10^{-6}$.
- (iv) The number X of parasitised caterpillars in a sample of 100 caterpillars. If each caterpillar has probability 0.4 of being parasitised then $X \sim \text{bin}(100, 0.4)$.

Recall (from Section 2.4) that

$$X \sim \text{bin}(n, p) \Rightarrow E(X) = np \text{ and } \text{var}(X) = np(1-p).$$

In Section 8.3, it was shown that

$$X \sim \text{bin}(n, p) \Rightarrow \frac{X - np}{\sqrt{np(1-p)}} \approx N(0, 1) \quad n \rightarrow \infty.$$

As a general guideline, this approximation is usually good if $\min(np, n(1-p)) > 5$ and p is not too far from 0.5. For p closer to 0 or 1, the approximation is reasonable if $\min(np, n(1-p)) > 9$.

Remember that the normal approximation to a binomial distribution approximates a discrete random variable X by a continuous random variable X^* and uses

$$\begin{aligned} P(X = x) &= P(x - 0.5 < X < x + 0.5) \\ &\approx P(x - 0.5 < X^* < x + 0.5) \end{aligned}$$

for integer x .

Examples (Multiple choice test): Consider a student taking a multiple choice test and guessing all the answers. If the test has 20 questions, each with 4 alternative answers, what is the probability of 10 or more correct answers?

Let X be the number of correct answers. Then

$$X \sim$$

so

$$P(X \geq 10) = 0.0139.$$

Now suppose that a test has 25 questions and 5 alternatives to choose from. What is the probability of getting 10 or more correct answers, by pure guesswork? (If $Z \sim N(0, 1)$ then $P(Z \geq 2.25) = 0.01222$.)

14.1.1 Hypothesis testing and confidence intervals

Testing hypotheses about a binomial probability p is straightforward. The binomial random variable X itself is an appropriate test statistic. For example, suppose that we wish to test

$$H_0 : p = 0.25 \text{ against } H_1 : p \neq 0.25$$

in $\text{bin}(18, p)$. Suppose that we observe $x = 3$. The left hand tail probability is

$$P_{H_0}(X \leq 3) = 1 - P_{H_0}(X \geq 4) = 1 - 0.6943 = 0.3057$$

(using `R` or Table 1 of K & Y). Since the alternative hypothesis is two-sided, we should use a 2-tailed test — both large X and small X would count against the null hypothesis. Thus the p -value is $2 \times 0.3057 = 0.6114$.

For larger n , it would be necessary to use a normal approximation (or use the `pbinom` command in `R`). Consider testing

$$H_0 : p = 0.2 \text{ against } H_1 : p \neq 0.2$$

for a binomial random variable X with $X \sim \text{bin}(300, p)$. Under H_0 ,

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\sim}{\sim} N(0, 1). \tag{14.1}$$

Suppose that we observe $x = 40$. Then, using a continuity correction, we have

$$P(X \leq 40) = P(X < 40.5) \simeq P\left(Z < \frac{40.5 - 60}{6.93}\right) = P(Z < -2.81) = 0.00248,$$

where $Z \sim N(0, 1)$. This gives a p -value of $2 \times 0.00248 \simeq 0.005$. There is strong evidence that $p \neq 0.2$.

To find a 95% confidence interval for p , we must find the range of values of p that we would accept when testing at the 5% level. If the normal approximation is appropriate then we can adapt confidence

intervals for a normal distribution with known variance (as considered in Section 9.2.3). We would accept all values of p such that

$$-1.96 < \frac{x - np}{\sqrt{np(1-p)}} < 1.96,$$

i.e.

$$-x - 1.96\sqrt{np(1-p)} < -np < -x + 1.96\sqrt{np(1-p)},$$

which is equivalent to

$$\frac{x}{n} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \frac{x}{n} + 1.96\sqrt{\frac{p(1-p)}{n}}.$$

This is not a confidence interval for p , as p occurs on both sides of both inequalities. There are two ways around this:

(i) solve the quadratic

$$p = \frac{x}{n} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

for p to obtain the endpoints of the confidence interval,

(ii) write $\hat{p} = x/n$ and use the approximate 95% interval

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

For large n and p not too close to 0 or 1, the two approaches give similar results. Otherwise, the quadratic method gives a better approximation.

Example: Under simple Mendelian inheritance, a cross between plants of two particular genotypes is expected to produce progeny 1/4 of which are ‘dwarf’ and 3/4 of which are ‘tall’. In an experiment, a cross resulted in 243 dwarf plants and 682 tall plants. Find a 95% confidence interval for p , the proportion of tall plants in an infinite population of crosses. (Use the second method.)

[The quadratic method yields a 95% confidence interval for p of (0.708, 0.765).]
Hence test at the 5% level $H_0 : p = 3/4$ against the alternative $H_1 : p \neq 3/4$.

14.1.2 Testing binomialness

When using a binomial distribution as a model for data, it is sensible to test how good a model it is, i.e. to assess whether or not the observations are consistent with a binomial distribution. Suppose that X_1, \dots, X_k are independent random variables and that we wish to test

$$\begin{array}{lll} H_0 : X_i \sim \text{bin}(n, p) & i = 1, \dots, k & \text{for some } p \\ & \text{against} & \\ H_1 : X_i \not\sim \text{bin}(n, p) & i = 1, \dots, k & \text{for any } p, \end{array}$$

where n is known but p is unknown. An appropriate test statistic is formed as follows. If we have observations x_1, \dots, x_k on X_1, \dots, X_k then p can be estimated by its maximum likelihood estimate $\hat{p} = \sum_{i=1}^k x_i / (nk) = \bar{x}/n$. If the data really are from a binomial distribution then \hat{p} leads to an estimate

$$\hat{\sigma}^2 = n\hat{p}(1 - \hat{p})$$

of the variance of the X_i . Thus, if H_0 is true then $\hat{\sigma}^2$ will tend to be quite close to the sample variance

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

Therefore an appropriate test statistic is the ratio

$$\frac{\sum_{i=1}^k (x_i - n\hat{p})^2}{n\hat{p}(1 - \hat{p})} \quad (14.2)$$

(where we have used $\bar{x} = n\hat{p}$). If the data come from $\text{bin}(n, p)$ then (14.2) will tend to be close to $k-1$, whereas if they do not then (14.2) will tend to be larger or smaller than $k-1$. The approximate distribution of the test statistic under the null hypothesis is obtainable from the normal approximation to the binomial distribution (which is based on the Central Limit Theorem). If H_0 is true then we can use the approximation

$$X_i \rightsquigarrow N(np, np(1-p)),$$

and so use (10.2) to obtain

$$\frac{1}{\sigma^2} \sum_{i=1}^k (X_i - np)^2 \rightsquigarrow \chi_{k-1}^2.$$

On substituting $\hat{\sigma}^2$ for σ^2 , we have

$$\frac{\sum_{i=1}^k (X_i - n\hat{p})^2}{n\hat{p}(1 - \hat{p})} \rightsquigarrow \chi_{k-1}^2 \quad \text{under } H_0. \quad (14.3)$$

An alternative derivation of (14.3) is

$$\begin{aligned} \frac{X_i - np}{\sqrt{np(1-p)}} &\rightsquigarrow N(0, 1) \\ \Rightarrow \frac{(X_i - np)^2}{np(1-p)} &\rightsquigarrow \chi_1^2 \\ \Rightarrow \frac{\sum_{i=1}^k (X_i - np)^2}{np(1-p)} &\rightsquigarrow \chi_k^2. \end{aligned} \quad (14.4)$$

Then p is replaced by \hat{p} and 'a degree of freedom is lost in the process', giving the result.

Example: We wish to test the null hypothesis that the number of piglets in a litter is a $\text{bin}(8, p)$ random variable. The numbers of piglets born in 32 litters were

1 2 2 3 3 3 3 4 4 4 4 4 4 4 4 4
4 4 4 5 5 5 5 5 5 6 6 6 6 6 7 7.

Let X_i denote the number of piglets in the i th litter (for $i = 1, \dots, 32$). Under the null hypothesis,

$$\frac{\sum_{i=1}^{32} (X_i - 8\hat{p})^2}{8\hat{p}(1 - \hat{p})} \rightsquigarrow \chi_{31}^2.$$

We estimate p by $\hat{p} = 139/256 = \sum_{i=1}^{32} X_i / (8 \times 32)$. (Here $\sum_{i=1}^{32} X_i$ is the total number of piglets, while 8×32 is the maximum possible number of piglets [which occurs if each sow has 8 piglets].) Also,

$$\sum_{i=1}^{32} (x_i - 8\hat{p})^2 = 61.22 \quad \text{and} \quad 8\hat{p}(1 - \hat{p}) = 1.985.$$

Do you accept or reject the null hypothesis? (When calculating the p -value, it may be useful to recall that if $V \sim \chi_\nu^2$ then $E(V) = \nu$.)

14.1.3 Comparison of two binomial distributions

Suppose that X_1 and X_2 are *independent* binomial random variables, with

$$X_1 \sim \text{bin}(n_1, p_1) \quad X_2 \sim \text{bin}(n_2, p_2).$$

From (14.1),

$$X_1 \overset{\sim}{\sim} N(n_1 p_1, n_1 p_1 (1 - p_1)) \quad X_2 \overset{\sim}{\sim} N(n_2 p_2, n_2 p_2 (1 - p_2)),$$

so that

$$\frac{X_1}{n_1} \overset{\sim}{\sim} N\left(p_1, \frac{p_1(1 - p_1)}{n_1}\right) \quad (14.5)$$

$$\frac{X_2}{n_2} \overset{\sim}{\sim} N\left(p_2, \frac{p_2(1 - p_2)}{n_2}\right), \quad (14.6)$$

the approximations being reasonable for $\min(n_i p_i, n_i(1 - p_i)) > 5$. (For simplicity, we do not bother with the continuity correction here.) Since X_1/n_1 and X_2/n_2 are independent,

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \overset{\sim}{\sim} N\left(p_1 - p_2, \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}\right). \quad (14.7)$$

Thus an approximate 95% confidence interval for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}},$$

where

$$\hat{p}_1 = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{x_2}{n_2}$$

are the maximum likelihood estimates of p_1 and p_2 .

Example: In a study of male baldness, 99 of the 310 men who used a proposed treatment experienced new hair growth. Of the 309 men who used another treatment (which was a placebo), 62 experienced new hair growth.

Test (against a 2-sided alternative) the null hypothesis that $p_T = p_P$, where p_T and p_P are the population proportions of those who experience new hair growth among those who use the proposed treatment and the placebo, respectively.

14.2 χ^2 tests

Many tests using count data are based on χ^2 distributions.

14.2.1 Multinomial distributions and the χ^2 goodness-of-fit test

The binomial $\text{bin}(n, p)$ distribution is the distribution of the number X of successes in n independent Bernoulli trials, each of which has two possible outcomes ('success' and 'failure') with probabilities p of 'success' and $1 - p$ of 'failure'. Often we are interested in n independent trials, each of which results in exactly one of k outcomes (or categories), e.g. the responses 'yes', 'no' and 'don't know' to a question in a questionnaire. For $i = 1, \dots, k$, let o_i denote the number of trials which result in outcome i . Then

$$o_1 + \dots + o_k = n. \quad (14.8)$$

The corresponding random variables (O_1, \dots, O_k) are said to have the *multinomial distribution* $\text{multino}(n; p_1, \dots, p_k)$, where p_i is the probability that any given trial results in outcome i . The only properties that we shall need of the multinomial distributions are

$$O_1 + \dots + O_k = n, \quad (14.9)$$

$$E(O_i) = np_i \quad i = 1, \dots, k, \quad (14.10)$$

$$E_1 + \dots + E_k = n, \quad (14.11)$$

where $E_i = E(O_i)$.

Multinomial distributions with $k = 2$ are equivalent to binomial distributions, since

$$Y \sim \text{bin}(n, p) \Leftrightarrow (Y, n - Y) \sim \text{multino}(n; p, 1 - p). \quad (14.12)$$

The tests on binomial distributions considered in Section 14.1 can be generalised readily to tests on multinomial distributions.

Suppose that O_1, \dots, O_k are random counts with $O_1 + \dots + O_k = n$ and that we wish to test

$$\begin{aligned} H_0 : (O_1, \dots, O_k) &\sim \text{multino}(n; p_1, \dots, p_k) \\ &\text{against} \\ H_1 : (O_1, \dots, O_k) &\not\sim \text{multino}(n; p_1, \dots, p_k), \end{aligned}$$

where the probabilities p_1, \dots, p_k are either specified fully or are specified as $p_1(\boldsymbol{\theta}), \dots, p_k(\boldsymbol{\theta})$, $\boldsymbol{\theta}$ being a vector of parameters with unknown values.

The expected values E_1, \dots, E_k (under the null hypothesis) of O_1, \dots, O_k are given by

$$E_i = \begin{cases} np_i & \text{if } p_1, \dots, p_k \text{ are specified fully,} \\ np_i(\hat{\boldsymbol{\theta}}) & \text{if } \boldsymbol{\theta} \text{ has to be estimated (as } \hat{\boldsymbol{\theta}}). \end{cases}$$

We compare the observed counts O_1, \dots, O_k with the corresponding expected values E_1, \dots, E_k using the statistic

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}. \quad (14.13)$$

(Here the symbol X denotes a capital Greek χ .)

If the null hypothesis specifies the distribution completely (i.e. the parameters are *known* if H_0 is true) then

$$X^2 \dot{\sim} \chi_{k-1}^2 \quad \text{under } H_0. \quad (14.14)$$

If q independent parameters in θ have to be estimated, then

$$X^2 \dot{\sim} \chi_{k-q-1}^2. \quad (14.15)$$

The usual rule of thumb for the approximations (14.14) and (14.15) is that they are reasonable if $E_i \geq 5$ for $i = 1, \dots, k$.

Since X^2 measures the discrepancy between the observed counts O_1, \dots, O_k and the expected counts E_1, \dots, E_k , H_0 is rejected for ‘large’ values of X^2 , so that this test is 1-tailed.

Example (The binomial case): As we saw in (14.12), if $Y \sim \text{bin}(n, p)$ then $(Y, n - Y) \sim \text{multino}(n; p, 1 - p)$. Then

$$\begin{aligned} X^2 &= \frac{(Y - np)^2}{np} + \frac{([n - Y] - [n(1 - p)])^2}{n(1 - p)} \\ &= \frac{(Y - np)^2}{np(1 - p)} \\ &\dot{\sim} \chi_1^2, \end{aligned}$$

which we recognise from (14.4).

Example: The following table gives the number of 6’s in 216 trials, each of which consisted of rolling 3 dice.

| | | | | |
|-----------|-----|----|----|---|
| count | 0 | 1 | 2 | 3 |
| frequency | 110 | 85 | 20 | 1 |

Test the null hypothesis that the dice are fair, i.e. test whether or not the data are consistent with a $\text{bin}(3, 1/6)$ distribution. (Hint: First calculate the expected frequencies of r 6’s for $r = 0, 1, 2, 3$.)

A proof of (14.14) will have to wait until Honours but here is a plausibility argument. The count o_j in interval j could be modelled as an observation of a Poisson random variable, i.e. $O_j \sim \mathcal{P}(np_j)$. (This is especially reasonable if the total number of observations is quite high and the probability of an observation being in any particular interval is quite low.) Thus, given moderately high expected values in each cell, we have

$$\frac{O_j - np_j}{\sqrt{np_j}} \dot{\sim} N(0, 1),$$

and so

$$\sum_{j=1}^k \frac{(O_j - np_j)^2}{np_j}$$

is a sum of squares of k standard normal random variables, suggesting that this has a χ^2 distribution. However, we cannot treat O_1, \dots, O_k as k independent random variables, since they satisfy (14.9), i.e.

$$O_1 + \dots + O_k = n.$$

Thus, if we know $k - 1$ of O_1, \dots, O_k then the remaining one is determined. This suggests that, although X^2 is a sum of squares of random variables which are approximately standard normal, it can have only $k - 1$ degrees of freedom, suggesting (correctly) that $X^2 \approx \chi_{k-1}^2$.

The χ^2 goodness-of-fit test

The χ^2 goodness-of-fit test which was described above was for testing hypotheses about multinomial distributions. This test can also be used for continuous distributions. This is done by first dividing the real line into k intervals and then considering that a trial results in outcome i if the observation falls in the i th interval. Although the division of the real line is arbitrary, it is sensible to choose the intervals so that each interval contains *at least* 5 observations.

14.2.2 Contingency tables

Many count data consist of counts of outcomes of events which can be classified in (at least) 2 ways. Consider observations classified according to two characteristics X and Y , where X takes r values (x_1, \dots, x_r , say) and Y takes c values (y_1, \dots, y_c , say). It is helpful to record them in an $r \times c$ *contingency table*. (Here the word ‘contingency’ means ‘outcome’.)

| | y_1 | y_2 | . | . | . | . | y_c | row total |
|--------------|----------|----------|---|---|---|---|----------|-----------|
| x_1 | O_{11} | O_{12} | . | . | . | . | O_{1c} | R_1 |
| x_2 | O_{21} | O_{22} | . | . | . | . | O_{2c} | R_2 |
| . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . |
| x_r | O_{r1} | O_{r2} | . | . | . | . | O_{rc} | R_r |
| column total | C_1 | C_2 | . | . | . | . | C_c | n |

Contingency tables arise in two ways:

- (i) n observations are taken on discrete random variables X and Y ,
- (ii) for $j = 1, \dots, c$, C_j observations are taken on a discrete random variable Y_j which takes values x_1, \dots, x_r .

In case (i) we wish to test

$$\begin{aligned} H_0: X \text{ and } Y \text{ are independent} \\ \text{against} \\ H_1: X \text{ and } Y \text{ are dependent} \end{aligned}$$

and the test is often called a *test of independence*. In case (ii) we wish to test

$$\begin{aligned} H_0: P(Y_j = x_i) = P(Y_k = x_i) \quad j, k = 1, \dots, c, \quad i = 1, \dots, r \\ \text{against} \\ H_1: P(Y_j = x_i) \neq P(Y_k = x_i) \quad \text{for some } j \neq k \text{ and } i \end{aligned}$$

and the test is often called a *test of homogeneity*.

Although the two cases arise in different contexts (in case (i) n is fixed, whereas in case (ii) C_1, \dots, C_c are fixed),

- (a) it is not possible to tell the context in which a table arose just from looking at the table,
- (b) the method of analysis is the same in the two cases.

The discussion will therefore concentrate on case (i).

Under the null hypothesis, the expected count E_{ij} in cell (i, j) (in row i and column j) of the table is

$$E_{ij} = n\hat{p}_i\hat{q}_j = \frac{R_i C_j}{n},$$

where

$$\hat{p}_i = \frac{R_i}{n} \quad \text{and} \quad \hat{q}_j = \frac{C_j}{n}$$

are the proportions of observations in categories x_i and y_j , respectively. Hence the table of expected values E_{ij} (under the null hypothesis) is

| | y_1 | y_2 | \cdot | \cdot | \cdot | \cdot | y_c | row total |
|--------------|-------------|-------------|---------|---------|---------|---------|-------------|-----------|
| x_1 | $C_1 R_1/n$ | $C_2 R_1/n$ | \cdot | \cdot | \cdot | \cdot | $C_c R_1/n$ | R_1 |
| x_2 | $C_1 R_2/n$ | $C_2 R_2/n$ | \cdot | \cdot | \cdot | \cdot | $C_c R_2/n$ | R_2 |
| \cdot | \cdot | \cdot | \cdot | \cdot | \cdot | \cdot | \cdot | \cdot |
| \cdot | \cdot | \cdot | \cdot | \cdot | \cdot | \cdot | \cdot | \cdot |
| x_r | $C_1 R_r/n$ | $C_2 R_r/n$ | \cdot | \cdot | \cdot | \cdot | $C_c R_r/n$ | R_r |
| column total | C_1 | C_2 | \cdot | \cdot | \cdot | \cdot | C_c | n |

From (14.13), the χ^2 statistic is

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \quad (14.16)$$

From (14.15), X^2 has a χ^2 distribution under the null hypothesis. The degrees of freedom need some explanation. The number k of observations is rc . It might seem that we have had to estimate $r + c$ parameters: p_1, \dots, p_r and q_1, \dots, q_c . However, since $p_1 + \dots + p_r = 1$ and $q_1 + \dots + q_c = 1$, p_1, \dots, p_r and q_1, \dots, q_c contain only $r - 1$ and $c - 1$ independent parameters, respectively. Thus the total number q of independent parameters to be estimated is $(r - 1) + (c - 1) = r + c - 2$. Then the appropriate degrees of freedom of the χ^2 statistic X^2 is

$$k - q - 1 = rc - (r + c - 2) - 1 = (r - 1)(c - 1),$$

so that, from (14.15),

$$X^2 \overset{\cdot}{\sim} \chi_{(r-1)(c-1)}^2 \quad \text{under } H_0. \quad (14.17)$$

Two independent binomial distributions

Two independent binomial random variables X_1 and X_2 with

$$X_1 \sim \text{bin}(n_1, p_1) \quad X_2 \sim \text{bin}(n_2, p_2)$$

(as in Section 14.1.3) give a 2×2 contingency table

| | | |
|-------------|---------------------------|-------------|
| X_1 | $n_1 - X_1$ | n_1 |
| X_2 | $n_2 - X_2$ | n_2 |
| $X_1 + X_2$ | $n_1 + n_2 - (X_1 + X_2)$ | $n_1 + n_2$ |

Some algebraic manipulation shows that

$$X^2 \approx \left(\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}} \right)^2,$$

where X^2 is the χ^2 statistic (which has an approximate χ_1^2 distribution in this case) and the expression inside the brackets on the right hand side is obtained by standardising (14.7) and substituting $\hat{p}_i = X_i/n_i$ for p_i . (This expression has an approximate $N(0, 1)$ distribution.)

Yates' correction in 2×2 contingency tables

In the case of 2×2 tables, the approximation (14.16) can be improved by Yates' correction, which replaces X^2 by

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(|O_{ij} - E_{ij}| - 1/2)^2}{E_{ij}}. \quad (14.18)$$

Example (Unemployment of school leavers):

The following table classifies 100 school leavers in the North of England and 50 school leavers in Scotland according to whether or not they had found work 6 months after leaving school.

| | Scotland | N. England |
|------------|----------|------------|
| Unemployed | 16 | 41 |
| Employed | 34 | 59 |

Do these data provide evidence of a difference in unemployment rates between Scottish and Northern English school leavers? In formal statistical language, we wish to test

$$H_0 : p_S = p_E \text{ against } H_1 : p_S \neq p_E,$$

where p_S and p_E are the probabilities of a school leaver being unemployed in Scotland and N. England, respectively. (Thus we wish to perform a test of homogeneity. Note that the proportions of Scottish and English are fixed by the survey design.)

Under H_0 , $p_S = p_E = p$ (say). We can estimate p from the data as $\hat{p} = 57/150 = 0.38$. Using this, calculate the expected number in each cell of the table under H_0 .

| | Scotland | N. England |
|------------|----------|------------|
| Unemployed | | |
| Employed | | |

Now calculate the value of the χ^2 test statistic (14.16) or (14.18):

$$\sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} =$$

and

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(|O_{ij} - E_{ij}| - 1/2)^2}{E_{ij}} =$$

Since $\chi_{1,0.05}^2 = 3.841$, the null hypothesis cannot be rejected at the 5% level. These data provide no evidence that unemployment rates of school leavers differ between the two sides of the border.

14.2.3 Contingency tables in R

The R command for carrying out a χ^2 test is `chisq.test`.

Example (Smoking and risk of heart attack):

Here are some data from the *Journal of Epidemiology and Community Health* (1989), **43**, 214–217, collected as part of a study to investigate smoking and risk of heart attack. The numbers are patients in each category.

| | Heart attack | Controls |
|--------------|--------------|----------|
| Had smoked | 172 | 173 |
| Never smoked | 90 | 346 |

The pattern here is so clear that formal hypothesis testing is hardly necessary. Nevertheless, the data serve to illustrate the way in which contingency tables can be analysed in R.

Putting the counts from the table into a matrix called `smoking` and applying the R command `chisq.test`, we get

```
> smoking<-matrix(c(172, 90,173,346), nc = 2)
> chisq.test(smoking)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: smoking
X-squared = 72.4241, df = 1, p-value = < 2.2e-16
```

Thus, there is extremely strong evidence of an association between smoking and heart attacks.

Example (Seat belts and car accidents):

The following data (from the Florida Department of Highway Safety, as reported in Agresti (1996) *An Introduction to Categorical Data Analysis*) classify driver and passenger injuries in car accidents by (i) whether or not the victim was wearing a seat belt, (ii) whether or not the victim was killed by the accident.

| | Fatal | Non fatal |
|---------|-------|-----------|
| No belt | 1601 | 162,527 |
| Belt | 510 | 412,368 |

The R output is

```
> seatbelt<-matrix(c(1601,510,162527,412368), nc = 2)
> seatbelt
      [,1] [,2]
[1,] 1601 162527
[2,]  510 412368
> chisq.test(seatbelt)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: seatbelt
X-squared = 2336.075, df = 1, p-value = < 2.2e-16
```

This is very strong evidence that the probability of being killed if involved in a car accident differs between those who wear seat belts and those who do not. (A slightly more careful analysis shows that this is very strong evidence *in favour* of wearing a seat belt.)

Example (Gender and political orientation): Here are some data from the US General Social Survey on party support and gender in 1991:

| | Democrat | Independent | Republican |
|--------|----------|-------------|------------|
| Female | 279 | 73 | 225 |
| Male | 165 | 47 | 191 |

```
> party<-matrix(c(279,165,73,47,225,191), nc = 3)
> party
      [,1] [,2] [,3]
[1,] 279  73 225
[2,] 165  47 191
> chisq.test(party)
```

Pearson's Chi-squared test

```
data: party
X-squared = 7.0095, df = 2, p-value = 0.03005
```

The small p -value indicates that there is evidence that political orientation differs between (American) men and women (with apparently greater support for the Democrats from women).
