A Handbook of Statistical Analyses Using R

SECOND EDITION

Brian S. Everitt and Torsten Hothorn

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CHAPTER 3

Simple Inference: Guessing Lengths, Wave Energy, Water Hardness, Piston Rings, and Rearrests of Juveniles

3.1 Introduction

Shortly after metric units of length were officially introduced in Australia in the 1970s, each of a group of 44 students was asked to guess, to the nearest metre, the width of the lecture hall in which they were sitting. Another group of 69 students in the same room was asked to guess the width in feet, to the nearest foot. The data were collected by Professor T. Lewis, and are given here in Table 3.1, which is taken from Hand et al. (1994). The main question is whether estimation in feet and in metres gives different results.

Table 3.1: `roomwidth` data. Room width estimates (`width`) in feet and in metres (`unit`).

`unit`	`width`	`unit`	`width`	`unit`	`width`	`unit`	`width`
metres	8	metres	16	feet	34	feet	45
metres	9	metres	16	feet	35	feet	45
metres	10	metres	17	feet	35	feet	45
metres	10	metres	17	feet	36	feet	45
metres	10	metres	17	feet	36	feet	45
metres	10	metres	17	feet	36	feet	46
metres	10	metres	18	feet	37	feet	46
metres	10	metres	18	feet	37	feet	47
metres	11	metres	20	feet	40	feet	48
metres	11	metres	22	feet	40	feet	48
metres	11	metres	25	feet	40	feet	50
metres	11	metres	27	feet	40	feet	50
metres	12	metres	35	feet	40	feet	50
metres	12	metres	38	feet	40	feet	51
metres	13	metres	40	feet	40	feet	54
metres	13	feet	24	feet	40	feet	54
metres	13	feet	25	feet	40	feet	54
metres	14	feet	27	feet	41	feet	55
metres	14	feet	30	feet	41	feet	55
metres	14	feet	30	feet	42	feet	60

Table 3.1: roomwidth data (continued).

unit	width	unit	width	unit	width	unit	width
metres	15	feet	30	feet	42	feet	60
metres	15	feet	30	feet	42	feet	63
metres	15	feet	30	feet	42	feet	70
metres	15	feet	30	feet	43	feet	75
metres	15	feet	32	feet	43	feet	80
metres	15	feet	32	feet	44	feet	94
metres	15	feet	33	feet	44		
metres	15	feet	34	feet	44		
metres	16	feet	34	feet	45		

In a design study for a device to generate electricity from wave power at sea, experiments were carried out on scale models in a wave tank to establish how the choice of mooring method for the system affected the bending stress produced in part of the device. The wave tank could simulate a wide range of sea states and the model system was subjected to the same sample of sea states with each of two mooring methods, one of which was considerably cheaper than the other. The resulting data (from Hand et al., 1994, giving root mean square bending moment in Newton metres) are shown in Table 3.2. The question of interest is whether bending stress differs for the two mooring methods.

Table 3.2: waves data. Bending stress (root mean squared bending moment in Newton metres) for two mooring methods in a wave energy experiment.

method1	method2	method1	method2	method1	method2
2.23	1.82	8.98	8.88	5.91	6.44
2.55	2.42	0.82	0.87	5.79	5.87
7.99	8.26	10.83	11.20	5.50	5.30
4.09	3.46	1.54	1.33	9.96	9.82
9.62	9.77	10.75	10.32	1.92	1.69
1.59	1.40	5.79	5.87	7.38	7.41

The data shown in Table 3.3 were collected in an investigation of environmental causes of disease and are taken from Hand et al. (1994). They show the annual mortality per 100,000 for males, averaged over the years 1958–1964, and the calcium concentration (in parts per million) in the drinking water for 61 large towns in England and Wales. The higher the calcium concentration, the harder the water. Towns at least as far north as Derby are identified in the

table. Here there are several questions that might be of interest including: are mortality and water hardness related, and do either or both variables differ between northern and southern towns?

Table 3.3: water data. Mortality (per 100,000 males per year, mortality) and water hardness for 61 cities in England and Wales.

location	town	mortality	hardness
South	Bath	1247	105
North	Birkenhead	1668	17
South	Birmingham	1466	5
North	Blackburn	1800	14
North	Blackpool	1609	18
North	Bolton	1558	10
North	Bootle	1807	15
South	Bournemouth	1299	78
North	Bradford	1637	10
South	Brighton	1359	84
South	Bristol	1392	73
North	Burnley	1755	12
South	Cardiff	1519	21
South	Coventry	1307	78
South	Croydon	1254	96
North	Darlington	1491	20
North	Derby	1555	39
North	Doncaster	1428	39
South	East Ham	1318	122
South	Exeter	1260	21
North	Gateshead	1723	44
North	Grimsby	1379	94
North	Halifax	1742	8
North	Huddersfield	1574	9
North	Hull	1569	91
South	Ipswich	1096	138
North	Leeds	1591	16
South	Leicester	1402	37
North	Liverpool	1772	15
North	Manchester	1828	8
North	Middlesbrough	1704	26
North	Newcastle	1702	44
South	Newport	1581	14
South	Northampton	1309	59
South	Norwich	1259	133
North	Nottingham	1427	27
North	Oldham	1724	6

Table 3.3: water data (continued).

location	town	mortality	hardness
South	Oxford	1175	107
South	Plymouth	1486	5
South	Portsmouth	1456	90
North	Preston	1696	6
South	Reading	1236	101
North	Rochdale	1711	13
North	Rotherham	1444	14
North	St Helens	1591	49
North	Salford	1987	8
North	Sheffield	1495	14
South	Southampton	1369	68
South	Southend	1257	50
North	Southport	1587	75
North	South Shields	1713	71
North	Stockport	1557	13
North	Stoke	1640	57
North	Sunderland	1709	71
South	Swansea	1625	13
North	Wallasey	1625	20
South	Walsall	1527	60
South	West Bromwich	1627	53
South	West Ham	1486	122
South	Wolverhampton	1485	81
North	York	1378	71

The two-way contingency table in Table 3.4 shows the number of piston-ring failures in each of three legs of four steam-driven compressors located in the same building (Haberman, 1973). The compressors have identical design and are oriented in the same way. The question of interest is whether the two categorical variables (compressor and leg) are independent.

The data in Table 3.5 (taken from Agresti, 1996) arise from a sample of juveniles convicted of felony in Florida in 1987. Matched pairs were formed using criteria such as age and the number of previous offences. For each pair, one subject was handled in the juvenile court and the other was transferred to the adult court. Whether or not the juvenile was rearrested by the end of 1988 was then noted. Here the question of interest is whether the true proportions rearrested were identical for the adult and juvenile court assignments?

Table 3.4: pistonrings data. Number of piston ring failures for three legs of four compressors.

		leg	
compressor	North	Centre	South
C1	17	17	12
C2	11	9	13
C3	11	8	19
C4	14	7	28

Source: From Haberman, S. J., *Biometrics*, 29, 205–220, 1973. With permission.

Table 3.5: rearrests data. Rearrests of juvenile felons by type of court in which they were tried.

	Juvenile court	
Adult court	Rearrest	No rearrest
Rearrest	158	515
No rearrest	290	1134

Source: From Agresti, A., *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New York, 1996. With permission.

3.2 Statistical Tests

Inference, the process of drawing conclusions about a population on the basis of measurements or observations made on a sample of individuals from the population, is central to statistics. In this chapter we shall use the data sets described in the introduction to illustrate both the application of the most common statistical tests, and some simple graphics that may often be used to aid in understanding the results of the tests. Brief descriptions of each of the tests to be used follow.

3.2.1 Comparing Normal Populations: Student's t-Tests

The t-test is used to assess hypotheses about two population means where the measurements are assumed to be sampled from a normal distribution. We shall describe two types of t-tests, the independent samples test and the paired test.

The independent samples t-test is used to test the null hypothesis that the means of two populations are the same, $H_0 : \mu_1 = \mu_2$, when a sample of observations from each population is available. The subjects of one population must not be individually matched with subjects from the other population and the subjects within each group should not be related to each other. The variable to be compared is assumed to have a normal distribution with the same standard deviation in both populations. The test statistic is essentially a standardised difference of the two sample means,

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{1/n_1 + 1/n_2}} \tag{3.1}$$

where $\bar{y}_i$ and n_i are the means and sample sizes in groups $i = 1$ and 2, respectively. The pooled standard deviation s is given by

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

where s_1 and s_2 are the standard deviations in the two groups.

Under the null hypothesis, the t-statistic has a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom. A $100(1 - \alpha)\%$ confidence interval for the difference between two means is useful in giving a plausible range of values for the differences in the two means and is constructed as

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha, n_1+n_2-2}\, s\sqrt{n_1^{-1} + n_2^{-1}}$$

where t_{α, n_1+n_2-2} is the percentage point of the t-distribution such that the cumulative distribution function, $\mathsf{P}(t \le t_{\alpha, n_1+n_2-2})$, equals $1 - \alpha/2$.

If the two populations are suspected of having different variances, a modified form of the t statistic, known as the Welch test, may be used, namely

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

In this case, t has a Student's t-distribution with ν degrees of freedom, where

$$\nu = \left(\frac{c}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}\right)^{-1}$$

with

$$c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

A paired t-test is used to compare the means of two populations when samples from the populations are available, in which each individual in one sample is paired with an individual in the other sample or each individual in the sample is observed twice. Examples of the former are anorexic girls and their healthy sisters and of the latter the same patients observed before and after treatment.

If the values of the variable of interest, y, for the members of the ith pair in groups 1 and 2 are denoted as y_{1i} and y_{2i}, then the differences $d_i = y_{1i} - y_{2i}$ are assumed to have a normal distribution with mean μ and the null hypothesis here is that the mean difference is zero, i.e., $H_0 : \mu = 0$. The paired t-statistic is

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

where $\bar{d}$ is the mean difference between the paired measurements and s is its standard deviation. Under the null hypothesis, t follows a t-distribution with $n - 1$ degrees of freedom. A $100(1 - \alpha)\%$ confidence interval for μ can be constructed by

$$\bar{d} \pm t_{\alpha, n-1}\, s/\sqrt{n}$$

where $\mathsf{P}(t \le t_{\alpha, n-1}) = 1 - \alpha/2$.

3.2.2 Non-parametric Analogues of Independent Samples and Paired t-Tests

One of the assumptions of both forms of t-test described above is that the data have a normal distribution, i.e., are unimodal and symmetric. When departures from those assumptions are extreme enough to give cause for concern, then it might be advisable to use the non-parametric analogues of the t-tests, namely the *Wilcoxon Mann-Whitney rank sum test* and the *Wilcoxon signed rank test*. In essence, both procedures throw away the original measurements and only retain the rankings of the observations.

For two independent groups, the Wilcoxon Mann-Whitney rank sum test applies the t-statistic to the joint ranks of all measurements in both groups instead of the original measurements. The null hypothesis to be tested is that the two populations being compared have identical distributions. For two normally distributed populations with common variance, this would be equivalent to the hypothesis that the means of the two populations are the same. The alternative hypothesis is that the population distributions differ in location, i.e., the median.

The test is based on the joint ranking of the observations from the two samples (as if they were from a single sample). The test statistic is the sum of the ranks of one sample (the lower of the two rank sums is generally used). A version of this test applicable in the presence of ties is discussed in Chapter 4.

For small samples, p-values for the test statistic can be assigned relatively simply. A large sample approximation is available that is suitable when the two sample sizes are greater and there are no ties. In R, the large sample approximation is used by default when the sample size in one group exceeds 50 observations.

In the paired situation, we first calculate the differences $d_i = y_{1i} - y_{2i}$ between each pair of observations. To compute the Wilcoxon signed-rank statistic, we rank the absolute differences $|d_i|$. The statistic is defined as the sum of the ranks associated with positive difference $d_i > 0$. Zero differences are discarded, and the sample size n is altered accordingly. Again, p-values for small sample sizes can be computed relatively simply and a large sample approximation is available. It should be noted that this test is valid only when the differences d_i are symmetrically distributed.

3.2.3 Testing Independence in Contingency Tables

When a sample of n observations in two nominal (categorical) variables are available, they can be arranged into a cross-classification (see Table 3.6) in which the number of observations falling in each cell of the table is recorded. Table 3.6 is an example of such a contingency table, in which the observations for a sample of individuals or objects are cross-classified with respect to two categorical variables. Testing for the independence of the two variables x and y is of most interest in general and details of the appropriate test follow.

Table 3.6: The general $r \times c$ table.

			y		
		1	...	c	
	1	n_{11}	...	n_{1c}	$n_{1\cdot}$
	2	n_{21}	...	n_{2c}	$n_{2\cdot}$
x	⋮	⋮	...	⋮	⋮
	r	n_{r1}	...	n_{rc}	$n_{r\cdot}$
		$n_{\cdot 1}$	...	$n_{\cdot c}$	n

Under the null hypothesis of independence of the row variable x and the column variable y, estimated expected values E_{jk} for cell (j, k) can be computed from the corresponding margin totals $E_{jk} = n_{j\cdot}n_{\cdot k}/n$. The test statistic for assessing independence is

$$X^2 = \sum_{j=1}^{r}\sum_{k=1}^{c} \frac{(n_{jk} - E_{jk})^2}{E_{jk}}.$$

Under the null hypothesis of independence, the test statistic X^2 is asymptotically distributed according to a χ^2-distribution with $(r-1)(c-1)$ degrees of freedom, the corresponding test is usually known as *chi-squared test.*

3.2.4 McNemar's Test

The chi-squared test on categorical data described previously assumes that the observations are independent. Often, however, categorical data arise from *paired* observations, for example, cases matched with controls on variables such as gender, age and so on, or observations made on the same subjects on two occasions (cf. paired t-test). For this type of paired data, the required procedure is McNemar's test. The general form of such data is shown in Table 3.7.

Table 3.7: Frequencies in matched samples data.

		Sample 1	
		present	absent
Sample 2	present	a	b
	absent	c	d

Under the hypothesis that the two populations do not differ in their probability of having the characteristic present, the test statistic

$$X^2 = \frac{(c-b)^2}{c+b}$$

has a χ^2-distribution with a single degree of freedom.

3.3 Analysis Using R

3.3.1 Estimating the Width of a Room

The data shown in Table 3.1 are available as `roomwidth` *data.frame* from the **HSAUR2** package and can be attached by using

```
R> data("roomwidth", package = "HSAUR2")
```

If we convert the estimates of the room width in metres into feet by multiplying each by 3.28 then we would like to test the hypothesis that the mean of the population of 'metre' estimates is equal to the mean of the population of 'feet' estimates. We shall do this first by using an independent samples t-test, but first it is good practise to check, informally at least, the normality and equal variance assumptions. Here we can use a combination of numerical and graphical approaches. The first step should be to convert the metre estimates into feet by a factor

```
R> convert <- ifelse(roomwidth$unit == "feet", 1, 3.28)
```

which equals one for all feet measurements and 3.28 for the measurements in metres. Now, we get the usual summary statistics and standard deviations of each set of estimates using

```
R> tapply(roomwidth$width * convert, roomwidth$unit, summary)
$feet
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   24.0    36.0    42.0    43.7    48.0    94.0

$metres
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.24   36.08   49.20   52.55   55.76  131.20
```

```
R> tapply(roomwidth$width * convert, roomwidth$unit, sd)

    feet   metres 
12.49742 23.43444 
```

where tapply applies summary, or sd, to the converted widths for both groups of measurements given by roomwidth$unit. A boxplot of each set of estimates might be useful and is depicted in Figure 3.1. The layout function (line 1 in Figure 3.1) divides the plotting area in three parts. The boxplot function produces a boxplot in the upper part and the two qqnorm statements in lines 8 and 11 set up the normal probability plots that can be used to assess the normality assumption of the *t*-test.

The boxplots indicate that both sets of estimates contain a number of outliers and also that the estimates made in metres are skewed and more variable than those made in feet, a point underlined by the numerical summary statistics above. Both normal probability plots depart from linearity, suggesting that the distributions of both sets of estimates are not normal. The presence of outliers, the apparently different variances and the evidence of non-normality all suggest caution in applying the *t*-test, but for the moment we shall apply the usual version of the test using the t.test function in R.

The two-sample test problem is specified by a *formula*, here by

```
I(width * convert) ~ unit
```

where the response, width, on the left hand side needs to be converted first and, because the star has a special meaning in formulae as will be explained in Chapter 5, the conversion needs to be embedded by I. The factor unit on the right hand side specifies the two groups to be compared.

From the output shown in Figure 3.2 we see that there is considerable evidence that the estimates made in feet are lower than those made in metres by between about 2 and 15 feet. The test statistic t from 3.1 is -2.615 and, with 111 degrees of freedom, the two-sided p-value is 0.01. In addition, a 95% confidence interval for the difference of the estimated widths between feet and metres is reported.

But this form of *t*-test assumes both normality and equality of population variances, both of which are suspect for these data. Departure from the equality of variance assumption can be accommodated by the modified *t*-test described above and this can be applied in R by choosing var.equal = FALSE (note that var.equal = FALSE is the default in R). The result shown in Figure 3.3 as well indicates that there is strong evidence for a difference in the means of the two types of estimate.

But there remains the problem of the outliers and the possible non-normality; consequently we shall apply the Wilcoxon Mann-Whitney test which, since it is based on the ranks of the observations, is unlikely to be affected by the outliers, and which does not assume that the data have a normal distribution. The test can be applied in R using the wilcox.test function.

Figure 3.4 shows a two-sided p-value of 0.028 confirming the difference in location of the two types of estimates of room width. Note that, due to ranking

```
1  R> layout(matrix(c(1,2,1,3), nrow = 2, ncol = 2, byrow = FALSE))
2  R> boxplot(I(width * convert) ~ unit, data = roomwidth,
3  +          ylab = "Estimated width (feet)",
4  +          varwidth = TRUE, names = c("Estimates in feet",
5  +          "Estimates in metres (converted to feet)"))
6  R> feet <- roomwidth$unit == "feet"
7  R> qqnorm(roomwidth$width[feet],
8  +         ylab = "Estimated width (feet)")
9  R> qqline(roomwidth$width[feet])
10 R> qqnorm(roomwidth$width[!feet],
11 +         ylab = "Estimated width (metres)")
12 R> qqline(roomwidth$width[!feet])
```

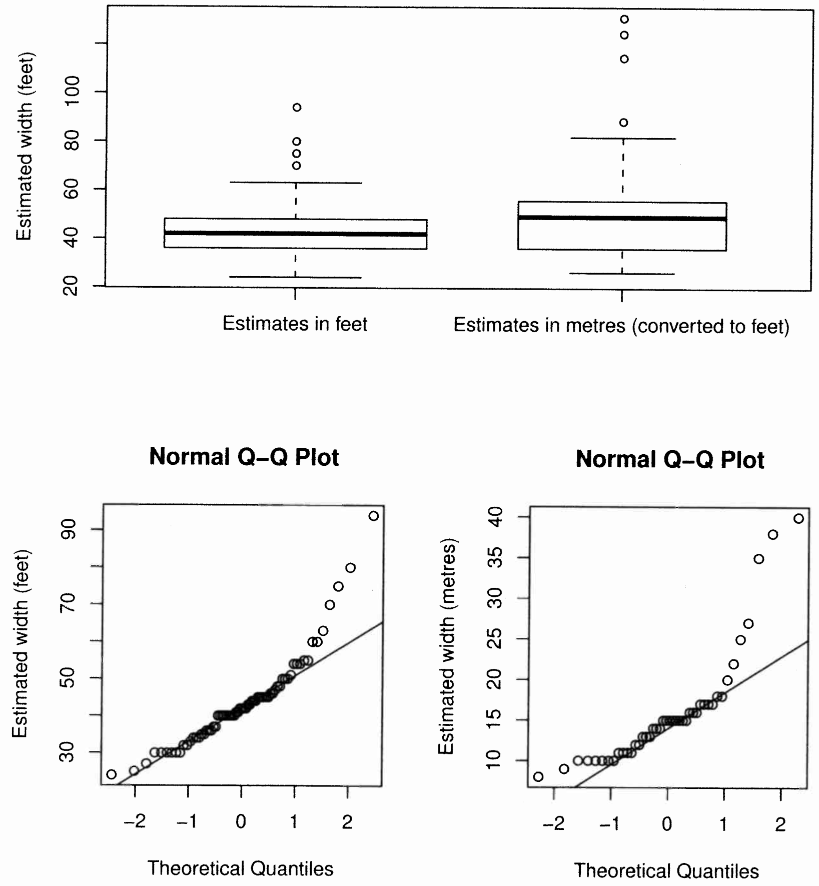

Figure 3.1 Boxplots of estimates of room width in feet and metres (after conversion to feet) and normal probability plots of estimates of room width made in feet and in metres.

```
R> t.test(I(width * convert) ~ unit, data = roomwidth,
+         var.equal = TRUE)

        Two Sample t-test

data:  I(width * convert) by unit
t = -2.6147, df = 111, p-value = 0.01017
95 percent confidence interval:
 -15.572734  -2.145052
sample estimates:
  mean in group feet mean in group metres
            43.69565             52.55455
```

Figure 3.2 R output of the independent samples t-test for the `roomwidth` data.

```
R> t.test(I(width * convert) ~ unit, data = roomwidth,
+         var.equal = FALSE)

        Welch Two Sample t-test

data:  I(width * convert) by unit
t = -2.3071, df = 58.788, p-value = 0.02459
95 percent confidence interval:
 -16.54308  -1.17471
sample estimates:
  mean in group feet mean in group metres
            43.69565             52.55455
```

Figure 3.3 R output of the independent samples Welch test for the `roomwidth` data.

the observations, the confidence interval for the median difference reported here is much smaller than the confidence interval for the difference in means as shown in Figures 3.2 and 3.3. Further possible analyses of the data are considered in Exercise 3.1 and in Chapter 4.

3.3.2 Wave Energy Device Mooring

The data from Table 3.2 are available as *data.frame* `waves`

```
R> data("waves", package = "HSAUR2")
```

and requires the use of a matched pairs t-test to answer the question of interest. This test assumes that the differences between the matched observations have a normal distribution so we can begin by checking this assumption by constructing a boxplot and a normal probability plot – see Figure 3.5.

The boxplot indicates a possible outlier, and the normal probability plot gives little cause for concern about departures from normality, although with

```
R> wilcox.test(I(width * convert) ~ unit, data = roomwidth,
+              conf.int = TRUE)

        Wilcoxon rank sum test with continuity correction

data:  I(width * convert) by unit
W = 1145, p-value = 0.02815
95 percent confidence interval:
 -9.3599953 -0.8000423
sample estimates:
difference in location
             -5.279955
```

Figure 3.4 R output of the Wilcoxon rank sum test for the `roomwidth` data.

only 18 observations it is perhaps difficult to draw any convincing conclusion. We can now apply the paired t-test to the data again using the `t.test` function. Figure 3.6 shows that there is no evidence for a difference in the mean bending stress of the two types of mooring device. Although there is no real reason for applying the non-parametric analogue of the paired t-test to these data, we give the R code for interest in Figure 3.7. The associated p-value is 0.316 confirming the result from the t-test.

3.3.3 Mortality and Water Hardness

There is a wide range of analyses we could apply to the data in Table 3.3 available from

```
R> data("water", package = "HSAUR2")
```

But to begin we will construct a scatterplot of the data enhanced somewhat by the addition of information about the marginal distributions of water hardness (calcium concentration) and mortality, and by adding the estimated linear regression fit (see Chapter 6) for mortality on hardness. The plot and the required R code is given along with Figure 3.8. In line 1 of Figure 3.8, we divide the plotting region into four areas of different size. The scatterplot (line 3) uses a plotting symbol depending on the location of the city (by the `pch` argument); a legend for the location is added in line 6. We add a least squares fit (see Chapter 6) to the scatterplot and, finally, depict the marginal distributions by means of a boxplot and a histogram. The scatterplot shows that as hardness increases mortality decreases, and the histogram for the water hardness shows it has a rather skewed distribution.

We can both calculate the Pearson's correlation coefficient between the two variables and test whether it differs significantly for zero by using the `cor.test` function in R. The test statistic for assessing the hypothesis that the population correlation coefficient is zero is

$$r/\sqrt{(1-r^2)/(n-2)}$$

```
R> mooringdiff <- waves$method1 - waves$method2
R> layout(matrix(1:2, ncol = 2))
R> boxplot(mooringdiff, ylab = "Differences (Newton metres)",
+          main = "Boxplot")
R> abline(h = 0, lty = 2)
R> qqnorm(mooringdiff, ylab = "Differences (Newton metres)")
R> qqline(mooringdiff)
```

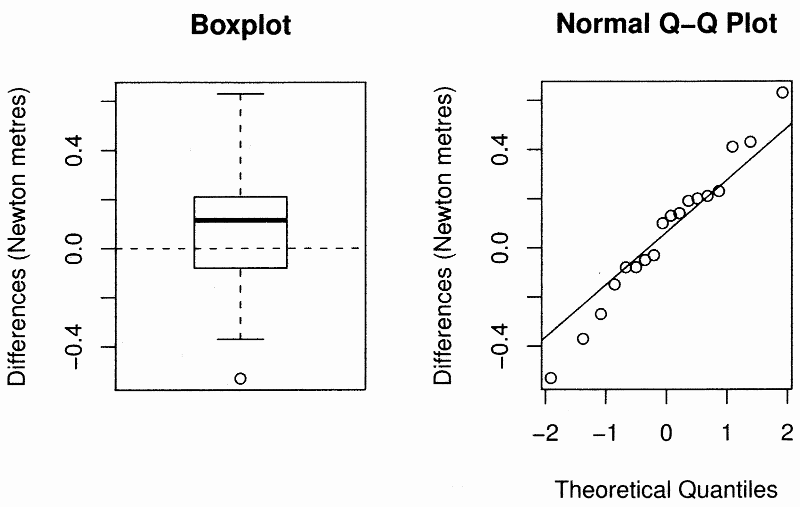

Figure 3.5 Boxplot and normal probability plot for differences between the two mooring methods.

where r is the sample correlation coefficient and n is the sample size. If the population correlation is zero and assuming the data have a bivariate normal distribution, then the test statistic has a Student's t distribution with $n - 2$ degrees of freedom.

The estimated correlation shown in Figure 3.9 is -0.655 and is highly significant. We might also be interested in the correlation between water hardness and mortality in each of the regions North and South but we leave this as an exercise for the reader (see Exercise 3.2).

3.3.4 Piston-ring Failures

The first step in the analysis of the `pistonrings` data is to apply the chi-squared test for independence. This we can do in R using the `chisq.test` function. The output of the chi-squared test, see Figure 3.10, shows a value of the X^2 test statistic of 11.722 with 6 degrees of freedom and an associated

```
R> t.test(mooringdiff)

        One Sample t-test

data:  mooringdiff
t = 0.9019, df = 17, p-value = 0.3797
95 percent confidence interval:
 -0.08258476  0.20591810
sample estimates:
 mean of x
0.06166667
```

Figure 3.6 R output of the paired t-test for the `waves` data.

```
R> wilcox.test(mooringdiff)

        Wilcoxon signed rank test with continuity correction

data:  mooringdiff
V = 109, p-value = 0.3165
```

Figure 3.7 R output of the Wilcoxon signed rank test for the `waves` data.

p-value of 0.068. The evidence for departure from independence of compressor and leg is not strong, but it may be worthwhile taking the analysis a little further by examining the estimated expected values and the differences of these from the corresponding observed value.

Rather than looking at the simple differences of observed and expected values for each cell which would be unsatisfactory since a difference of fixed size is clearly more important for smaller samples, it is preferable to consider a *standardised residual* given by dividing the observed minus the expected difference by the square root of the appropriate expected value. The X^2 statistic for assessing independence is simply the sum, over all the cells in the table, of the squares of these terms. We can find these values extracting the `residuals` element of the object returned by the `chisq.test` function

```
R> chisq.test(pistonrings)$residuals

          leg
compressor       North      Centre       South
        C1   0.6036154   1.6728267  -1.7802243
        C2   0.1429031   0.2975200  -0.3471197
        C3  -0.3251427  -0.4522620   0.6202463
        C4  -0.4157886  -1.4666936   1.4635235
```

A graphical representation of these residuals is called an *association plot* and is available via the `assoc` function from package **vcd** (Meyer et al., 2009) applied to the contingency table of the two categorical variables. Figure 3.11

```
R> nf <- layout(matrix(c(2, 0, 1, 3), 2, 2, byrow = TRUE),
+                c(2, 1), c(1, 2), TRUE)
R> psymb <- as.numeric(water$location)
R> plot(mortality ~ hardness, data = water, pch = psymb)
R> abline(lm(mortality ~ hardness, data = water))
R> legend("topright", legend = levels(water$location),
+          pch = c(1,2), bty = "n")
R> hist(water$hardness)
R> boxplot(water$mortality)
```

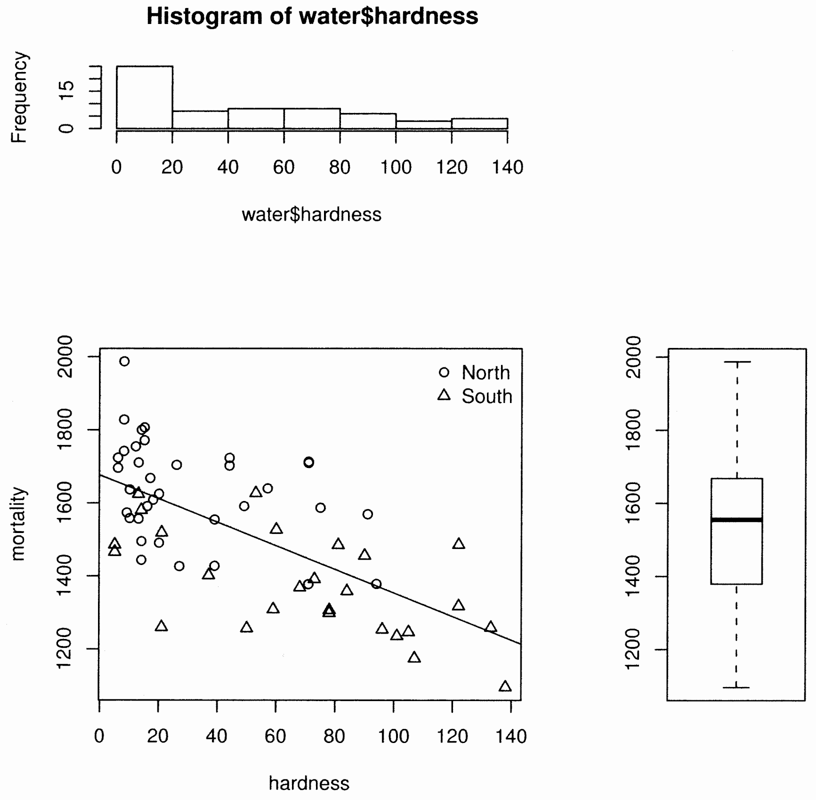

Figure 3.8 Enhanced scatterplot of water hardness and mortality, showing both the joint and the marginal distributions and, in addition, the location of the city by different plotting symbols.

```
R> cor.test(~ mortality + hardness, data = water)

        Pearson's product-moment correlation

data:  mortality and hardness
t = -6.6555, df = 59, p-value = 1.033e-08
95 percent confidence interval:
 -0.7783208 -0.4826129
sample estimates:
       cor
-0.6548486
```

Figure 3.9 R output of Pearsons' correlation coefficient for the water data.

```
R> data("pistonrings", package = "HSAUR2")
R> chisq.test(pistonrings)

        Pearson's Chi-squared test

data:  pistonrings
X-squared = 11.7223, df = 6, p-value = 0.06846
```

Figure 3.10 R output of the chi-squared test for the pistonrings data.

depicts the residuals for the piston ring data. The deviations from independence are largest for C1 and C4 compressors in the centre and south leg.

It is tempting to think that the size of these residuals may be judged by comparison with standard normal percentage points (for example greater than 1.96 or less than 1.96 for significance level $\alpha = 0.05$). Unfortunately it can be shown that the variance of a standardised residual is always less than or equal to one, and in some cases considerably less than one, however, the residuals are asymptotically normal. A more satisfactory 'residual' for contingency table data is considered in Exercise 3.3.

3.3.5 Rearrests of Juveniles

The data in Table 3.5 are available as *table* object via

```
R> data("rearrests", package = "HSAUR2")
R> rearrests

             Juvenile court
Adult court   Rearrest No rearrest
  Rearrest         158         515
  No rearrest      290        1134
```

and in rearrests the counts in the four cells refer to the matched pairs of subjects; for example, in 158 pairs both members of the pair were rearrested.

```
R> library("vcd")
R> assoc(pistonrings)
```

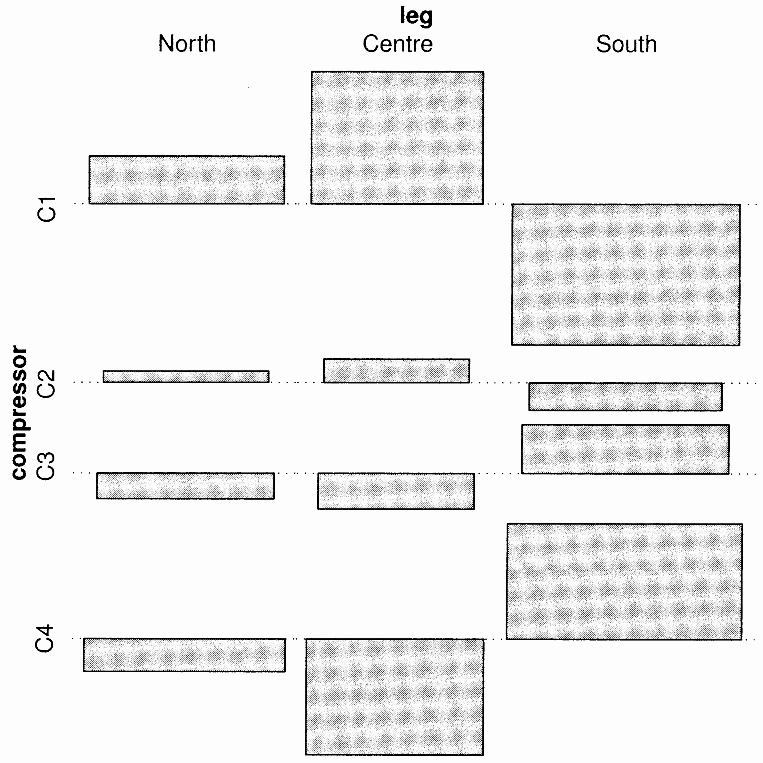

Figure 3.11 Association plot of the residuals for the `pistonrings` data.

Here we need to use McNemar's test to assess whether rearrest is associated with the type of court where the juvenile was tried. We can use the R function `mcnemar.test`. The test statistic shown in Figure 3.12 is 62.89 with a single degree of freedom – the associated p-value is extremely small and there is strong evidence that type of court and the probability of rearrest are related. It appears that trial at a juvenile court is less likely to result in rearrest (see Exercise 3.4). An exact version of McNemar's test can be obtained by testing whether b and c are equal using a binomial test (see Figure 3.13).

```
R> mcnemar.test(rearrests, correct = FALSE)

        McNemar's Chi-squared test

data:  rearrests
McNemar's chi-squared = 62.8882, df = 1, p-value =
2.188e-15
```

Figure 3.12 R output of McNemar's test for the `rearrests` data.

```
R> binom.test(rearrests[2], n = sum(rearrests[c(2,3)]))

        Exact binomial test

data:  rearrests[2] and sum(rearrests[c(2, 3)])
number of successes = 290, number of trials = 805,
p-value = 1.918e-15
95 percent confidence interval:
 0.3270278 0.3944969
sample estimates:
probability of success
             0.3602484
```

Figure 3.13 R output of an exact version of McNemar's test for the `rearrests` data computed via a binomial test.

3.4 Summary

Significance tests are widely used and they can easily be applied using the corresponding functions in R. But they often need to be accompanied by some graphical material to aid in interpretation and to assess whether assumptions are met. In addition, p-values are never as useful as confidence intervals.

Exercises

Ex. 3.1 After the students had made the estimates of the width of the lecture hall the room width was accurately measured and found to be 13.1 metres (43.0 feet). Use this additional information to determine which of the two types of estimates was more precise. ✓

Ex. 3.2 For the mortality and water hardness data calculate the correlation between the two variables in each region, north and south. ✓

Ex. 3.3 The standardised residuals calculated for the piston ring data are not entirely satisfactory for the reasons given in the text. An alternative residual suggested by Haberman (1973) is defined as the ratio of the standardised residuals and an adjustment:

$$\frac{\sqrt{(n_{jk} - E_{jk})^2/E_{jk}}}{\sqrt{(1 - n_{j\cdot}/n)(1 - n_{\cdot k}/n)}}.$$

When the variables forming the contingency table are independent, the adjusted residuals are approximately normally distributed with mean zero and standard deviation one. Write a general R function to calculate both standardised and adjusted residuals for any $r \times c$ contingency table and apply it to the piston ring data.

Ex. 3.4 For the data in table `rearrests` estimate the difference between the probability of being rearrested after being tried in an adult court and in a juvenile court, and find a 95% confidence interval for the population difference.

CHAPTER 4

Conditional Inference: Guessing Lengths, Suicides, Gastrointestinal Damage, and Newborn Infants

4.1 Introduction

There are many experimental designs or studies where the subjects are not a random sample from some well-defined population. For example, subjects recruited for a clinical trial are hardly ever a random sample from the set of all people suffering from a certain disease but are a selection of patients showing up for examination in a hospital participating in the trial. Usually, the subjects are randomly assigned to certain groups, for example a control and a treatment group, and the analysis needs to take this randomisation into account. In this chapter, we discuss such test procedures usually known as *(re)-randomisation* or *permutation tests*.

In the room width estimation experiment reported in Chapter 3, 40 of the estimated widths (in feet) of 69 students and 26 of the estimated widths (in metres) of 44 students are tied. In fact, this violates one assumption of the *unconditional* test procedures applied in Chapter 3, namely that the measurements are drawn from a continuous distribution. In this chapter, the data will be reanalysed using conditional test procedures, i.e., statistical tests where the distribution of the test statistics under the null hypothesis is determined *conditionally* on the data at hand. A number of other data sets will also be considered in this chapter and these will now be described.

Mann (1981) reports a study carried out to investigate the causes of jeering or baiting behaviour by a crowd when a person is threatening to commit suicide by jumping from a high building. A hypothesis is that baiting is more likely to occur in warm weather. Mann (1981) classified 21 accounts of threatened suicide by two factors, the time of year and whether or not baiting occurred. The data are given in Table 4.1 and the question is whether they give any evidence to support the hypothesis? The data come from the northern hemisphere, so June–September are the warm months.