A Handbook of Statistical Analyses Using R

Brian S. Everitt and Torsten Hothorn

CHAPTER 2

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

2.1 Introduction

2.2 Statistical Tests

2.3 Analysis Using R

2.3.1 Estimating the Width of a Room

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

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

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 practice to, informally at least, check 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, i.e., 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 2.1. The `layout` function (line 1 in Figure 2.1) 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 4, the conversion needs to be embedded by `I`. The factor `unit` on the right hand side specifies the two groups to be compared.

2.3.2 Wave Energy Device Mooring

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

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

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 2.5.

2.3.3 Mortality and Water Hardness

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

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

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 5) for mortality on hardness. The plot and the required R code is given along with Figure 2.8. In line 1 of Figure 2.8, we divide the plotting region into four areas of different size. The scatterplot

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

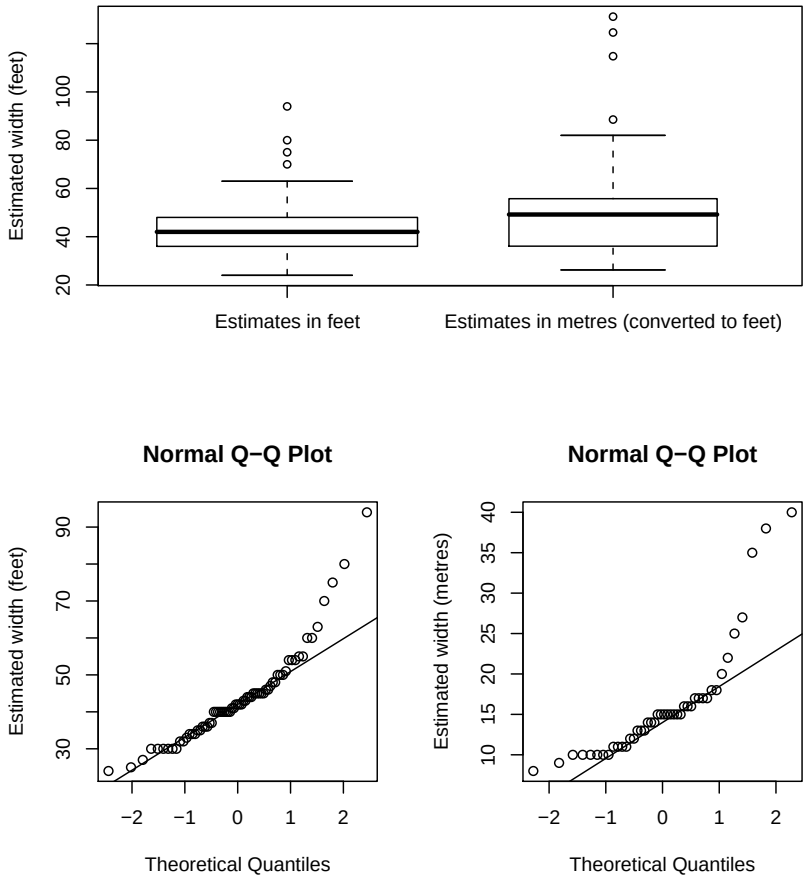

Figure 2.1 Boxplots of estimates of width of room in feet and metres (after conversion to feet) and normal probability plots of estimates of room width made in feet and in metres.

```
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
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.572734  -2.145052
sample estimates:
  mean in group feet mean in group metres
            43.69565             52.55455
```

Figure 2.2 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
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.54308  -1.17471
sample estimates:
  mean in group feet mean in group metres
            43.69565             52.55455
```

Figure 2.3 R output of the independent samples Welch test for the `roomwidth` data.

(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 5) 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.

2.3.4 Piston-ring Failures

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 expected difference by the square root of the appropriate expected value. The X^2 statistic for

```
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
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -9.3599953 -0.8000423
sample estimates:
difference in location
             -5.279955
```

Figure 2.4 R output of the Wilcoxon rank sum test for the `roomwidth` data.

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 *association plot* and is available via the `assoc` function from package *vcd* (Meyer et al., 2006) applied to the contingency table of the two categorical variables. Figure 2.11 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.

2.3.5 Rearrests of Juveniles

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

```
R> data("rearrests", package = "HSAUR")
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. Here we need to use McNemar's test to assess whether rearrest is associated

```
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)
```

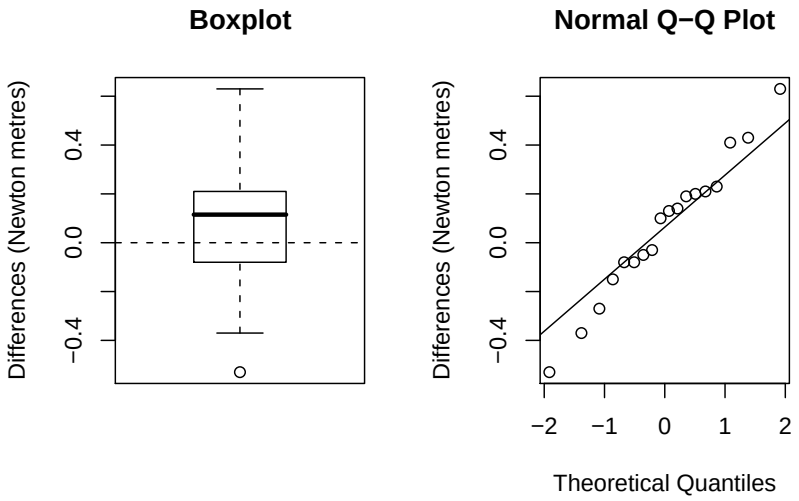

Figure 2.5 Boxplot and normal probability plot for differences between the two mooring methods.

with type of court where the juvenile was tried. We can use the R function `mcnemar.test`. The test statistic shown in Figure 2.12 is 62.888 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 2.4). An exact version of McNemar's test can be obtained by testing whether b and c are equal using a binomial test (see Figure 2.13).

```
R> t.test(mooringdiff)

        One Sample t-test

data:  mooringdiff
t = 0.9019, df = 17, p-value = 0.3797
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.08258476  0.20591810
sample estimates:
 mean of x
0.06166667
```

Figure 2.6 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
alternative hypothesis: true location is not equal to 0
```

Figure 2.7 R output of the Wilcoxon signed rank test for the `waves` data.

```
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)
```

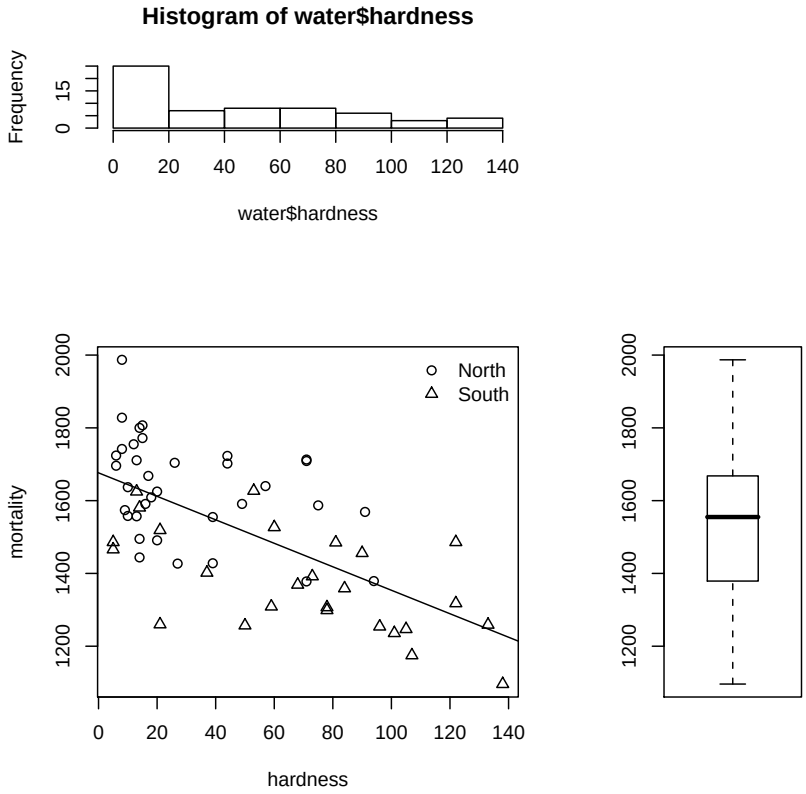

Figure 2.8 Enhanced scatterplot of water hardness and mortality, showing both the joint and the marginal distributions and, in addition, the location of the city by different plotting symbols.

```
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
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7783208 -0.4826129
sample estimates:
       cor
-0.6548486
```

Figure 2.9 R output of Pearsons' correlation coefficient for the `water` data.

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

        Pearson's Chi-squared test

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

Figure 2.10 R output of the chi-squared test for the `pistonrings` data.

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

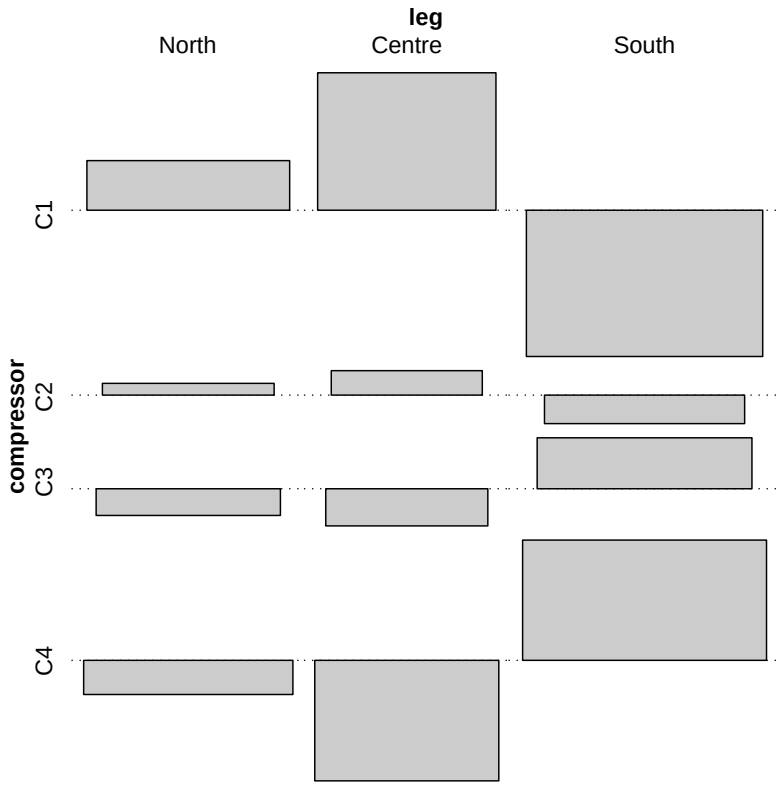

Figure 2.11 Association plot of the residuals for the `pistonrings` data.

```
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 2.12 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
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3270278 0.3944969
sample estimates:
probability of success
             0.3602484
```

Figure 2.13 R output of an exact version of McNemar's test for the `rearrests` data computed via a binomal test.

Bibliography

Meyer, D., Zeileis, A., Karatzoglou, A., and Hornik, K. (2006), *vcd: Visualizing Categorical Data*, URL http://CRAN.R-project.org, R package version 1.0-2.