1

Basics

The purpose of this chapter is to get you started using R. It is assumed that you have a working installation of the software and of the `ISwR` package that contains the data sets for this book. Instructions for obtaining and installing the software are given in Appendix A.

The text that follows describes R version 2.6.2. As of this writing, that is the latest version of R. As far as possible, I present the issues in a way that is independent of the operating system in use and assume that the reader has the elementary operational knowledge to select from menus, move windows around, etc. I do, however, make exceptions where I am aware of specific difficulties with a particular platform or specific features of it.

1.1 First steps

This section gives an introduction to the R computing environment and walks you through its most basic features.

Starting R is straightforward, but the method will depend on your computing platform. You will be able to launch it from a system menu, by double-clicking an icon, or by entering the command "`R`" at the system command line. This will either produce a console window or cause R to start up as an interactive program in the current terminal window. In

P. Dalgaard, *Introductory Statistics with R*,
DOI: 10.1007/978-0-387-79054-1_1,

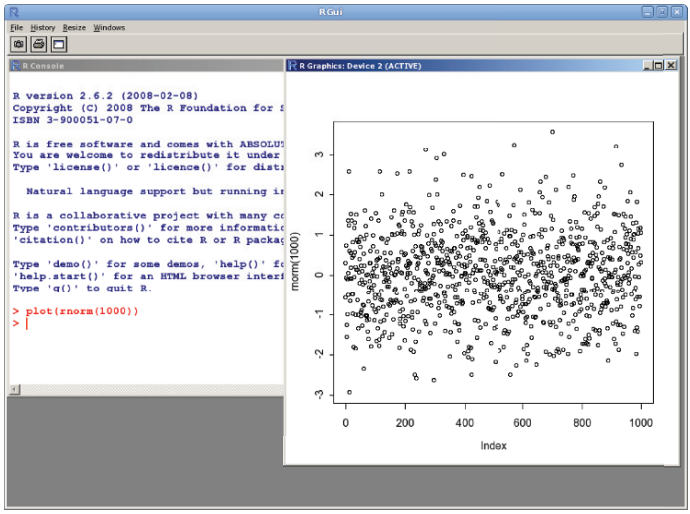

Figure 1.1. Screen image of R for Windows.

either case, R works fundamentally by the question-and-answer model: You enter a line with a command and press Enter (↩). Then the program does something, prints the result if relevant, and asks for more input. When R is ready for input, it prints out its prompt, a ">". It is possible to use R as a text-only application, and also in batch mode, but for the purposes of this chapter, I assume that you are sitting at a graphical workstation.

All the examples in this book should run if you type them in exactly as printed, *provided* that you have the `ISwR` package not only installed but also loaded into your current search path. This is done by entering

```
> library(ISwR)
```

at the command prompt. You do not need to understand what the command does at this point. It is explained in Section 2.1.5.

For a first impression of what R can do, try typing the following:

```
> plot(rnorm(1000))
```

This command draws 1000 numbers at random from the normal distribution (rnorm = *r*andom *norm*al) and plots them in a pop-up graphics window. The result on a Windows machine can be seen in Figure 1.1.

Of course, you are not expected at this point to guess that you would obtain this result in that particular way. The example is chosen because it shows several components of the user interface in action. Before the style

of commands will fall naturally, it is necessary to introduce some concepts and conventions through simpler examples.

Under Windows, the graphics window will have taken the keyboard focus at this point. Click on the console to make it accept further commands.

1.1.1 An overgrown calculator

One of the simplest possible tasks in R is to enter an arithmetic expression and receive a result. (The second line is the answer from the machine.)

```
> 2 + 2
[1] 4
```

So the machine knows that 2 plus 2 makes 4. Of course, it also knows how to do other standard calculations. For instance, here is how to compute e^{-2}:

```
> exp(-2)
[1] 0.1353353
```

The `[1]` in front of the result is part of R's way of printing numbers and vectors. It is not useful here, but it becomes so when the result is a longer vector. The number in brackets is the index of the first number on that line. Consider the case of generating 15 random numbers from a normal distribution:

```
> rnorm(15)
 [1] -0.18326112 -0.59753287 -0.67017905  0.16075723  1.28199575
 [6]  0.07976977  0.13683303  0.77155246  0.85986694 -1.01506772
[11] -0.49448567  0.52433026  1.07732656  1.09748097 -1.09318582
```

Here, for example, the `[6]` indicates that 0.07976977 is the sixth element in the vector. (For typographical reasons, the examples in this book are made with a shortened line width. If you try it on your own machine, you will see the values printed with six numbers per line rather than five. The numbers themselves will also be different since random number generation is involved.)

1.1.2 Assignments

Even on a calculator, you will quickly need some way to store intermediate results, so that you do not have to key them in over and over again. R, like other computer languages, has *symbolic variables*, that is names that

can be used to represent values. To assign the value 2 to the variable `x`, you can enter

```
> x <- 2
```

The two characters <− should be read as a single symbol: an arrow pointing to the variable to which the value is assigned. This is known as the *assignment operator.* Spacing around operators is generally disregarded by R, but notice that adding a space in the middle of a <− changes the meaning to "less than" followed by "minus" (conversely, omitting the space when comparing a variable to a negative number has unexpected consequences!).

There is no immediately visible result, but from now on, `x` has the value 2 and can be used in subsequent arithmetic expressions.

```
> x
[1] 2
> x + x
[1] 4
```

Names of variables can be chosen quite freely in R. They can be built from letters, digits, and the period (dot) symbol. There is, however, the limitation that the name must not start with a digit or a period followed by a digit. Names that start with a period are special and should be avoided. A typical variable name could be `height.1yr`, which might be used to describe the height of a child at the age of 1 year. Names are case-sensitive: `WT` and `wt` do not refer to the same variable.

Some names are already used by the system. This can cause some confusion if you use them for other purposes. The worst cases are the single-letter names `c`, `q`, `t`, `C`, `D`, `F`, `I`, and `T`, but there are also `diff`, `df`, and `pt`, for example. Most of these are functions and do not usually cause trouble when used as variable names. However, `F` and `T` are the standard abbreviations for `FALSE` and `TRUE` and no longer work as such if you redefine them.

1.1.3 Vectorized arithmetic

You cannot do much statistics on single numbers! Rather, you will look at data from a group of patients, for example. One strength of R is that it can handle entire *data vectors* as single objects. A data vector is simply an array of numbers, and a vector variable can be constructed like this:

```
> weight <- c(60, 72, 57, 90, 95, 72)
> weight
[1] 60 72 57 90 95 72
```

The construct `c(...)` is used to define vectors. The numbers are made up but might represent the weights (in kg) of a group of normal men.

This is neither the only way to enter data vectors into R nor is it generally the preferred method, but short vectors are used for many other purposes, and the `c(...)` construct is used extensively. In Section 2.4, we discuss alternative techniques for reading data. For now, we stick to a single method.

You can do calculations with vectors just like ordinary numbers, as long as they are of the same length. Suppose that we also have the heights that correspond to the weights above. The body mass index (BMI) is defined for each person as the weight in kilograms divided by the square of the height in meters. This could be calculated as follows:

```
> height <- c(1.75, 1.80, 1.65, 1.90, 1.74, 1.91)
> bmi <- weight/height^2
> bmi
[1] 19.59184 22.22222 20.93664 24.93075 31.37799 19.73630
```

Notice that the operation is carried out elementwise (that is, the first value of `bmi` is $60/1.75^2$ and so forth) and that the `^` operator is used for raising a value to a power. (On some keyboards, `^` is a "dead key" and you will have to press the spacebar afterwards to make it show.)

It is in fact possible to perform arithmetic operations on vectors of different length. We already used that when we calculated the `height^2` part above since `2` has length 1. In such cases, the shorter vector is *recycled*. This is mostly used with vectors of length 1 (scalars) but sometimes also in other cases where a repeating pattern is desired. A warning is issued if the longer vector is not a multiple of the shorter in length.

These conventions for vectorized calculations make it very easy to specify typical statistical calculations. Consider, for instance, the calculation of the mean and standard deviation of the `weight` variable.

First, calculate the mean, $\bar{x} = \sum x_i/n$:

```
> sum(weight)
[1] 446
> sum(weight)/length(weight)
[1] 74.33333
```

Then save the mean in a variable `xbar` and proceed with the calculation of SD $= \sqrt{(\sum(x_i - \bar{x})^2)/(n-1)}$. We do this in steps to see the individual components. The deviations from the mean are

```
> xbar <- sum(weight)/length(weight)
> weight - xbar
```

```
[1] -14.333333  -2.333333 -17.333333  15.666667  20.666667
[6]  -2.333333
```

Notice how `xbar`, which has length 1, is recycled and subtracted from each element of `weight`. The squared deviations will be

```
> (weight - xbar)^2
[1] 205.444444   5.444444 300.444444 245.444444 427.111111
[6]   5.444444
```

Since this command is quite similar to the one before it, it is convenient to enter it by editing the previous command. On most systems running R, the previous command can be recalled with the up-arrow key.

The sum of squared deviations is similarly obtained with

```
> sum((weight - xbar)^2)
[1] 1189.333
```

and all in all the standard deviation becomes

```
> sqrt(sum((weight - xbar)^2)/(length(weight) - 1))
[1] 15.42293
```

Of course, since R is a statistical program, such calculations are already built into the program, and you get the same results just by entering

```
> mean(weight)
[1] 74.33333
> sd(weight)
[1] 15.42293
```

1.1.4 Standard procedures

As a slightly more complicated example of what R can do, consider the following: The rule of thumb is that the BMI for a normal-weight individual should be between 20 and 25, and we want to know if our data deviate systematically from that. You might use a one-sample *t* test to assess whether the six persons' BMI can be assumed to have mean 22.5 given that they come from a normal distribution. To this end, you can use the function `t.test`. (You might not know the theory of the *t* test yet. The example is included here mainly to give some indication of what "real" statistical output looks like. A thorough description of `t.test` is given in Chapter 5.)

```
> t.test(bmi, mu=22.5)
        One Sample t-test
data:  bmi
t = 0.3449, df = 5, p-value = 0.7442
alternative hypothesis: true mean is not equal to 22.5
95 percent confidence interval:
 18.41734 27.84791
sample estimates:
mean of x
 23.13262
```

The argument `mu=22.5` attaches a value to the formal argument `mu`, which represents the Greek letter μ conventionally used for the theoretical mean. If this is not given, `t.test` would use the default `mu=0`, which is not of interest here.

For a test like this, we get a more extensive printout than in the earlier examples. The details of the output are explained in Chapter 5, but you might focus on the p-value which is used for testing the hypothesis that the mean is 22.5. The p-value is not small, indicating that it is not at all unlikely to get data like those observed if the mean were in fact 22.5. (Loosely speaking; actually p is the probability of obtaining a t value bigger than 0.3449 or less than -0.3449.) However, you might also look at the 95% confidence interval for the true mean. This interval is quite wide, indicating that we really have very little information about the true mean.

1.1.5 Graphics

One of the most important aspects of the presentation and analysis of data is the generation of proper graphics. R — like S before it — has a model for constructing plots that allows simple production of standard plots as well as fine control over the graphical components.

If you want to investigate the relation between `weight` and `height`, the first idea is to plot one versus the other. This is done by

```
> plot(height,weight)
```

leading to Figure 1.2.

You will often want to modify the drawing in various ways. To that end, there are a wealth of plotting parameters that you can set. As an example, let us try changing the plotting symbol using the keyword `pch` ("plotting character") like this:

```
> plot(height, weight, pch=2)
```

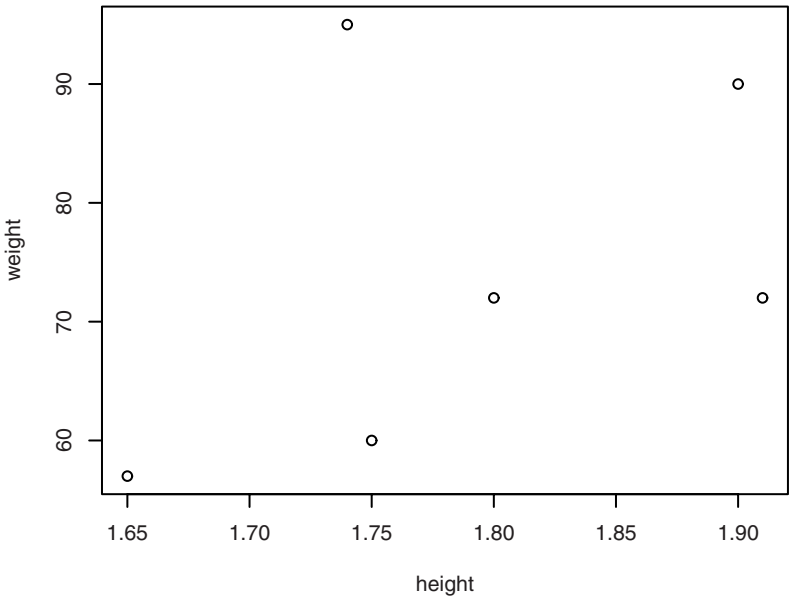

Figure 1.2. A simple *x*–*y* plot.

This gives the plot in Figure 1.3, with the points now marked with little triangles.

The idea behind the BMI calculation is that this value should be independent of the person's height, thus giving you a single number as an indication of whether someone is overweight and by how much. Since a normal BMI should be about 22.5, you would expect that $weight \approx 22.5 \times height^2$. Accordingly, you can superimpose a curve of expected weights at BMI 22.5 on the figure:

```
> hh <- c(1.65, 1.70, 1.75, 1.80, 1.85, 1.90)
> lines(hh, 22.5 * hh^2)
```

yielding Figure 1.4. The function `lines` will *add* (x, y) values joined by straight lines to an existing plot.

The reason for defining a new variable (`hh`) with heights rather than using the original `height` vector is twofold. First, the relation between height and weight is a quadratic one and hence nonlinear, although it can be difficult to see on the plot. Since we are approximating a nonlinear curve with a piecewise linear one, it will be better to use points that are spread evenly along the *x*-axis than to rely on the distribution of the original data. Sec-

ond, since the values of `height` are not sorted, the line segments would not connect neighbouring points but would run back and forth between distant points.

1.2 R language essentials

This section outlines the basic aspects of the R language. It is necessary to do this in a slightly superficial manner, with some of the finer points glossed over. The emphasis is on items that are useful to know in interactive usage as opposed to actual programming, although a brief section on programming is included.

1.2.1 Expressions and objects

The basic interaction mode in R is one of expression evaluation. The user enters an expression; the system evaluates it and prints the result. Some expressions are evaluated not for their result but for *side effects* such as

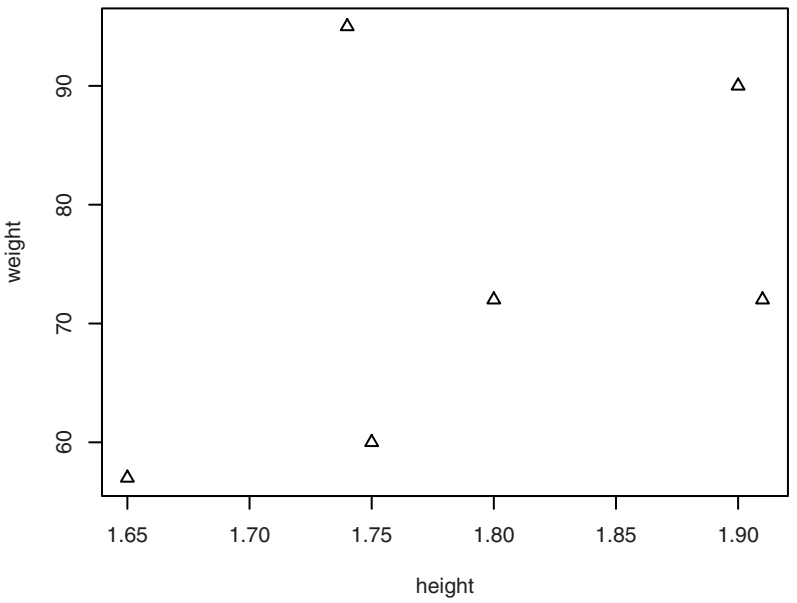

Figure 1.3. Plot with `pch = 2`.

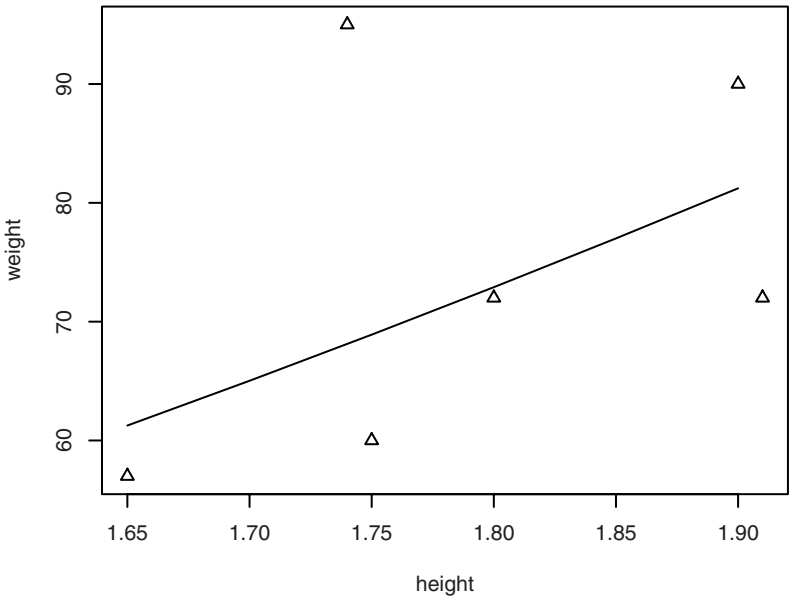

Figure 1.4. Superimposed reference curve, using `lines(...)`.

putting up a graphics window or writing to a file. All R expressions return a value (possibly `NULL`), but sometimes it is "invisible" and not printed.

Expressions typically involve variable references, operators such as `+`, and function calls, as well as some other items that have not been introduced yet.

Expressions work on *objects*. This is an abstract term for anything that can be assigned to a variable. R contains several different types of objects. So far, we have almost exclusively seen numeric vectors, but several other types are introduced in this chapter.

Although objects can be discussed abstractly, it would make a rather boring read without some indication of how to generate them and what to do with them. Conversely, much of the expression syntax makes little sense without knowledge of the objects on which it is intended to work. Therefore, the subsequent sections alternate between introducing new objects and introducing new language elements.

1.2.2 Functions and arguments

At this point, you have obtained an impression of the way R works, and we have already used some of the special terminology when talking about the plot *function*, etc. That is exactly the point: Many things in R are done using *function calls*, commands that look like an application of a mathematical function of one or several variables; for example, `log(x)` or `plot(height, weight)`.

The format is that a function name is followed by a set of parentheses containing one or more arguments. For instance, in `plot(height,weight)` the function name is `plot` and the arguments are `height` and `weight`. These are the *actual arguments*, which apply only to the current call. A function also has *formal arguments*, which get connected to actual arguments in the call.

When you write `plot(height, weight)`, R assumes that the first argument corresponds to the x-variable and the second one to the y-variable. This is known as *positional matching*. This becomes unwieldy if a function has a large number of arguments since you have to supply every one of them and remember their position in the sequence. Fortunately, R has methods to avoid this: Most arguments have sensible defaults and can be omitted in the standard cases, and there are nonpositional ways of specifying them when you need to depart from the default settings.

The `plot` function is in fact an example of a function that has a large selection of arguments in order to be able to modify symbols, line widths, titles, axis type, and so forth. We used the alternative form of specifying arguments when setting the plot symbol to triangles with `plot(height, weight, pch=2)`.

The `pch=2` form is known as a *named actual argument*, whose name can be matched against the formal arguments of the function and thereby allow *keyword matching* of arguments. The keyword `pch` was used to say that the argument is a specification of the plotting character. This type of function argument can be specified in arbitrary order. Thus, you can write `plot(y=weight,x=height)` and get the same plot as with `plot(x=height,y=weight)`.

The two kinds of argument specification — positional and named — can be mixed in the same call.

Even if there are no arguments to a function call, you have to write, for example, `ls()` for displaying the contents of the workspace. A common error is to leave off the parentheses, which instead results in the display of a piece of R code since `ls` entered by itself indicates that you want to see the definition of the function rather than execute it.

The *formal arguments* of a function are part of the function definition. The set of formal arguments to a function, for instance `plot.default` (which is the function that gets called when you pass `plot` an `x` argument for which no special plot method exists), may be seen with

```
> args(plot.default)
function (x, y = NULL, type = "p", xlim = NULL, ylim = NULL,
    log = "", main = NULL, sub = NULL, xlab = NULL, ylab = NULL,
    ann = par("ann"), axes = TRUE, frame.plot = axes,
    panel.first = NULL, panel.last = NULL, asp = NA, ...)
```

Notice that most of the arguments have defaults, meaning that if you do not specify (say) the `type` argument, the function will behave as if you had passed `type="p"`. The `NULL` defaults for many of the arguments really serve as indicators that the argument is unspecified, allowing special behaviour to be defined inside the function. For instance, if they are not specified, the `xlab` and `ylab` arguments are constructed from the actual arguments passed as `x` and `y`. (There are some very fine points associated with this, but we do not go further into the topic.)

The triple-dot (`...`) argument indicates that this function will accept additional arguments of unspecified name and number. These are often meant to be passed on to other functions, although some functions treat it specially. For instance, in `data.frame` and `c`, the names of the `...`-arguments become the names of the elements of the result.

1.2.3 Vectors

We have already seen numeric vectors. There are two further types, character vectors and logical vectors.

A *character vector* is a vector of text strings, whose elements are specified and printed in quotes:

```
> c("Huey","Dewey","Louie")
[1] "Huey"  "Dewey" "Louie"
```

It does not matter whether you use single- or double-quote symbols, as long as the left quote is the same as the right quote:

```
> c('Huey','Dewey','Louie')
[1] "Huey"  "Dewey" "Louie"
```

However, you should avoid the *acute accent* key (´), which is present on some keyboards. Double quotes are used throughout this book to prevent mistakes. *Logical vectors* can take the value `TRUE` or `FALSE` (or `NA`; see below). In input, you may use the convenient abbreviations `T` and `F` (if you

are careful not to redefine them). Logical vectors are constructed using the `c` function just like the other vector types:

```
> c(T,T,F,T)
[1]  TRUE  TRUE FALSE  TRUE
```

Actually, you will not often have to specify logical vectors in the manner above. It is much more common to use single logical values to turn an option on or off in a function call. Vectors of more than one value most often result from *relational expressions*:

```
> bmi > 25
[1] FALSE FALSE FALSE FALSE  TRUE FALSE
```

We return to relational expressions and logical operations in the context of conditional selection in Section 1.2.12.

1.2.4 Quoting and escape sequences

Quoted character strings require some special considerations: How, for instance, do you put a quote symbol inside a string? And what about special characters such as newlines? This is done using *escape sequences*. We shall look at those in a moment, but first it will be useful to observe the following.

There is a distinction between a text string and the way it is printed. When, for instance, you give the string `"Huey"`, it is a string of four characters, not six. The quotes are not actually part of the string, they are just there so that the system can tell the difference between a string and a variable name.

If you print a character vector, it usually comes out with quotes added to each element. There is a way to avoid this, namely to use the `cat` function. For instance,

```
> cat(c("Huey","Dewey","Louie"))
Huey Dewey Louie>
```

This prints the strings without quotes, just separated by a space character. There is no newline following the string, so the prompt (>) for the next line of input follows directly at the end of the line. (Notice that when the character vector is printed by `cat` there is no way of telling the difference from the single string `"Huey Dewey Louie"`.)

To get the system prompt onto the next line, you must include a newline character

```
> cat("Huey","Dewey","Louie", "\n")
Huey Dewey Louie
>
```

Here, `\n` is an example of an escape sequence. It actually represents a single character, the linefeed (LF), but is represented as two. The backslash (`\`) is known as the *escape character*. In a similar vein, you can insert quote characters with `\"`, as in

```
> cat("What is \"R\"?\n")
What is "R"?
```

There are also ways to insert other control characters and special glyphs, but it would lead us too far astray to discuss it in full detail. One important thing, though: What about the escape character itself? This, too, must be escaped, so to put a backslash in a string, you must double it. This is important to know when specifying file paths on Windows, see also Section 2.4.1.

1.2.5 Missing values

In practical data analysis, a data point is frequently unavailable (the patient did not show up, an experiment failed, etc.). Statistical software needs ways to deal with this. R allows vectors to contain a special `NA` value. This value is carried through in computations so that operations on `NA` yield `NA` as the result. There are some special issues associated with the handling of missing values; we deal with them as we encounter them (see "missing values" in the index).

1.2.6 Functions that create vectors

Here we introduce three functions, `c`, `seq`, and `rep`, that are used to create vectors in various situations.

The first of these, `c`, has already been introduced. It is short for "concatenate", joining items end to end, which is exactly what the function does:

```
> c(42,57,12,39,1,3,4)
[1] 42 57 12 39  1  3  4
```

You can also concatenate vectors of more than one element as in

```
> x <- c(1, 2, 3)
> y <- c(10, 20)
```

```
> c(x, y, 5)
[1]  1  2  3 10 20  5
```

However, you do not need to use `c` to create vectors of length 1. People sometimes type, for example, `c(1)`, but it is the same as plain `1`.

It is also possible to assign names to the elements. This modifies the way the vector is printed and is often used for display purposes.

```
> x <- c(red="Huey", blue="Dewey", green="Louie")
> x
    red    blue   green
 "Huey" "Dewey" "Louie"
```

(In this case, it *does* of course make sense to use `c` even for single-element vectors.)

The names can be extracted or set using `names`:

```
> names(x)
[1] "red"   "blue"  "green"
```

All elements of a vector have the same type. If you concatenate vectors of different types, they will be converted to the least "restrictive" type:

```
> c(FALSE, 3)
[1] 0 3
> c(pi, "abc")
[1] "3.14159265358979" "abc"
> c(FALSE, "abc")
[1] "FALSE" "abc"
```

That is, logical values may be converted to 0/1 or `"FALSE"`/`"TRUE"` and numbers converted to their printed representations.

The second function, `seq` ("sequence"), is used for equidistant series of numbers. Writing

```
> seq(4,9)
[1] 4 5 6 7 8 9
```

yields, as shown, the integers from 4 to 9. If you want a sequence in jumps of 2, write

```
> seq(4,10,2)
[1]  4  6  8 10
```

This kind of vector is frequently needed, particularly for graphics. For example, we previously used `c(1.65,1.70,1.75,1.80,1.85,1.90)` to define the x-coordinates for a curve, something that could also have been

written `seq(1.65,1.90,0.05)` (the advantage of using `seq` might have been more obvious if the heights had been in steps of 1 cm rather than 5 cm!).

The case with step size equal to 1 can also be written using a special syntax:

```
> 4:9
[1] 4 5 6 7 8 9
```

The above is exactly the same as `seq(4,9)`, only easier to read.

The third function, `rep` ("replicate"), is used to generate repeated values. It is used in two variants, depending on whether the second argument is a vector or a single number:

```
> oops <- c(7,9,13)
> rep(oops,3)
[1]  7  9 13  7  9 13  7  9 13
> rep(oops,1:3)
[1]  7  9  9 13 13 13
```

The first of the function calls above repeats the entire vector `oops` three times. The second call has the number 3 replaced by a vector with the three values (1, 2, 3); these values correspond to the elements of the `oops` vector, indicating that 7 should be repeated once, 9 twice, and 13 three times. The `rep` function is often used for things such as group codes: If it is known that the first 10 observations are men and the last 15 are women, you can use

```
> rep(1:2,c(10,15))
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

to form a vector that for each observation indicates whether it is from a man or a woman.

The special case where there are equally many replications of each value can be obtained using the `each` argument. E.g., `rep(1:2,each=10)` is the same as `rep(1:2,c(10,10))`.

1.2.7 Matrices and arrays

A *matrix* in mathematics is just a two-dimensional array of numbers. Matrices are used for many purposes in theoretical and practical statistics, but it is not assumed that the reader is familiar with matrix algebra, so many special operations on matrices, including matrix multiplication, are skipped. (The document "An Introduction to R", which comes with

the installation, outlines these items quite well.) However, matrices and also higher-dimensional arrays do get used for simpler purposes as well, mainly to hold tables, so an elementary description is in order.

In R, the matrix notion is extended to elements of any type, so you could have, for instance, a matrix of character strings. Matrices and arrays are represented as vectors with dimensions:

```
> x <- 1:12
> dim(x) <- c(3,4)
> x
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

The `dim` assignment function sets or changes the *dimension attribute* of `x`, causing R to treat the vector of 12 numbers as a 3×4 matrix. Notice that the storage is column-major; that is, the elements of the first column are followed by those of the second, etc.

A convenient way to create matrices is to use the `matrix` function:

```
> matrix(1:12,nrow=3,byrow=T)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
```

Notice how the `byrow=T` switch causes the matrix to be filled in a rowwise fashion rather than columnwise.

Useful functions that operate on matrices include `rownames`, `colnames`, and the transposition function `t` (notice the lowercase `t` as opposed to uppercase `T` for `TRUE`), which turns rows into columns and vice versa:

```
> x <- matrix(1:12,nrow=3,byrow=T)
> rownames(x) <- LETTERS[1:3]
> x
  [,1] [,2] [,3] [,4]
A    1    2    3    4
B    5    6    7    8
C    9   10   11   12
> t(x)
     A B  C
[1,] 1 5  9
[2,] 2 6 10
[3,] 3 7 11
[4,] 4 8 12
```

The character vector `LETTERS` is a built-in variable that contains the capital letters A–Z. Similar useful vectors are `letters`, `month.name`, and `month.abb` with lowercase letters, month names, and abbreviated month names.

You can "glue" vectors together, columnwise or rowwise, using the `cbind` and `rbind` functions.

```
> cbind(A=1:4,B=5:8,C=9:12)
     A B  C
[1,] 1 5  9
[2,] 2 6 10
[3,] 3 7 11
[4,] 4 8 12
> rbind(A=1:4,B=5:8,C=9:12)
  [,1] [,2] [,3] [,4]
A    1    2    3    4
B    5    6    7    8
C    9   10   11   12
```

We return to table operations in Section 4.5, which discusses tabulation of variables in a data set.

1.2.8 Factors

It is common in statistical data to have categorical variables, indicating some subdivision of data, such as social class, primary diagnosis, tumor stage, Tanner stage of puberty, etc. Typically, these are input using a numeric code.

Such variables should be specified as *factors* in R. This is a data structure that (among other things) makes it possible to assign meaningful names to the categories.

There are analyses where it is essential for R to be able to distinguish between categorical codes and variables whose values have a direct numerical meaning (see Chapter 7).

The terminology is that a factor has a set of *levels* — say four levels for concreteness. Internally, a four-level factor consists of two items: (a) a vector of integers between 1 and 4 and (b) a character vector of length 4 containing strings describing what the four levels are. Let us look at an example:

```
> pain <- c(0,3,2,2,1)
> fpain <- factor(pain,levels=0:3)
> levels(fpain) <- c("none","mild","medium","severe")
```

The first command creates a numeric vector `pain`, encoding the pain levels of five patients. We wish to treat this as a categorical variable, so we create a factor `fpain` from it using the function `factor`. This is called with one argument in addition to `pain`, namely `levels=0:3`, which indicates that the *input* coding uses the values 0–3. The latter can in principle be left out since R by default uses the values in `pain`, suitably sorted, but it is a good habit to retain it; see below. The effect of the final line is that the level names are changed to the four specified character strings.

The result should be apparent from the following:

```
> fpain
[1] none   severe medium medium mild
Levels:  none mild medium severe
> as.numeric(fpain)
[1] 1 4 3 3 2
> levels(fpain)
[1] "none"   "mild"   "medium" "severe"
```

The function `as.numeric` extracts the numerical coding as numbers 1–4 and `levels` extracts the names of the levels. Notice that the original input coding in terms of numbers 0–3 has disappeared; the internal representation of a factor always uses numbers starting at 1.

R also allows you to create a special kind of factor in which the levels are ordered. This is done using the `ordered` function, which works similarly to `factor`. These are potentially useful in that they distinguish nominal and ordinal variables from each other (and arguably `text.pain` above ought to have been an ordered factor). Unfortunately, R defaults to treating the levels as if they were *equidistant* in the modelling code (by generating polynomial contrasts), so it may be better to ignore ordered factors at this stage.

1.2.9 Lists

It is sometimes useful to combine a collection of objects into a larger composite object. This can be done using *lists*.

You can construct a list from its components with the function `list`.

As an example, consider a set of data from Altman (1991, p. 183) concerning pre- and postmenstrual energy intake in a group of women. We can place these data in two vectors as follows:

```
> intake.pre <- c(5260,5470,5640,6180,6390,
+ 6515,6805,7515,7515,8230,8770)
> intake.post <- c(3910,4220,3885,5160,5645,
+ 4680,5265,5975,6790,6900,7335)
```

Notice how input lines can be broken and continue on the next line. If you press the Enter key while an expression is syntactically incomplete, R will assume that the expression continues on the next line and will change its normal > prompt to the *continuation prompt* +. This often happens inadvertently due to a forgotten parenthesis or a similar problem; in such cases, either complete the expression on the next line or press ESC (Windows and Macintosh) or Ctrl-C (Unix). The "Stop" button can also be used under Windows.

To combine these individual vectors into a list, you can say

```
> mylist <- list(before=intake.pre,after=intake.post)
> mylist
$before
 [1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770

$after
 [1] 3910 4220 3885 5160 5645 4680 5265 5975 6790 6900 7335
```

The components of the list are named according to the argument names used in `list`. Named components may be extracted like this:

```
> mylist$before
 [1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770
```

Many of R's built-in functions compute more than a single vector of values and return their results in the form of a list.

1.2.10 Data frames

A data frame corresponds to what other statistical packages call a "data matrix" or a "data set". It is a list of vectors and/or factors of the same length that are related "across" such that data in the same position come from the same experimental unit (subject, animal, etc.). In addition, it has a unique set of row names.

You can create data frames from preexisting variables:

```
> d <- data.frame(intake.pre,intake.post)
> d
   intake.pre intake.post
1        5260        3910
2        5470        4220
3        5640        3885
4        6180        5160
5        6390        5645
6        6515        4680
7        6805        5265
```

```
8        7515        5975
9        7515        6790
10       8230        6900
11       8770        7335
```

Notice that these data are paired, that is, the same woman has an intake of 5260 kJ premenstrually and 3910 kJ postmenstrually.

As with lists, components (i.e., individual variables) can be accessed using the `$` notation:

```
> d$intake.pre
 [1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770
```

1.2.11 Indexing

If you need a particular element in a vector, for instance the premenstrual energy intake for woman no. 5, you can do

```
> intake.pre[5]
[1] 6390
```

The brackets are used for selection of data, also known as *indexing* or *subsetting*. This also works on the left-hand side of an assignment (so that you can say, for instance, `intake.pre[5] <- 6390`) if you want to modify elements of a vector.

If you want a subvector consisting of data for more than one woman, for instance nos. 3, 5, and 7, you can index with a vector:

```
> intake.pre[c(3,5,7)]
[1] 5640 6390 6805
```

Note that it is necessary to use the `c(...)`-construction to define the vector consisting of the three numbers 3, 5, and 7. `intake.pre[3,5,7]` would mean something completely different. It would specify indexing into a three-dimensional array.

Of course, indexing with a vector also works if the index vector is stored in a variable. This is useful when you need to index several variables in the same way.

```
> v <- c(3,5,7)
> intake.pre[v]
[1] 5640 6390 6805
```

It is also worth noting that to get a sequence of elements, for instance the first five, you can use the `a:b` notation:

```
> intake.pre[1:5]
[1] 5260 5470 5640 6180 6390
```

A neat feature of R is the possibility of negative indexing. You can get all observations *except* nos. 3, 5, and 7 by writing

```
> intake.pre[-c(3,5,7)]
[1] 5260 5470 6180 6515 7515 7515 8230 8770
```

It is not possible to mix positive and negative indices. That would be highly ambiguous.

1.2.12 Conditional selection

We saw in Section 1.2.11 how to extract data using one or several indices. In practice, you often need to extract data that satisfy certain criteria, such as data from the males or the prepubertal or those with chronic diseases, etc. This can be done simply by inserting a relational expression instead of the index,

```
> intake.post[intake.pre > 7000]
[1] 5975 6790 6900 7335
```

yielding the postmenstrual energy intake for the four women who had an energy intake above 7000 kJ premenstrually.

Of course, this kind of expression makes sense only if the variables that go into the relational expression have the same length as the variable being indexed.

The comparison operators available are < (less than), > (greater than), == (equal to), <= (less than or equal to), >= (greater than or equal to), and != (not equal to). Notice that a double equal sign is used for testing equality. This is to avoid confusion with the = symbol used to match keywords with function arguments. Also, the != operator is new to some; the ! symbol indicates negation. The same operators are used in the C programming language.

To combine several expressions, you can use the logical operators & (logical "and"), | (logical "or"), and ! (logical "not"). For instance, we find the postmenstrual intake for women with a premenstrual intake between 7000 and 8000 kJ with

```
> intake.post[intake.pre > 7000 & intake.pre <= 8000]
[1] 5975 6790
```

There are also `&&` and `||`, which are used for flow control in R programming. However, their use is beyond what we discuss here.

It may be worth taking a closer look at what actually happens when you use a logical expression as an index. The result of the logical expression is a logical vector as described in Section 1.2.3:

```
> intake.pre > 7000 & intake.pre <= 8000
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  FALSE
[11] FALSE
```

Indexing with a logical vector implies that you pick out the values where the logical vector is `TRUE`, so in the preceding example we got the 8th and 9th values in `intake.post`.

If missing values (`NA`; see Section 1.2.5) appear in an indexing vector, then R will create the corresponding elements in the result but set the values to `NA`.

In addition to the relational and logical operators, there are a series of functions that return a logical value. A particularly important one is `is.na(x)`, which is used to find out which elements of `x` are recorded as missing (`NA`).

Notice that there is a real need for `is.na` because you cannot make comparisons of the form `x==NA`. That simply gives `NA` as the result for any value of `x`. The result of a comparison with an unknown value is unknown!

1.2.13 Indexing of data frames

We have already seen how it is possible to extract variables from a data frame by typing, for example, `d$intake.post`. However, it is also possible to use a notation that uses the matrix-like structure directly:

```
> d <- data.frame(intake.pre,intake.post)
> d[5,1]
[1] 6390
```

gives fifth row, first column (that is, the "pre" measurement for woman no. 5), and

```
> d[5,]
  intake.pre intake.post
5       6390        5645
```

gives *all* measurements for woman no. 5. Notice that the comma in `d[5,]` is required; without the comma, for example `d[2]`, you get the data frame

consisting of the second *column* of `d` (that is, more like `d[,2]`, which is the column itself).

Other indexing techniques also apply. In particular, it can be useful to extract all data for cases that satisfy some criterion, such as women with a premenstrual intake above 7000 kJ:

```
> d[d$intake.pre>7000,]
   intake.pre intake.post
8        7515        5975
9        7515        6790
10       8230        6900
11       8770        7335
```

Here we extracted the rows of the data frame where `intake.pre>7000`. Notice that the row names are those of the original data frame.

If you want to understand the details of this, it may be a little easier if it is divided into smaller steps. It could also have been done like this:

```
> sel <- d$intake.pre>7000
> sel
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
[11]  TRUE
> d[sel,]
   intake.pre intake.post
8        7515        5975
9        7515        6790
10       8230        6900
11       8770        7335
```

What happens is that `sel` *(select)* becomes a logical vector with the value `TRUE` for to the four women consuming more than 7000 kJ premenstrually. Indexing as `d[sel,]` yields data from the rows where `sel` is `TRUE` and from all columns because of the empty field after the comma.

It is often convenient to look at the first few cases in a data set. This can be done with indexing, like this:

```
> d[1:2,]
   intake.pre intake.post
1        5260        3910
2        5470        4220
```

This is such a frequent occurrence that a convenience function called `head` exists. By default, it shows the first six lines.

```
> head(d)
   intake.pre intake.post
1        5260        3910
2        5470        4220
```

```
3        5640        3885
4        6180        5160
5        6390        5645
6        6515        4680
```

Similarly, `tail` shows the last part.

1.2.14 *Grouped data and data frames*

The natural way of storing grouped data in a data frame is to have the data themselves in one vector and parallel to that have a factor telling which data are from which group. Consider, for instance, the following data set on energy expenditure for lean and obese women.

```
> energy
   expend stature
1    9.21   obese
2    7.53    lean
3    7.48    lean
4    8.08    lean
5    8.09    lean
6   10.15    lean
7    8.40    lean
8   10.88    lean
9    6.13    lean
10   7.90    lean
11  11.51   obese
12  12.79   obese
13   7.05    lean
14  11.85   obese
15   9.97   obese
16   7.48    lean
17   8.79   obese
18   9.69   obese
19   9.68   obese
20   7.58    lean
21   9.19   obese
22   8.11    lean
```

This is a convenient format since it generalizes easily to data classified by multiple criteria. However, sometimes it is desirable to have data in a separate vector for each group. Fortunately, it is easy to extract these from the data frame:

```
> exp.lean <- energy$expend[energy$stature=="lean"]
> exp.obese <- energy$expend[energy$stature=="obese"]
```

Alternatively, you can use the `split` function, which generates a list of vectors according to a grouping.

```
> l <- split(energy$expend, energy$stature)
> l
$lean
 [1]  7.53  7.48  8.08  8.09 10.15  8.40 10.88  6.13  7.90  7.05
[11]  7.48  7.58  8.11

$obese
[1]  9.21 11.51 12.79 11.85  9.97  8.79  9.69  9.68  9.19
```

1.2.15 *Implicit loops*

The looping constructs of R are described in Section 2.3.1. For the purposes of this book, you can largely ignore their existence. However, there is a group of R functions that it will be useful for you to know about.

A common application of loops is to apply a function to each element of a set of values or vectors and collect the results in a single structure. In R this is abstracted by the functions `lapply` and `sapply`. The former always returns a list (hence the 'l'), whereas the latter tries to simplify (hence the 's') the result to a vector or a matrix if possible. So, to compute the mean of each variable in a data frame of numeric vectors, you can do the following:

```
> lapply(thuesen, mean, na.rm=T)
$blood.glucose
[1] 10.3

$short.velocity
[1] 1.325652

> sapply(thuesen, mean, na.rm=T)
 blood.glucose short.velocity
     10.300000       1.325652
```

Notice how both forms attach meaningful names to the result, which is another good reason to prefer to use these functions rather than explicit loops. The second argument to `lapply/sapply` is the function that should be applied, here `mean`. Any further arguments are passed on to the function; in this case we pass `na.rm=T` to request that missing values be removed (see Section 4.1).

Sometimes you just want to repeat something a number of times but still collect the results as a vector. Obviously, this makes sense only when the repeated computations actually give different results, the common case being simulation studies. This can be done using `sapply`, but there is a simplified version called `replicate`, in which you just have to give a count and the expression to evaluate:

```
> replicate(10,mean(rexp(20)))
 [1] 1.0677019 1.2166898 0.8923216 1.1281207 0.9636017 0.8406877
 [7] 1.3357814 0.8249408 0.9488707 0.5724575
```

A similar function, `apply`, allows you to apply a function to the rows or columns of a matrix (or over indices of a multidimensional array in general) as in

```
> m <- matrix(rnorm(12),4)
> m
           [,1]       [,2]       [,3]
[1,] -2.5710730 0.2524470 -0.16886795
[2,]  0.5509498 1.5430648  0.05359794
[3,]  2.4002722 0.1624704 -1.23407417
[4,]  1.4791103 0.9484525 -0.84670929
> apply(m, 2, min)
[1] -2.5710730  0.1624704 -1.2340742
```

The second argument is the index (or vector of indices) that defines what the function is applied to; in this case we get the columnwise minima.

Also, the function `tapply` allows you to create tables (hence the 't') of the value of a function on subgroups defined by its second argument, which can be a factor or a list of factors. In the latter case a cross-classified table is generated. (The grouping can also be defined by ordinary vectors. They will be converted to factors internally.)

```
> tapply(energy$expend, energy$stature, median)
 lean obese
 7.90  9.69
```

1.2.16 Sorting

It is trivial to sort a vector. Just use the `sort` function. (We use the built-in data set `intake` here; it contains the same data that were used in Section 1.2.9.)

```
> intake$post
 [1] 3910 4220 3885 5160 5645 4680 5265 5975 6790 6900 7335
> sort(intake$post)
 [1] 3885 3910 4220 4680 5160 5265 5645 5975 6790 6900 7335
```

(`intake$pre` could not be used for this example since it is sorted already!)

However, sorting a single vector is not always what is required. Often you need to sort a series of variables according to the values of some *other* variables — blood pressures sorted by sex and age, for instance. For this

purpose, there is a construction that may look somewhat abstract at first but is really very powerful. You first compute an *ordering* of a variable.

```
> order(intake$post)
 [1]  3  1  2  6  4  7  5  8  9 10 11
```

The result is the numbers 1 to 11 (or whatever the length of the vector is), sorted according to the size of the argument to `order` (here `intake$post`). Interpreting the result of `order` is a bit tricky — it should be read as follows: You sort `intake$post` by placing its values in the order no. 3, no. 1, no. 2, no. 6, etc.

The point is that, by indexing with this vector, other variables can be sorted by the same criterion. Note that indexing with a vector containing the numbers from 1 to the number of elements exactly once corresponds to a reordering of the elements.

```
> o <- order(intake$post)
> intake$post[o]
 [1] 3885 3910 4220 4680 5160 5265 5645 5975 6790 6900 7335
> intake$pre[o]
 [1] 5640 5260 5470 6515 6180 6805 6390 7515 7515 8230 8770
```

What has happened here is that `intake$post` has been sorted — just as in `sort(intake$post)` — while `intake$pre` has been sorted by the size of the corresponding `intake$post`.

It is of course also possible to sort the entire data frame `intake`

```
> intake.sorted <- intake[o,]
```

Sorting by several criteria is done simply by having several arguments to `order`; for instance, `order(sex,age)` will give a main division into men and women, and within each sex an ordering by age. The second variable is used when the order cannot be decided from the first variable. Sorting in reverse order can be handled by, for example, changing the sign of the variable.

1.3 Exercises

1.1 How would you check whether two vectors are the same if they may contain missing (`NA`) values? (Use of the `identical` function is considered cheating!)

1.2 If `x` is a factor with `n` levels and `y` is a length `n` vector, what happens if you compute `y[x]`?

1.3 Write the logical expression to use to extract girls between 7 and 14 years of age in the `juul` data set.

1.4 What happens if you change the levels of a factor (with `levels`) and give the same value to two or more levels?

1.5 On p. 27, `replicate` was used to simulate the distribution of the mean of 20 random numbers from the exponential distribution by repeating the operation 10 times. How would you do the same thing with `sapply`?