Statistics with R

Hugo Quené

2020-04-11

Contents

Preface

This booklet is written as accompaniment to the online tutorial on *Statistics with R (Basics)*, to be held as part of the workshop on Experimental Methods in Language Acquisition Research (EMLAR, https://emlar.wp.hum.uu.nl/), Utrecht, on 17 April 2020.

Here's the abstract for the tutorial:

> This tutorial will introduce the R programming environment for statistical analysis. Contrary to SPSS which is procedure-oriented (commands are verbs, e.g. "compute"), R is object-oriented (objects are nouns, e.g. "factor"). In this workshop, we will try to ease the learning curve of using R for your data analysis. Experience with statistical software is NOT required! We will use data simulation as well as real data sets, to explore topics like t-tests, χ^2 tests, and regression. We will also show how R produces publication-quality figures.

Hugo Quené, Utrecht institute of Linguistics OTS, Utrecht University

https://www.hugoquene.nl

Chapter 1

Introduction

This tutorial offers a first introduction into R. R is available as freeware from http://www.r-project.org, where one can also find a wealth of information and documentation.

This tutorial assumes that R is already properly installed on your computer. It is further assumed that the reader has some basic knowledge about statistics, equivalent to an introductory course in statistics. This tutorial introduces the R software for statistical analyses, and not the statistical analyses themselves. This tutorial occasionally mentions differences with SPSS, but the tutorial is also intended for novice users of statistical software.

1.1 What is R ?

Perhaps surprisingly, R is several things at once:

- a program for statistical analyses

```
firstmodel.lm <- lm( syldur~age, data=talkers ) # linear-model regression
```

- a calculator

```
# 440 Hz is `how many` semitones above 110 Hz?
(log(440)-log(110)) / log(2^(1/12))
```

```
## [1] 24
```

- a programming language

```
# function to convert hertz to semitones, relative to `base`, by Mark Liberman
h2st <- function( h, base=50 ) {
  semi1<-log(2^(1/12)); # log of frequency ratio of 1 semitone
  return( (log(h)-log(base)) / semi1 ) }
```

The assignment operator `<-` is further explained in section 3.1 below.

The hash `#` indicates comment, which is not processed.

1.2 Object-oriented philosophy

R works in an object-oriented way. This means that *objects* are the most important things in R , and *not* the actions we perform with these objects. Let's use a culinary example to illustrate this. In order to obtain pancakes, a cook needs flour, milk, eggs, some mixing utensils, a pan, oil, and a fire. An object-oriented approach places primary focus on these six objects. If the relations between these are properly specified, then a good pancake will result. Provided that the necessary objects (ingredients) are available, the R syntax could be as follows:

```
# batter <- mixed( flour,  milk/2 ) # mix flour and half of milk
# batter <- mixed( batter, egg*2 ) # add 2 eggs
# batter <- mixed( batter, milk/2, use=whisk) # add other half of milk
# while (enough(batter)) # FALSE if insufficient for next
#  pancake <- baked( batter, in=oil, with=pan, temp=max(fire) )
```

This example illustrates that R is indeed a full programming language (but see footnote [1]). In fact, there is no recipe, in the traditional sense. This "pancake" script merely specifies the relations between the ingredients and the result. Note that some relations are recursive: batter can be both input and output of the mixing operation. Also note that the `mixed` relation takes an optional argument `use=whisk`, which will produce a fatal error message if there is no whisk in the kitchen. Such arguments, however, allow for greater flexibility of the `mixed` relation. Likewise, we might specify `baked(in=grease)` if there is no oil in the kitchen. The only requirement for the object supplied as `in` argument is that one can bake in it, so this object must have some attribute `goodforbaking==TRUE`.

For contrast, we might imagine how the pancake recipe would be formulated in a more traditional, procedure-oriented approach. Ingredients and a spoon are again assumed to be provided.

```
MIX batter = flour + milk/2 . # what utensil?
MIX batter = batter + eggs .
```

[1] Technically speaking, R is an interpreted language, and not a compiled language

```
MIX batter = batter + milk/2 .
BAKE batter IN oil .
BAKE batter IN water . # garbage in garbage out
```

The programmer of this recipe has defined the key procedures `MIX` and `BAKE`, and has stipulated boundary conditions such as utensils and temperatures. Optional arguments are allowed for the `BAKE` command, but only within the limits set by the programmer (see footnote 2).

So far, you may have thought that the difference between the two recipes was semantic rather than pragmatic. To demonstrate the greater flexibility of an object-oriented approach, let us consider the following variant of the recipe, again in R syntax:

```
# batter is done
while (number(pancakes)<2) # first bake 2 pancakes
    pancake <- baked(batter,in=oil,with=pan,temp=max(fire))
feed(pancake,child) # feed one to hungry spectator
# define new function, data 'x' split into 'n' pieces
chopped <- function(x,n=1000) { return(split(x,n)) }
pieces <- chopped(pancake) # new data object, array of 1000 pieces
batter <- mixed(batter,pieces) # mix pancake pieces into batter
# etc
```

Such complex relations between objects are quite difficult to specify, if there are strong a priori limits to what one can `MIX` or `BAKE`. Thus, object-oriented programs such as R allow for greater flexibility than procedure-oriented programs.

If you are a user of the `Praat` software (http://www.praat.org) then you are already familiar with this basic idea. `Praat` has an object window, listing the known objects. These objects are the output of previous operations (e.g. Create, Read, ToSpectrum), as well as input for subsequent operations (e.g. Write, Draw). The classes or types of these objects are pre-defined (Sound, Spectrum, Periodicity, etc). R takes the same idea even further: users may create their own *classes* of data objects (e.g. class `SuperData`) and may create their own methods or relations to work with such objects [2].

This object-oriented philosophy results in a different behavior than observed in procedure-oriented software:

> There is an important difference in philosophy between S (and hence R) and the other main statistical systems. In S a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give copious output from a regression or discriminant analysis, R will give

[2] Praat allows the latter but not the former.

minimal output and store the results in a fit object for subsequent interrogation by further R functions.

from: http://cran.r-project.org/doc/manuals/R-intro.html

Chapter 2

Objects

2.1 Vectors

A vector is a simple, one-dimensional list of data, like a single column in Excel or in SPSS. Typically a single vector holds a single variable of interest. The data in a vector can be of various classes: numeric, character (strings of letters, always enclosed in double quotes), or logical (i.e., boolean, `TRUE` or `FALSE`, may be abbreviated to `T` or `F`).

- `c`: Atomic data are combined into a vector by means of the `c` (combine, concatenate) operator.
- `seq` The sequence operator, also abbreviated as a colon `:`, creates subsequent values.

```
x <- 1:5
print(x)
```

```
## [1] 1 2 3 4 5
```

```
2*(x-1)
```

```
## [1] 0 2 4 6 8
```

Computations are also done on whole vectors, as exemplified above. In the last example, we see that the result of the computation is *not* assigned to a new object. Hence the result is displayed — and then lost. This may still be useful however when you use R as a pocket calculator.

- `rep` Finally, the repeat operator is very useful in creating repetitive sequences, e.g. for levels of an independent variable.

```
x <- rep( 1:5, each=2 )
x
```

```
##  [1] 1 1 2 2 3 3 4 4 5 5
```

2.2 Factors

Factors constitute a special class of variables. A factor is a variable that holds categorical, character-like data. R realizes that variables of this class hold categorical data, and that the values are category labels or *levels* rather than real characters or digits, as illustrated in the examples below.

```
x1 <- rep( 1:4, each=2 ) # create vector of numbers
print(x1) # numeric
```

```
## [1] 1 1 2 2 3 3 4 4
```

```
summary(x1) # of numeric vector
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.75    2.50    2.50    3.25    4.00
```

```
x2 <- as.character(x1) # convert to char
print(x2) # character vector
```

```
## [1] "1" "1" "2" "2" "3" "3" "4" "4"
```

```
x3 <- as.factor(x1) # convert to factor
print(x3) # factor
```

```
## [1] 1 1 2 2 3 3 4 4
## Levels: 1 2 3 4
```

```
summary(x3) # compare against summary(x1) above
```

```
## 1 2 3 4
## 2 2 2 2
```

2.3 Complex objects

Simple objects, like the ones introduced above, may be combined into composite objects. For example, we may combine all pancake ingredients into a complex object of class `list`.

In R we often use a particular complex object, a *data frame*, to hold various data together. A data frame is a complex object like an Excel worksheet or SPSS data sheet. The columns represent variables, and the rows represent single observations — these may be "cases" or sampling units, or single measurements repeated for each sampling unit, depending on the study [1].

The easiest way to create a data object is to read it from a plain-text (ASCII) file, using the command `read.table`. (Windows users must remember to use double backslashes in the file specification string). An optional `header=TRUE` argument indicates whether the first line contains the names of the variables; argument `sep` specifies the character(s) that separate the variables in the input file. The `file` argument can be a string specifying a local file, or a `url` to a web-based file, or a call of function `file.choose()` to select a file interactively. Argument `na.strings` specifies the character string(s) that indicate missing values in the input file.

```
# in Windows system
myexp <- read.table(
  file="f:\\temp\\mydata.txt", header=T, sep="," )
```

```
nlspkr <- read.table(
  file=url("http://www.hugoquene.nl/emlar/intra.bysubj.txt"),
  header=TRUE, na.strings=c("NA","MISSING") )
```

It is also possible to read so-called CSV files (comma-separated values) saved from Excel or SPSS (`read.csv`), and it is also possible to read Excel or SPSS data files directly using extension packages (`readxl::readxl`, `foreign::read.spss`, see Chapter 8).

The basic R and extension packages already have many datasets pre-defined, for immediate use. To see a long(!) overview of these datasets, enter the command `data()`.

[1] For repeated measures analyses, R does not require a multivariate or "wide" layout, with repeated measures for each participant on a single row, as SPSS does. Instead R always uses a univariate or "long" layout, with each measurement on a single row of input. See the `reshape` command to convert between layouts, section @**??**sec:split.merge.reshape).

Chapter 3

Basic operations

3.1 Basics

3.1.1 `<-`

This is the assignment operator: the expression to its right is evaluated (if applicable) and then assigned to the object on the left of the operator. Hence the expression `a<-10` means that the object `a`, a single number, "gets" the value of 10, i.e. the value of 10 is assigned to `a`. The symbol resembles an arrow in the direction of assignment. The assignment may also be in the other direction, with symbol `->` (and see note [1]). There should be no space between the two characters making up the arrow. Use spaces or brackets to avoid ambiguities and errors:

```
x <- 10 # assignment of atomic value 10 to object x
x < -10 # is value of x less than -10 ?
```

```
## [1] FALSE
```

3.1.2 `#`

indicates a comment: everything following this symbol, on the same line of input, is ignored.

[1] The equal sign `=` is also available for assignment. Using it is somewhat dangerous, however, because the equal sign does not specify the direction of assignment explicitly.

3.1.3 `scan`

This command reads a simple vector from the keyboard. Make sure to assign the result to a new object! Read in the numbers 1 to 10, and assign them to a new object.

3.1.4 `objects`

This command shows a list of all objects in memory (similar to the contents of the `Praat` Objects window). With `objects(pattern="abc")` the list is filtered so that only the objects matching the pattern string `"abc"` are shown.

3.1.5 `rm`

Objects are removed *forever* with this command.

3.1.6 `print`

Contents of an object can be inspected with this command, or by just entering the name of the object, as in some examples above.

3.1.7 `summary`

This command offers a summary of an object. The result depends on the data class of the object, as illustrated in section REF above.

3.1.8 Workspace:

R holds its objects in memory. The whole workspace, containing all data objects, can be stored from the Rstudio window (`Session > Save Workspace As...`). This allows you to save a session, and continue your work later (`Session > Load Workspace...`).

3.1.9 `save`

(to write) and

3.1.10 `read`

(to read) an object from/to memory to hard disk. By default, R data objects have the extension `.Rda`.

3.1.11 No undo

Remember that there is *no* `undo` command, nor such a menu option. Save your work regularly. If in doubt, work with scratch copies of your data sets.

3.2 Subselection

Subselection within an object is a very powerful tool in R. The subselection operator `x[…]` selects only those data from object `x` that match the expression within square brackets. This expression can be a single index number, a sequence or list of numbers, or an evaluated expression, as illustrated in the following example.

In the following example, variable `x` contains 30 numbers, but 3 of these are `NA`. Notice that the output of `is.na` is the input of `table`.

```
# is.na() returns TRUE/FALSE for each element of 'x'.
# table() summarizes categorical data
table( is.na(x) )
```

```
##
## FALSE  TRUE
##    27     3
```

```
ok <- !is.na( x ) # exclamation mark means NOT
which( !ok ) # which index numbers are NOT ok? inspect!
```

```
## [1] 11 13 19
```

```
mean( x[ok] ) # select ok values, compute mean, display
```

```
## [1] 1.015252
```

Subselection can also be achieved by using the function `subset(data, subset, select)`. The first argument is the input data (set), the second argument is the selector condition, and the optional third argument indicates which columns of a data frame should be kept in the output.

```
require(hqmisc)
data(talkers)
subset( talkers, subset=( age<45 & region=="W" ) )
```

```
##      id sex age region syldur  nsyl
## 1    60   1  38      W 0.1940 13.56
## 3    62   1  36      W 0.2331 11.73
## 4   112   1  33      W 0.2633 11.67
## 45  153   0  39      W 0.2676  6.36
## 50  158   1  40      W 0.2131  7.99
## 51  159   0  25      W 0.2152  8.11
## 52  160   0  26      W 0.2104  8.54
## 53  161   0  27      W 0.2459  8.89
## 55  163   0  33      W 0.2287  7.60
## 80  391   1  34      W 0.2225  8.89
```

This command selects rows from data frame `talkers` from the package `hqmisc` (see 8) corresponding to speakers who are under 45 years of age, and who are from the West region.

3.3 Split, merge, reshape

There are useful functions available to split and merge data frames. First we create two example data frames. The first data frame has a list of English vowels, with a phonological feature for each vowel, and with the average frequency of the second formant[2] of each vowel (Peterson and Barney, 1952) spoken by male speakers. The second data frame has a partially overlapping list of vowels, with key words by John Wells [3].

```
vowelsymb <- c( "i","I","e","E","ae", "A","V","o","U","u", "@" )
v1df <- data.frame( vowel=vowelsymb,
                    feat=factor( c(rep("front",5),
                                   rep("back",5),
                                   "central" ) ),
                    F2=c( 2290,1990,NA,1840,1720,
                          1090,1190,NA,1020,870,
                          NA ) )
v2df <- data.frame( vowel=vowelsymb[1:10],
                    word=c("fleece","kit","face","dress","trap",
                           "lot","strut","goat","foot","goose") )
```

3.3.1 `split`

This command divides the data in the first argument (column `F2` in data frame `v1df`) into the groups defined by the second argument.

[2]A formant is a resonant frequency of the vocal tract, counted from low to high. The second formant or F2 is related to the front–back articulatory dimension.

[3]See http://en.wikipedia.org/wiki/Lexical_set

```
with( v1df, split(F2,feat) ) -> v1list
```

This is particularly helpful in combination with `lapply` to apply a function to these grouped data, e.g. to compute the mean of F2 for each vowel category separately:

```
with(v1df, lapply( split(F2,feat), mean, na.rm=T ) )
```

```
## $back
## [1] 1042.5
##
## $central
## [1] NaN
##
## $front
## [1] 1960
```

Also see `unsplit`.

3.3.2 `merge`

This command merges two data frames, by common columns. Specify argument `all=TRUE` if you want to include non-matching rows in the output, with NA's in the appropriate columns.

```
m1 <- merge(v1df, v2df)
```

The resulting merged data frame is also sorted on the common columns, unless argument `sort=FALSE`.

3.3.3 `reshape`

If you need to perform a Repeated Measures (within-subjects) analysis of variance (RM-ANOVA) in SPSS, your data have to be in "wide" data layout, with all observations from one subject on a single data line. R on the other hand uses the "long" data layout, with one observation per line, and with all descriptors of that observation repeated for each line. There is a convenient command `reshape` to convert data between the wide layout (of SPSS RM-ANOVA) and the long layout (of R). To illustrate, first we read a wide data set:

```
widedata <- read.table(
  file=url("http://www.hugoquene.nl/emlar/widedata.txt"),
  header=T)
```

The wide data show the subject id, between-subject group, and three within-subject observations, for 6 subjects (with leading row numbers):

```
head(widedata)
```

```
##   subject group item1 item2 item3
## 1       1     1     2     3     4
## 2       2     1     3     4     6
## 3       3     1     1     3     6
## 4       4     2     2     4     5
## 5       5     2     4     5     6
## 6       6     2     2     5     7
```

These data are then reshaped to long layout with the following command:

```
longdata <- reshape( widedata, direction="long",
                     varying=c("item1","item2","item3"),
                     timevar="item", times=c("1","2","3"),
                     v.names="resp", idvar="subject")
```

The observations from multiple columns `varying` are collected into a new single column `v.names`, using identifiers in column `idvar`. The information contained in the multiple column names of `varying` is stored in a new single column `timevar`, using the values `times`. Inspect the two data frames to verify this.

Chapter 4

Exploratory data analyses

R is more graphically oriented than most other statistical packages; it relies more on plots and figures for initial exploratory data analysis. Numerical summaries are of course also available.

4.1 `hist`

This command produces a histogram. There is a useful optional argument `breaks` to specify the number of bins (bars), or a vector of breaks between bins.

4.2 `plot`

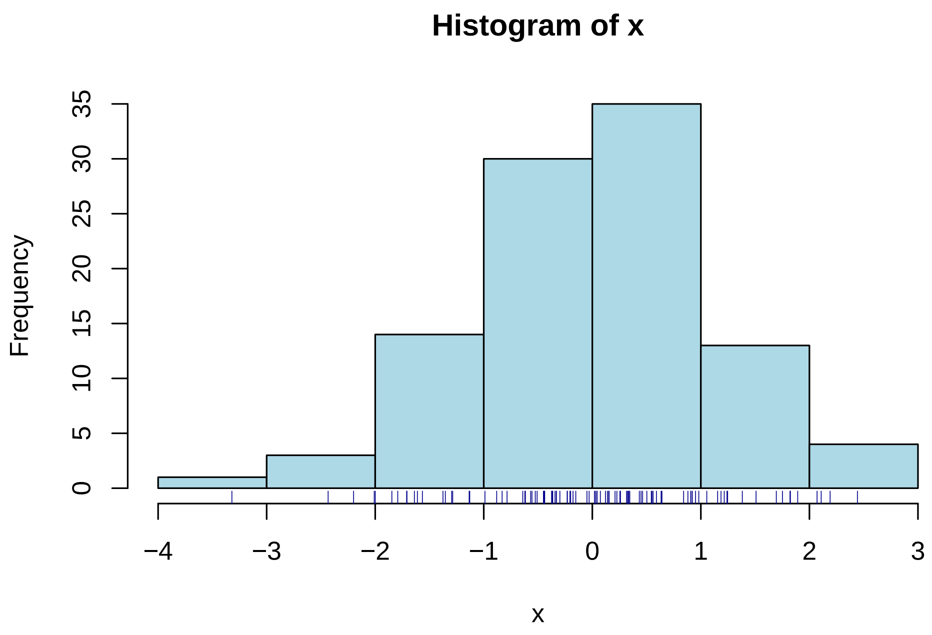
The default version of this command produces a scattergram of two variables. If you enter just one variable, then the index numbers of the observations are used on the horizontal axis, and the values on the vertical axis. Useful arguments are `title`, and `xlab` and `ylab` for axis labels. In addition, you can use a third variable to vary the plot symbols.

4.3 `rug`

This command produces tick marks at the actual data values, yielding the visual effect of a rug. This is useful in combination with a scattergram or histogram. Try it out, with the following commands [1]:

[1] The semicolon separates multiple R commands on a single line of input.

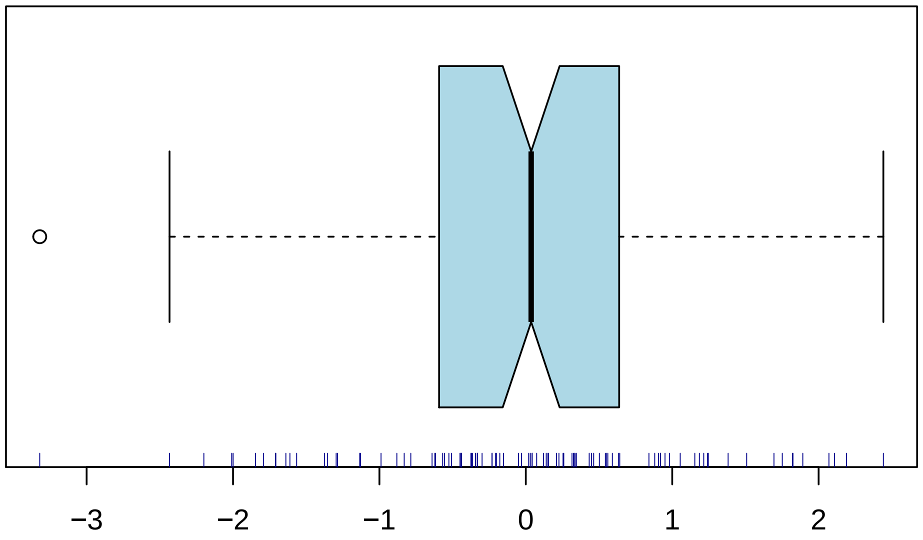
```
x<-rnorm(100); hist(x,col="lightblue"); rug(x,col="darkblue")
```



4.4 boxplot

This yields a boxplot summary (Tukey, 1977) of one variable. You can also specify the dependent and independent variable, with argument `dv~iv1`; this will produce multiple boxplots for the dependent variable, broken down by the independent variable(s). Two useful arguments for this command are:
`notch=T` to give additional information about the distribution, and
`varwidth=T` to scale the width of the boxes to the numbers of observations.

```
boxplot( x, col="lightblue", horizontal=TRUE, notch=TRUE, varwidth=TRUE )
rug(x, col="darkblue" )
```



4.5 `qqnorm`

This produces a quantile–quantile (QQ) plot. This plots the observed quantiles against the expected quantiles *if* the argument variable would be distributed normally. If the variable is indeed distributed normally, then the data should fall on an approximately straight line. Deviations of this line indicate deviations from normality. You can also add the expected line with `qqline`.

4.6 `summary`

This command produces a numerical summary of the argument variable. However it does not supply standard measures of variability. We often need

4.7 `var`

to compute the variance of the argument variable. Related functions are `sd` to compute standard deviation, `cov` to compute the covariance between two variables, and `cor` to compute their correlation.

4.8 length

returns the length of the argument variable, i.e. the number of observations in that vector. This is useful for checking the number of data, as a preliminary for further analyses.

```
valid.n <- function(x) {
  length(x)-length(which(is.na(x)))
  }
```

In the last command above, we have programmed a new function `valid.n`, using standard functions provided with R.

4.9 unique

returns a vector containing the (unsorted) unique values of the argument variable, without duplicates. This is also useful for checking your data.

```
x <- c(3,3,3,1,1,2); unique(x); sort(unique(x))
```

```
## [1] 3 1 2
```

```
## [1] 1 2 3
```

4.10 table

returns a frequency table, i.e. reports the number of observations for each (combination of) value(s) of one or more variables. This again is helpful for checking your data, to inspect the numbers of observations per cell (see also 3.2). Tables can be one-dimensional (see example below) or multidimensional. It is also possible to make a frequency table of a frequency table!

```
x <- rep(1:3,each=7); table(x) # three cells, each 7 observations, N=21
```

```
## x
## 1 2 3
## 7 7 7
```

```r
table( table(x) ) # ok: all three cells have 7 observations, N=21
```

```
##
## 7
## 3
```

A useful trick: if all cells should have the same number of observations, then the table-of-table should contain a single number representing the number of *cells*.

```r
x[ sample(1:3*7, size=2) ] <- NA # replace 2 out of 21 obs by NA
table( table(x) ) # not ok: not all three cells have 7 observations, N<21
```

```
##
## 6 7
## 2 1
```

The `lattice` package (see Chapter 8) contains additional functions for higher-order data visualization: `xyplot`, `histogram`, `densityplot`, `cloud`, and more. For more information, enter the command `help(lattice)`.

Now that we have obtained some insightful figures, we might want to include these in our documents. The best procedure in Rstudio is to go to the Plots tab in the lower right pane, and select `Export > Save as image...` or `Export > Save as PDF...`. Figures in PDF format (extension `pdf`) are easy to include in LaTeX documents, and are best for publication. Other formats (e.g. `png`) are easy to include in LaTeX, MS Word and web documents.

Chapter 5

Testing hypotheses

5.1 formula

When testing hypotheses, and building regression models, we need to specify the relations between variables. This is done in R by means of a *formula*, which is needed in many statistical functions. In general, such a formula consists of a response variable, followed by the tilde symbol `~`, followed by a list of independent variables and/or factors (Wilkinson and Rogers, 1973). In this list, the colon `:` indicates an interaction effect (instead of the sequence operator), and the asterisk `*` is shorthand for main effects plus interactions (instead of the multiplication operator). By default, the intercept `~1` is included in the formula, unless suppressed explicitly (`-1`). We have already encountered such a formula in the boxplot example above.

```
y ~ x1 + x2 # only main effects
y ~ x1 * x2 # shorthand for x1 + x2 + (x1:x2)
```

Consult the help files for further information on how to specify complex models.

5.2 t test

There are three ways to use the t test.

In a one-sample test, the mean is compared against an expected mean, with

```
t.test( x1, mu=0.80 )
```

```
##
##  One Sample t-test
##
## data:  x1
## t = 1.8691, df = 99, p-value = 0.06456
## alternative hypothesis: true mean is not equal to 0.8
## 95 percent confidence interval:
##  0.7878051 1.2082871
## sample estimates:
## mean of x
## 0.9980461
```

In a two-sample test with independent observations, we often compare the same dependent variable, broken down by an independent variable.

```
t.test( y[x1<median(x1)], y[x1>median(x1)] ) # y split up by median of x1
```

```
##
##  Welch Two Sample t-test
##
## data:  y[x1 < median(x1)] and y[x1 > median(x1)]
## t = -3.3044, df = 94.766, p-value = 0.001345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0476023 -0.2612324
## sample estimates:
## mean of x mean of y
##  2.749787  3.404204
```

This could also be achieved by specifying the dependent and independent variables in a formula:

```
t.test( y ~ (x1<median(x1)) ) # equivalent
```

```
##
##  Welch Two Sample t-test
##
## data:  y by x1 < median(x1)
## t = 3.3044, df = 94.766, p-value = 0.001345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2612324 1.0476023
## sample estimates:
```

```
## mean in group FALSE  mean in group TRUE
##             3.404204            2.749787
```

In a two-sample test with paired observations, we often compare two different observations, stored in two different variables.

```
t.test( x1, x2 )
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -6.3729, df = 197.68, p-value = 1.279e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2764242 -0.6731481
## sample estimates:
## mean of x mean of y
## 0.9980461 1.9728323
```

Note that the number of observations in the test (and hence d.f.) varies in these examples.

5.3 chisq.test

First, let us create two categorical variables, derived from a speaker's `age` (in years) and average `phraselength` (in syllables), for 80 speakers in the Corpus of Spoken Dutch (`talkers` data set; (Quené, 2014)). Categorical variables are created here with the `cut` function, to create `breaks=2` categories of `age` (young and old) and of `phraselength` (short and long).

```
require(hqmisc)
data(talkers)
age.cat <- cut( talkers$age, breaks=2 )
phraselength.cat <- cut( talkers$nsyl, breaks=2 )
```

The hypothesis under study is that older speakers tend to produce shorter phrases. This hypothesis may be tested with a χ^2 (chi square) test.

```
table( age.cat, phraselength.cat ) # show 2x2 table
```

```
##          phraselength.cat
## age.cat   (4.44,9] (9,13.6]
##   (21,40]       28       12
##   (40,59]       32        8
```

```
chisq.test( age.cat, phraselength.cat )
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  age.cat and phraselength.cat
## X-squared = 0.6, df = 1, p-value = 0.4386
```

The data in the table show that, as hypothesized, the odds of older talkers producing short phrases (32/8 or 4.0 : 1) are indeed higher than the odds of younger talkers producing short phrases (28/12 or 2.3 : 1). The effect is far from significant, however, and H_0 is not rejected.

5.4 aov

This function performs a between-subjects analysis of variance, with only fixed factors (Johnson, 2008) (For more complex analyses of variance having repeated measures, see Johnson 2008; for mixed effects models, see Chapter 7 and references cited there.) In the example below we create a response variable `aa` which is not normally distributed (check with `hist`, `qqnorm`, etc).

```
a1<-rpois(20,lambda=2); a2<-rpois(20,lambda=4); a3<-rpois(20,lambda=6)
aa <- c(a1,a2,a3)
x1 <- as.factor(rep(1:3,each=20))
# x1 corresponds with the three different poisson distributions within aa
x2 <- as.factor(rep( rep(1:2,each=10), 3)) # no effect expected
summary( model1.aov <- aov(aa~x1*x2) )
```

```
##             Df Sum Sq Mean Sq F value     Pr(>F)
## x1           2 109.73   54.87  15.512 0.00000475 ***
## x2           1   0.27    0.27   0.075      0.785
## x1:x2        2   6.93    3.47   0.980      0.382
## Residuals   54 191.00    3.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 6

Regression

6.1 `lm`

This function is used for regression according to a linear model, i.e. linear regression. It returns a model-class object. There are specialized functions for such models, e.g. to extract residuals (`resid`), to extract regression coefficients (`coef`), to modify (`update`) the model, etc.

In the following example, we construct two regression models. The first step in your regression analysis should always be a thorough visual and numerical inspection, including histograms and scatterplots of the variables under study.

The first model is

$\texttt{nsyl} = b_0$

having only a constant intercept. The second model includes the speakers' age as a predictor, i.e.

$\texttt{nsyl} = b_0 + b_1 \texttt{age}$

(The intercept is included in this model too, by default, unless suppressed explicitly with `-1` or `~0` in the regression formula). The key question here is whether inclusion of a predictor yields a better model, with significantly smaller residuals and significantly higher R^2. The intercept-only model and the linear-regression model are compared with the `anova` function.

```r
require(hqmisc)
data(talkers)
model1.lm <- lm( nsyl~1, data=talkers ) # only intercept
model2.lm <- lm( nsyl~age, data=talkers ) # with intercept
anova( model1.lm, model2.lm ) # compare two models
```

```
## Analysis of Variance Table
##
## Model 1: nsyl ~ 1
## Model 2: nsyl ~ age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     79 381.85
## 2     78 375.06  1    6.7823 1.4105 0.2386
```

Including the `age` predictor does improve the model a little bit, as indicated by the somewhat smaller residual sums-of-squares (`RSS`). The improvement, however, is too small to be of significance. The linear effect of a speaker's age on his or her average phrase length (in syllables) is not significant.

6.2 `glm`

For logistic regression we use function `glm(family=binomial)`, again with a regression formula as an obligatory argument. Logistic regression can be imagined as computing the logit of the hit-rate for each cell, and then regressing these logits on the predictor(s). Here is an annotated example (Moore and McCabe, 2003) (Exercise 15.25).

The response variable `outcome` indicates the death (`0`) or survival (`1`) of 2900 patients in two hospitals.

```
ips1525 <- read.table(
  file=url("http://www.hugoquene.nl/emlar/ipsex1525.txt"),
  header=T, sep=",")
with( ips1525, table(outcome) )
```

```
## outcome
##    0    1
##   79 2821
```

```
log(2821/79) # log of odds of survival is 3.575
```

```
## [1] 3.575399
```

```
# intercept-only logistic-regression model
model1.glm <- glm( outcome~1, data=ips1525, family=binomial )
summary(model1.glm)
```

```
##
```

```
## Call:
## glm(formula = outcome ~ 1, family = binomial, data = ips1525)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.684   0.235   0.235   0.235   0.235
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.5754     0.1141   31.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 725.1  on 2899  degrees of freedom
## Residual deviance: 725.1  on 2899  degrees of freedom
## AIC: 727.1
##
## Number of Fisher Scoring iterations: 6
```

The intercept coefficient of the intercept-only logistic regression model equals the overall log odds of survival; this intercept is significantly different from zero [1]. Thus the intercept-only logistic regression model captures the overall log odds of survival. Next, let's try to improve this model, by including two predictors: first, the `hospital` where the patient was treated, and second, the patient's `condition` at intake, classified as bad (`0`) or good (`1`).

```
model2.glm <- glm( outcome~hospital, data=ips1525, family=binomial )
model3.glm <- glm( outcome~hospital*condition,
                   data=ips1525, family=binomial )
```

The deviance among logistic-regression models follows a χ^2 distribution. Hence we can compare models by computing the χ^2 probability of their deviances. Both model 2 and model 3 are compared here against model 1.

```
anova( model1.glm, model2.glm, model3.glm, test="Chisq" )
```

```
## Analysis of Deviance Table
##
## Model 1: outcome ~ 1
## Model 2: outcome ~ hospital
## Model 3: outcome ~ hospital * condition
```

[1]If the odds are $1:1$, then the log of the odds is $\log(1) = 0$.

```
##   Resid. Df Resid. Dev Df Deviance   Pr(>Chi)
## 1      2899     725.10
## 2      2898     722.78  1   2.3253     0.1273
## 3      2896     703.96  2  18.8222 0.00008181 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results indicate that there is no significant difference among hospitals in their survival rates (model 2, $p > .10$), but there is a significant effect of intake condition on the outcome (model 3, $p < .001$). Of course, you should also inspect the models themselves before drawing conclusions.

Chapter 7

Mixed-effects modeling

Many language (acquisition) studies are based on samples of two random factors: a sample of participants (subjects) and a sample of language items (words, sentences, texts). The two random factors are *crossed*, i.e., each item is presented to each participant — often only once, so that a subject does not respond to the same item repeatedly in multiple conditions. The analysis methods shown above (`aov`, `lm`, `glm`) all fail to acknowledge this particular structure in the random part of the design. They include a single random factor (named `Residual`) that aggregates all random effects.

A better method is to use *mixed-effects modeling*, which may be done in R by using the `lmer` command. Key advantages of this method[^15] are (a) it allows multiple random factors, crossed and/or nested, (b) it does not require homogeneity of variance, (c) it is robust against missing data. Hence mixed-effects modeling is quickly gaining in popularity (Quené and van den Bergh, 2004; Quené and Van den Bergh, 2008; Baayen, 2008; Hox et al., 2018).

For mixed-effects modeling, you need to install two add-on packages to R , named `lme4` and `languageR` (Baayen et al., 2008). For more information on packages, see Chapter 8 below. After activation of these packages, we can simply perform a mixed-effects analysis. First, we read in an example dataset (Quené and Van den Bergh, 2008) in long data layout:

```
x24 <- read.table( file=url("http://www.hugoquene.nl/emlar/x24r2.txt"), header=T )
```

These fictitious responses were provided by 24 subjects, for 36 items, in 3 conditions, with rotation of items over conditions. This rotation may be inspected for a small subset of the data frame:

```
with( subset(x24, subj<=3&item<=6), table(subj,item,cond) )
```

```
## , , cond = 1
##
##     item
## subj 1 2 3 4 5 6
##    1 1 1 1 1 1 1
##    2 0 0 0 0 0 0
##    3 0 0 0 0 0 0
##
## , , cond = 2
##
##     item
## subj 1 2 3 4 5 6
##    1 0 0 0 0 0 0
##    2 1 1 1 1 1 1
##    3 0 0 0 0 0 0
##
## , , cond = 3
##
##     item
## subj 1 2 3 4 5 6
##    1 0 0 0 0 0 0
##    2 0 0 0 0 0 0
##    3 1 1 1 1 1 1
```

Next, we need to specify that `cond` is a categorical factor, and not a continuous predictor. In addition, we specify the levels of the factor, we specify its contrasts, and indicate that the second level is the baseline or reference level.

```
x24$cond <- as.factor(x24$cond)
contrasts(x24$cond) <- contr.treatment( c(".A",".B",".C"), base=2 )
```

After these preliminary steps we can estimate an appropriate mixed-effects model in a single command. The estimated model is also stored as an object, and a summary is displayed.

```
summary( x24.m1 <- lmer(resp ~ 1+cond+(1|subj)+(1|item),
                        data=x24, REML=FALSE) )
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: resp ~ 1 + cond + (1 | subj) + (1 | item)
##    Data: x24
```

```
##
##      AIC      BIC   logLik deviance df.resid
##   2046.8   2075.4  -1017.4   2034.8      858
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1641 -0.6394 -0.0077  0.6416  2.7157
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  item     (Intercept) 0.2579   0.5078
##  subj     (Intercept) 0.2891   0.5377
##  Residual             0.5103   0.7144
## Number of obs: 864, groups:  item, 36; subj, 24
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.04569    0.14485   0.315
## cond.A       0.17037    0.05953   2.862
## cond.C      -0.23696    0.05953  -3.980
##
## Correlation of Fixed Effects:
##        (Intr) cond.A
## cond.A -0.205
## cond.C -0.205  0.500
```

The output correctly shows that there are two unrelated random effects, plus unexplained residual variance. Each response is now modeled as a unique combination of the intercept (mean of baseline condition B), item effect, subject effect, condition effect, and residual. The average response in the baseline condition B is 0.046 units. Responses in condition A are 0.170 units higher than baseline, and in condition C they are -0.237 units higher than baseline, i.e. 0.237 units lower.

For reasons not discussed here [1], the significance levels of the fixed effects are not reported in the output of `lmer`. There are several solutions to obtain these significance levels.

- The most conservative option (Hox et al., 2018) is to use the critical t value associated with the random effect that has the fewest levels (here `subj`), corrected for the number of fixed coefficients including the intercept (here 3). If a fixed effect is significant by this very conservative criterion, then it will also be significant by any other criterion that is less conservative and more liberal.

[1] See Frequently Asked Questions about R, Question 7.35, at http://cran.r-project.org/doc/FAQ/R-FAQ.html

```
qt( p=1-.05/2, df=24-3 ) # critical value t*, alpha=.05, two-sided
```

```
## [1] 2.079614
```

Comparison of the fixed effects with this critical $t^* = 2.08$ shows that both conditions A and C differ significantly from the baseline condition B.

- A second option is to estimate 95% confidence intervals for all coefficients in the resulting mixed-effects model. This can be time-consuming, as the mixed-effects model will be re-fit many times on slightly varying datasets.

```
# warnings may occur
print( x24.m1.ci <- confint( x24.m1, method="boot", nsim=250 ))
```

```
## Computing bootstrap confidence intervals ...
```

```
##
## 6 warning(s): Model failed to converge with max|grad| = 0.00267345 (tol = 0.002, co
```

```
##                   2.5 %     97.5 %
## .sig01       0.35950709  0.6747204
## .sig02       0.36333157  0.6844652
## .sigma       0.68054619  0.7474401
## (Intercept) -0.23533348  0.3634542
## cond.A       0.05695928  0.3035280
## cond.C      -0.36262503 -0.1261797
```

As the interval for `cond.A` is entirely positive, we may conclude with 95% confidence that condition A yields higher scores than the baseline condition B, and mutatis mutandis that condition C yields lower scores than condition B.

Chapter 8

Packages

Packages are user-provided extensions to the basic R system (comparable to an "add-on" or "extension" for some web browsers). Packages may contain custom datasets, additional functions, re-formulations of existing functions, and more. There are by now thousands of useful packages extending R.

A package named `ABC` can be installed by entering `install.packages("ABC")` (with double quotes) and then `require(ABC)` (without double quotes). If a package in turn requires other packages, these are also installed and loaded.

Some useful packages are the following:

- `datasets`: A wide variety of datasets, for exploration and education. This package is now integrated within R (see Help on `datasets`).
- `foreign`: Functions for reading and writing data stored by statistical packages such as Minitab, S, SAS, SPSS, Stata, Systat, and for reading CSV files (comma separated values) created by Microsoft Excel, and for reading and writing dBase files (R Core Team, 2020).
- `lattice`: Trellis graphics for R. The functions provide a powerful, elegant and flexible high-level data visualization system, using Trellis graphics, with an emphasis on multivariate data (Sarkar, 2018).
- `MASS`: Datasets and functions accompanying (Venables and Ripley, 2002; Ripley, 2019)
- `languageR`: Datasets and functions accompanying (Baayen, 2008; Baayen and Shafaei-Bajestan, 2019).
- `hqmisc`: Some convenience functions and an example dataset, by the present author (Quené, 2014), and used in this booklet.

- `tidyverse`: A meta-package consisting of many other packages supporting your data science, such as `dplyr` for data transformation, `ggplot2` for data visualization, and `rmarkdown` for reporting (also used for this booklet) (Wickham et al., 2019); all component packages can also be used separately.

Packages are stored on a so-called *repository*; the CRAN repository is the most important one (https://cran.r-project.org/). You should use a nearby mirror site of the CRAN repository, by giving the command `chooseCRANmirror()`. Rstudio and R remember your chosen mirror site over multiple sessions.

Finally, to inspect the status of your current session in R, use the command `sessionInfo()`. This will return a listing of technical information, locale settings, all attached packages, and all loaded packages, with version info for each.

Chapter 9

Further reading

A wealth of useful documentation is available through the `Help` tab in the lower right pane of RStudio. Browse in the FAQ files, the help files, the manuals and the vignettes that come with R and Rstudio.

More help can be called from within R by giving the command `help(ABC)` or `?ABC` with a command or operator as argument. (This can be typed in the `Console` tab in the lower left pane of RStudio.)

There is also a lot more help available on the internet, in particular from the R project website https://www.r-project.org and the RStudio website https://www.Rstudio.com.

A few other useful web resources are:

- Quick-R, http://statmethods.net
- http://math.illinoisstate.edu/dhkim/Rstuff/Rtutor.html

The following three books on using R and on its applications in linguistics are highly recommended: (Baayen, 2008; Johnson, 2008; Adler, 2010).

Happy analyses!

Bibliography

Adler, J. (2010). *R in a Nutshell.* O'Reilly.

Baayen, R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge University Press.

Baayen, R., Davidson, D., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390—412.

Baayen, R. H. and Shafaei-Bajestan, E. (2019). *languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics.* R package version 1.5.0.

Hox, J. J., Moerbeek, M., and van de Schoot, R. (2018). *Multilevel Analysis: Techniques and Applications.* Routledge, 3rd edition.

Johnson, K. (2008). *Quantitative Methods in Linguistics.* Blackwell.

Moore, D. S. and McCabe, G. P. (2003). *Introduction to the Practice of Statistics.* Freeman, 4th edition.

Peterson, G. and Barney, H. (1952). Control methods in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184.

Quené, H. (2014). *hqmisc: Miscellaneous convenience functions and dataset.* R package version 0.1-1.

Quené, H. and van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, 43(1–2):103–121.

Quené, H. and Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59(4):413–425.

R Core Team (2020). *foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...* R package version 0.8-75.

Ripley, B. (2019). *MASS: Support Functions and Datasets for Venables and Ripley's MASS.* R package version 7.3-51.5.

Sarkar, D. (2018). *lattice: Trellis Graphics for R.* R package version 0.20-38.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S.* Springer, 4th edition.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wilkinson, G. N. and Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3):392—399.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.18.