Appendix: Introduction to *R*

A

Background

The open-source software *R* was developed as a free implementation of the language *S* which was designed as a language for statistical computation, statistical programming, and graphics. The main intention was to allow users to explore data in an easy and interactive way, supported by meaningful graphical representations. The statistical software *R* was originally created by Ross Ihaka and Robert Gentleman (University of Auckland, New Zealand).

Installation and Basic Functionalities

- The "base" *R* version, i.e. the software with its most relevant commands, can be downloaded from https://www.r-project.org/. After installing *R*, it is recommended to install an editor too. An editor allows the user to conveniently save and display *R*-code, submit this code to the *R* console (i.e. the *R* software itself), and control the settings and the output. A popular choice of editor is *R*Studio (free of charge) which can be downloaded from https://www.rstudio.com/ (see Fig. A.1 for a screenshot of the software). There are alternative good editors, for example "Tinn-R" (http://sourceforge.net/projects/tinn-r/).
- A lot of additional user-written packages are available online and can be installed within the *R* console or using the *R* menu. Within the console, the `install.packages("package to install")` function can be used. Please note that an internet connection is required.

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5

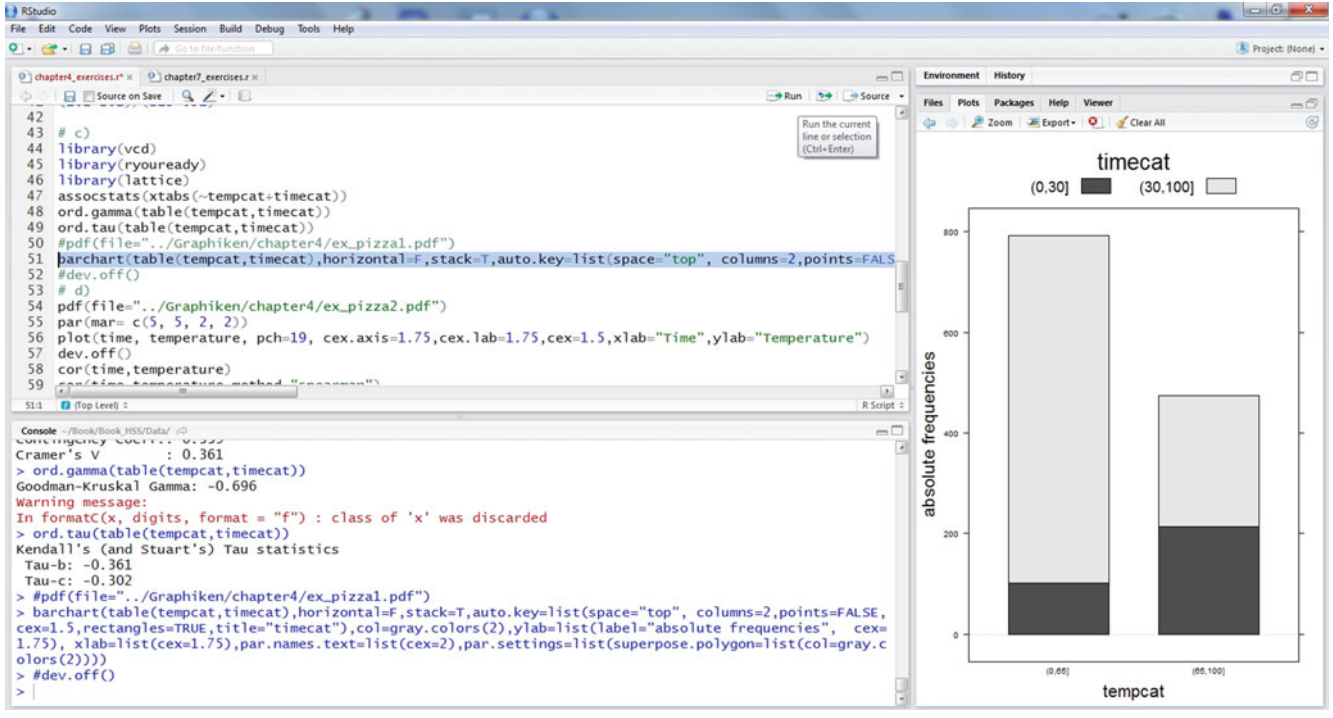


Fig. A.1 Screenshot of *R* Studio with the command window (*top left*), the output window ("console", i.e. *R* itself, *bottom left*), and a plot window (*right*). Other windows (e.g. about the environment or package updates) are closed

Statistics has a close relationship to algebra: data sets can be seen as matrices, and variables as vectors. *R* makes use of these structures and this is why we first introduce data structure functionalities before explaining some of the most relevant basic statistical commands.

R as a Calculator, Basic Data Structures and Arithmetic Operations

- The character # marks the beginning of a comment. All characters until the end of the *line* are ignored by *R*. We use # to comment on our *R*-code.
- If we know the name of a command we would like to use, and we want to learn about its functionality, typing `?command` in the *R* command line prompt displays a help page, e.g.

```
?mean
```

  displays a help page for the arithmetic mean function.
- Using

```
example(mean)
```

  shows application examples of the respective function.
- The command `c(1,2,3,4,5)` combines the numbers 1, 2, 3, 4 and 5 into a vector (e.g. a variable).
- Vectors can be assigned to an "object". For example,

```
X <- c(2,12,22,32)
```

  assigns a numeric vector of length 4 to the object `X`. In general, the arrow sign ($<-$) is a very important concept to store data, summaries, and outputs in objects (i.e. the name in the front of the $<-$ sign). Note that *R* is case sensitive; i.e. `X` and `x` are two different variable names. Instead of "$<-$", one can also use "$=$".
- Sequences of numbers can be created using the `seq` and `rep` commands. For example,

```
seq(1,10)
```

  and

```
rep(1,10)
```

  yield

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

  and

```
 [1] 1 1 1 1 1 1 1 1 1 1
```

  respectively.

- Basic data structures are **vectors, matrices, arrays, lists**, and **data frames**. They can contain numeric values, logical values or even characters (strings). In the latter case, arithmetic operations are not allowed.

  – A numeric vector of length 5 can be constructed by the command

  ```
  x <- vector(mode="numeric", length=5)
  ```

  The elements can be accessed by squared brackets: [ ]. For example,

  ```
  x[3] <- 4
  ```

  assigns the value 4 to the third element of the object `x`. Logical vectors containing the values `TRUE` and `FALSE` are also possible. In arithmetic operations, `TRUE` corresponds to 1 and `FALSE` corresponds to 0 (which is the default). Consider the following example:

  ```
  x.log <- vector(mode="logical", length=4)
  x.log[1] = x.log[3] = x.log[4] = TRUE
  mean(x.log)
  ```

  returns as output `0.75` because the mean of (1, 0, 1, 1) =(TRUE, FALSE, TRUE, TRUE) is 0.75.

  – A matrix can be constructed by the `matrix()` command:

  ```
  x <- matrix(nrow=4, ncol=2, data=1:8, byrow=T)
  ```

  creates a 4×2 matrix, where the data values are the natural numbers $1, 2, 3, \ldots, 8$ which are stored row-wise in the matrix,

  ```
  [,1] [,2]
  [1,]    1    2
  [2,]    3    4
  [3,]    5    6
  [4,]    7    8
  ```

  because of the parameter `byrow=T` (which is equivalent to `byrow=TRUE`). The default is `byrow=F` which would store the data column-wise.

  – Arrays are more general data structures than vectors and matrices in the sense that they can have more than two dimensions. For instance,

  ```
  x <- array(data=1:12, dim=c(3,2,2))
  ```

  creates a three-dimensional array with $3 \cdot 2 \cdot 2 = 12$ elements.

  – A list can contain objects of different types. For example, a list element can be a vector or matrix. Lists can be initialized by the command `list` and can grow dynamically. It is important to understand that list elements should be accessed by the name of the entry via the dollar sign or using double brackets:

```
 x <- list(one=c(1,2,3,4,5),two=c("Hello", "world", "!"))
 x
$one
[1] 1 2 3 4 5

$two
[1] "Hello" "world" "!"

x[[2]]
[1] "Hello" "world" "!"
x$one
    [1] 1 2 3 4 5
```

– A data frame is the standard data structure for storing a data set with rows as observations and columns as variables. Many statistical procedures in *R* (such as the `lm` function for linear models) expect a data frame. A data frame is conceptually not much different from a matrix and can either be initialized by reading a data set from an external file or by binding several column vectors. As an example, we consider three variables (age, favourite hobby, and favourite animal) each with five observations:

```
age    <- c(25,33,30,40,28)
hobby  <- c("Reading","Sports","Games","Reading","Games")
animal <- c("Elephant", "Giraffe", NA, "Monkey", "Cat")
dat    <- data.frame(age,hobby,animal)
names(dat) <- c("Age","Favourite.hobby","Favourite.animal")
dat
```

The resulting output is

```
 > dat
  Age Favourite.hobby Favourite.animal
1  25        Reading          Elephant
2  33         Sports          Giraffe
3  30          Games             <NA>
4  40        Reading           Monkey
5  28          Games              Cat
```

where `NA` denotes a missing value. With `write.table` or a specialized version thereof such as `write.csv` (for writing the data in a file using comma-separated fields), a data frame can be saved in a file. The command sequence

```
write.csv(x=dat,file="toy.csv",row.names=FALSE)
read.dat <- read.csv(file="toy.csv")
read.dat
```

saves the data frame as an external (comma-separated) file and then loads the data again using `read.csv`.

Individual elements of the data frame can be accessed using squared brackets, as for matrices. For example, `dat[1,2]` returns the first observation of the second variable column for the data set `dat`. Individual columns (variables) can also be selected using the $ sign:

```
dat$Age
```

returns the age column:

```
[1] 25 33 30 40 28
```

- The `factor` command is very useful to store nominal variables, and the command `ordered` is ideal for ordinal variables. Both commands are extremely important since factor variables with more than two categories are automatically expanded into several columns of dummy variables if necessary, e.g. if they are included as covariates in a linear model. In the previous paragraph, two factor variables have already been created. This can be confirmed by typing

```
is.factor(dat$Favourite.hobby)
is.factor(dat$Favourite.animal)
```

which return the value `TRUE`. Have a look at the following two factor variables:

```
sex <- factor("female","male","male","female","female")
grade <- ordered(c("low", "medium", "low", "high", "high"),
         levels=c("low", "medium","high"))
```

Please note that by default alphabetical order is used to order the categories (e.g. female is coded as 1 and male as 2). However, the mapping of integers to strings can be controlled by the user as seen for the variable "grade":

```
grade
```

returns

```
[1] low    medium low    high   high
Levels: low < medium < high
```

- Basic arithmetic operations can be applied directly to a numeric vector. Basic operations are addition $+$, subtraction $-$, multiplication $*$ and division $/$, integer division $\%/\%$, modulo operation $\%\%$, and exponentiation with two possible notations: $**$ or ^. Examples are given as:

```
2^3                    # command
[1] 8                  # output
2**3                   # command
[1] 8                  # output
2^0.5                  # command
[1] 1.414214           # output
c(2,3,5,7)^2           # command: application to a vector
[1]  4  9 25 49        # output
c(2,3,5,7)^c(2,3)      # command: !!!  ATTENTION!
[1]   4  27  25 343    # output
c(1,2,3,4,5,6)^c(2,3,4) # command
[1]    1    8   81   16  125 1296 #output
c(2,3,5,7)^c(2,3,4)    # command: !!! WARNING MESSAGE!
[1]   4  27 625  49
Warning message:
longer object length
        is not a multiple of shorter object length
        in: c(2, 3, 5, 7)^c(2, 3, 4)
```

The last four commands show the "recycling property" of *R*. It tries to match the vectors with respect to the length if possible. In fact,

```
c(2,3,5,7)^c(2,3)
```

is expanded to

```
c(2,3,5,7)^c(2,3,2,3)
```

The last example shows that *R* gives a warning if the length of the shorter vector cannot be expanded to the length of the longer vector by a simple multiplication with a natural number (2, 3, 4, ...). Here

```
c(2,3,5,7)^c(2,3,4)
```

is expanded to

```
c(2,3,5,7)^c(2,3,4,2)
```

such that not all elements of

```
c(2,3,4)
```

are "recycled".

More on indexing

The standard ways of accessing/indexing elements in a vector, matrix, list, or data frame have already been introduced above, but *R* allows a lot more flexible accessing of elements.

1. Selecting elements using vectors of positive numbers (`letters` and `LETTERS` show the 26 letters of the alphabet)

```
letters[1:3]
letters[ c(2,4,6) ]
```

```
[1] "a" "b" "c"
[1] "b" "d" "f"
```

2. Selecting elements using logical vectors

```
x <- 1:10          # numbers 1 to 10
x[ (x>5) ]         # selecting any number >5
x[ (x%%2==0) ]     # numbers that are divisible by 2
x[(x%%2==1)]       # numbers that are not divisible by 2
x[5] <- NA         # 5th element of x is defined
                   # to be missing (NA)
x
y <- x[!is.na(x)]  # all x which are not missing
y
```

returns the output

```
[1]  6  7  8  9 10
[1]  2  4  6  8 10
[1] 1 3 5 7 9
[1]  1  2  3  4 NA  6  7  8  9 10
[1]  1  2  3  4  6  7  8  9 10
```

3. Selecting (deleting) elements using negative numbers

```
x <- 1:10
x[-(1:5)]    # x, but delete first five entries of x
```

returns the output

```
[1]  6  7  8  9 10
```

because the first five elements have been removed.

4. Selecting elements using characters

```
x <- c(Water=1, Juice=2, Lemonade=3 )
names(x)
x["Juice"]
```

returns the output

```
[1] "Water"   "Juice"     "Lemonade"
Juice
    2
```

Standard Functions

Some standard functions and their roles in *R* are

`abs()`	Absolute value
`sqrt()`	Square root
`round()`, `floor()`, `ceiling()`	Rounding, up and down
`sum()`, `prod()`	Sum and product
`log()`, `log10()`, `log2()`	Logarithms
`exp()`	Exponential function
`sin()`, `cos()`, `tan()`, `asin()`, `acos()`, `atan()`	Trigonometric functions
`sinh()`, `cosh()`, `tanh()`, `asinh(x)`, `acosh()`, `atanh(x)`	Hyperbolic functions

All functions can again be applied directly to numeric vectors.

Statistical Functions

Some statistical functions and their roles in *R* are

`mean()`, `var()`	Mean and variance
`cov()`, `cor()`	Covariance and correlation
`min()`, `max()`	Minimum and maximum

Note: the arguments of the functions vary depending on the chosen method. For example, the `mean()` function can be applied to general *R* objects where averaging makes sense (numeric or logical vectors, but also, e.g. matrices). The functions `var(), cov(), cor()` expect one or two numeric vectors, matrices, or data frames. Minimum and maximum functions work also with a comma-separated list of values, i.e.

```
min(2, 6.7, 1.2, 8.0)
```

provides the same result (`1.2`) as

```
min(c(2, 6.7, 1.2, 8.0))
```

Examples:

```
mean( c(1,2,5,6) )
[1] 3.5



var( c(1,2,5,6) )
[1] 5.666667
```

Note that `var(), cov()` use the factor $1/(n-1)$ for the unbiased estimate of the variance instead of $1/n$ for the empirical variance, i.e. $1/3$ in the example above. Both functions can also be applied to several vectors or a matrix. Then the covariance matrix (and correlation matrix in case of `cor()`) is computed. For example, consider two variables

```
age.v <- c(25,30,35,40)
income.v <- c(2000, 2500, 2800, 3200)
```

Then both commands return the symmetric covariance matrix (with variances as the diagonal entries and covariances as the non-diagonal entries).

```
var(cbind(age.v, income.v))
             age.v income.v
age.v      41.66667   3250.0
income.v 3250.00000 255833.3
```

```
cov(cbind(age.v, income.v))
             age.v income.v
age.v      41.66667   3250.0
income.v 3250.00000 255833.3
```

The (Pearson) correlation between the two variables is calculated as 0.9954293.

```
cor(cbind(age.v, income.v))
            age.v  income.v
age.v    1.0000000 0.9954293
income.v 0.9954293 1.0000000
```

The Spearman rank correlation is perfectly 1, since both vectors are in increasing order:

```
cor(cbind(age.v, income.v), method="spearman")
         age.v income.v
age.v        1        1
income.v     1        1
```

More Useful Functions

Some more commonly used standard functions and their roles in *R* are as follows:

- Cumulative sum and product:

```
x <- c( 1,3, 2, 5)
cumsum(x)          # 1, 1+3, 1+3+2, 1+3+2+5
cumprod(x)         # 1, 1*3, 1*3*2, 1*3*2*5
```

give the output

```
[1]  1  4  6 11
[1]  1  3  6 30
```

- Factorial:

```
factorial(5)
```

returns 5! as

```
[1] 120
```

- Binomial coefficient $\binom{n}{k}$:

```
choose(4,2)
```

returns $\binom{4}{2}$ as

```
[1] 6
```

Mathematical Constants

The number π is a mathematical constant, the ratio of a circle's circumference to its diameter, and is commonly approximated as 3.14159. It is directly available in *R* as `pi`.

```
 pi
[1] 3.141593
```

Other "constants" are

`Inf, -Inf`	$\infty, -\infty$
`NaN`	Not a Number: e.g. 0/0[1] NaN
`NA`	Not Available: missing values
`NULL`	empty set

Assignment Operator for Creating Functions

The assignment operator $<-$ ("less than" sign followed by hyphen) has already been introduced above in the context of variables. Alternatively, $=$ (equality sign) can be used. One can create one's own functions: the function is an object with a name which takes values specified in the round brackets and returns what has been specified in the curly braces. For example, the following function `myfunction` returns a polynomial of degree 3 for a given argument x. Note that by default all four coefficients are equal to 1.

```
my.function <- function(x,a=1,b=1,c=1,d=1){
  h <- a+b*x+c*x^2+d*x^3
  return(h)
}

my.function(2)
[1] 15

my.function(x=2, a=4, b=3)
[1] 22
```

Loops and Conditions

The concept of loops is convenient when some operation has to be repeated. Loops can be utilized in various ways, for example, via `for` or `while`. Conditions are specified with the `if` statement. For example,

```
x <- 1:10
for(i in 1:10){ if(x[i]>5){x[i] <- x[i]+i}
}
x
```

returns

```
 [1]  1  2  3  4  5 12 14 16 18 20
```

In this example, x is a vector with 10 elements: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. For each element `x[i]`, we replace `x[i]` with `x[i]+i` if the condition `x[i]>5` is true; otherwise we do not.

Statistical Functions

Now we consider some basic statistical functions in *R*. For illustration, we use the `painters` data in the following example. This data is available after loading the library `MASS` (only a subset is shown below). The data lists the subjective assessment, on a 0 to 20 integer scale, of 54 classical painters. The painters were assessed on four characteristics: composition, drawing, colour, and expression. The data is due to the eighteenth-century art critic, de Piles. Use `?painters` for more information on the data set.

```
library(MASS)
painters
```

shows

```
                Composition Drawing Colour Expression School
Da Udine                 10       8     16          3      A
Da Vinci                 15      16      4         14      A
Del Piombo                8      13     16          7      A
Del Sarto                12      16      9          8      A
Fr. Penni                 0      15      8          0      A
```

The `Summary` Function

The summary function allows a quick overview of a data frame. For numeric variables, the five-point summary (which is also used in a simple box plot, see Sect. 3.3) is calculated together with the arithmetic mean. For factor variables, the absolute frequencies of the most frequent categories are printed. If the factor has more than six categories, the other categories are summarized in a separate category—`Other`.

```
summary(painters)
```

yields

```
  Composition       Drawing         ...        School
 Min.   : 0.00   Min.   : 6.00     ...   A      :10
 1st Qu.: 8.25   1st Qu.:10.00     ...   D      :10
 Median :12.50   Median :13.50     ...   E      : 7
 Mean   :11.56   Mean   :12.46     ...   G      : 7
 3rd Qu.:15.00   3rd Qu.:15.00     ...   B      : 6
 Max.   :18.00   Max.   :18.00     ...   C      : 6
                                   ...  (Other) : 8
```

The summary function can also be applied to a single variable:

```
summary(painters$School)
```

returns

```
 A  B  C  D  E  F  G  H
10  6  6 10  7  4  7  4
```

Accessing Subgroups in Data Frames

Subgroups, i.e. groups of observations which share the same feature(s) of one or several variables, can be accessed using the `subset` command.

```
subset(painters, School=="F")
```

accesses all painters for which `School==''F''` holds.

```
           Composition Drawing Colour Expression School
Durer                8      10     10          8      F
Holbein              9      10     16         13      F
Pourbus              4      15      6          6      F
Van Leyden           8       6      6          4      F
```

This is a more elegant method than selecting these observations by specifying a condition in squared brackets via the `[rows,columns]` argument.

```
painters[ painters[["School"]] == "F", ]
```

Note that the explained structure is an important one: we access the rows and columns of a matrix or data set by using the `[rows,columns]` argument. Here we access all rows for which the variable "school" is "F". If, in addition, we also want to restrict the data set to the first two variables, we can write:

```
painters[ painters[["School"]] == "F", c(1,2)]
```

Similarly,

```
subset(painters, Composition <= 6)
```

gives the output

```
              Composition Drawing Colour Expression School
Fr. Penni               0      15      8          0      A
Perugino                4      12     10          4      A
Bassano                 6       8     17          0      D
Bellini                 4       6     14          0      D
Murillo                 6       8     15          4      D
Palma Vecchio           5       6     16          0      D
Caravaggio              6       6     16          0      E
Pourbus                 4      15      6          6      F
```

Uninteresting columns can be eliminated using negative indexing. For instance, in the following example,

```
subset(painters, School=="F", select=c(-3,-5) )
```

```
           Composition Drawing Expression
Durer                8      10          8
Holbein              9      10         13
Pourbus              4      15          6
Van Leyden           8       6          4
```

the third and the fifth columns (Colour and School) are not shown.

The operator `%in%` allows for more complex searches. For instance,

```
subset(painters, Drawing %in% c(6,7,8,9) & Composition==10)
```

returns the following output:

```
             Composition Drawing Colour Expression School
Da Udine              10       8     16          3      A
J. Jordaens           10       8     16          6      G
Bourdon               10       8      8          4      H
```

i.e. those painters with a drawing score between 6 and 9 (= any number which matches 6, or 7, or 8, or 9).

Stratifying a Data Frame and Applying Commands to a List

Sometimes it is of interest to apply statistical commands (such as `summary`) to several subgroups. If this is the case, the data is partitioned into different groups using `split` and then `lapply` applies a function to each of these groups. The command `split` partitions the data set by values of a specific variable. For example, we first stratify the `painters` data with respect to the painter's school:

```
splitted <- split(painters, painters$School)
splitted
```

```
$A
                Composition Drawing Colour Expression School
Da Udine                 10       8     16          3      A
Da Vinci                 15      16      4         14      A
Del Piombo                8      13     16          7      A
Del Sarto                12      16      9          8      A
Fr. Penni                 0      15      8          0      A
Guilio Romano            15      16      4         14      A
```

```
Michelangelo              8      17      4          8      A
Perino del Vaga          15      16      7          6      A
Perugino                  4      12     10          4      A
Raphael                  17      18     12         18      A


$B
             Composition Drawing Colour Expression School
F. Zucarro            10      13      8          8      B
Fr. Salviata          13      15      8          8      B
Parmigiano            10      15      6          6      B
Primaticcio           15      14      7         10      B
T. Zucarro            13      14     10          9      B
Volterra              12      15      5          8      B

$C
...
```

Note, that `splitted` is now a list,

```
is.list(splitted)
```

returns

```
[1] TRUE
```

while the objects `splitted$A` to `splitted$H` are data frames.

```
is.data.frame(splitted$A)
```

returns

```
[1] TRUE
```

Secondly, as indicated above, the command `lapply` allows us to apply a function to a list. For instance,

```
lapply(splitted, summary)
```

applies the summary function to all data frames in the list `splitted` (output not shown). See also `?apply, ?sapply, ?tapply`, and `?mapply` for similar operations.

Sorting, Ranking, Finding Duplicates, and Unique Values

- Sorting a vector:

```
x <- c( 1,3, 2, 5)
sort(x)
sort(x, decreasing=TRUE)
```

  returns the ordered values in decreasing order as

```
[1] 5 3 2 1
```

  See also the command `order` for showing the order of vector elements.

- Calculating ranks;

```
x <- c( 10,30, 20, 50, 20)
rank(x)
```

  returns the following output:

```
[1] 1.0 4.0 2.5 5.0 2.5
```

- Finding duplicate values:

```
x <- c( 1,3, 2, 5, 2)
duplicated(x)
```

  indicates which values occur more than once:

```
[1] FALSE FALSE FALSE FALSE  TRUE
```

- Removing duplicates:

```
x <- c( 1,3, 2, 5, 2)
unique(x)
```

  shows the output as

```
[1] 1 3 2 5
```

  This means `unique` finds out how many *different* values a vector has.

Categorizing Numeric Variables

Continuous variables (vectors) can be categorized using the `cut` command.

```
x <- c(1.3, 1.5, 2.5, 3.8, 4.1, 5.9, 7.1, 8.4, 9.0)
xdiscrete <- cut(x, breaks=c(-Inf, 2, 5, 8, Inf) )
is.factor(xdiscrete)
xdiscrete
table(xdiscrete)
```

returns

```
[1] TRUE
[1] (-Inf,2] (-Inf,2] (2,5] (2,5] (2,5] (5,8] (5,8] (8,Inf]
[9] (8,Inf]
Levels: (-Inf,2] (2,5] (5,8] (8,Inf]
(-Inf,2]    (2,5]    (5,8]  (8,Inf]
 2        3        2        2
```

Random Variables

- *R* has built-in functions for several probability density/mass functions (PMF/PDF), (probability) distribution function (i.e. the CDF), quantile functions and for generating random numbers.
- The function names use the following scheme:

First letter	Function	Further letters
`d`	density	`distribution name`
`p`	probability	`distribution name`
`q`	quantiles	`distribution name`
`r`	random number	`distribution name`

- Examples:

```
dnorm(x=0)
[1] 0.3989423
```

returns the value of the density function (i.e. $P(X = x)$) of a $N(0, 1)$-distribution at $x = 0$, which is $1/\sqrt{2\pi}$.

```
pnorm(q=0)
pnorm(q=1.96)
[1] 0.5
[1] 0.9750021
```

returns the value of the CDF of a $N(0, 1)$-distribution at q, i.e. $\Phi(q) = P(X \leq q)$.

```
qnorm(p=0.95)
```

returns the value

```
[1] 1.644854
```

which is the 95 % quantile of a $N(0, 1)$-distribution.

```
X <- rnorm(n=4)
X
```

returns a vector of four normal random numbers of a $N(0, 1)$-distribution:

```
[1] -0.90826678 -0.09089654 -0.47679821  1.22137230
```

Note that a repeated application of this function leads to different random numbers. To get a reproducible sequence of random numbers, a `seed` value needs to be set:

```
set.seed(89234512)
X <- rnorm(n=4)
X
```

If all three commands are executed, then the sequence is (using the standard random generator)

```
[1] -1.07628865  0.37797715  0.04925738 -0.22137107
```

- The following functions for distributions can be used:

Model distributions

Function	Distribution
`beta`	Beta
`binom`	Binomial
`cauchy`	Cauchy
`exp`	Exponential
`gamma`	Gamma
`geom`	Geometric
`hyper`	Hypergeometric
`lnorm`	Log–normal
`norm`	Normal
`pois`	Poisson
`unif`	Uniform
`mvnorm`	Multivariate normal (in package `mvtnorm`)

Test distributions

Function	Distribution
`chisq`	χ^2
`f`	F
`signrank`	Wilcoxon signed rank (1 sample)
`t`	t
`wilcox`	Wilcoxon rank sum (2 samples)

- For convenience, we list a few important PDF and CDF values in Sect. C.

Key Points and Further Issues

Note:

- ✓ *R* uses the following data structures: vectors, matrices, arrays, lists, and data frames.
- ✓ Entries of matrices and data frames can be accessed using squared brackets. For example, `data[1:5,2]` refers to the first five observations of the second column (variable) of the data. Variables of data frames can also be accessed via the $ sign, e.g. via `data$variable`.
- ✓ If operations such as statistical analyses have to be repeated on several subgroups, using `split` together with `lapply` is a viable option. Alternatively, loops (e.g. `for`) together with conditions (such as `if`) can be used.
- ✓ *R* contains the most relevant statistical functions needed for descriptive and inductive analyses (as shown throughout the book). User-written packages can be installed using `install.packages(''package_name'')`.
- ✓ Readers who are interested in learning more about *R* are referred to Albert and Rizzo (2012), Crawley (2013), Dalgaard (2008), Ligges (2008), and Everitt and Hothorn (2011).

Data Sets

From the data used in this book, we publish some relevant data sets, along with solutions of the *R*-exercises, on https://chris.userweb.mwn.de/book/. The important data sets are explained in the following paragraphs.

Pizza Delivery Data

The pizza delivery data (`pizza_delivery.csv`, see also Table A.1) is a simulated data set. The data refers to an Italian restaurant which offers home delivery of pizza. It contains the orders received during a period of one month: May 2014. There are three branches of the restaurant. The pizza delivery is centrally managed: an operator receives a phone call and forwards the order to the branch which is nearest to the customer's address. One of the five drivers (two of whom only work part time at the weekend) delivers the order. The data set captures the number of pizzas ordered as well as the final bill (in €) which may also include drinks, salads, and pasta dishes. The owner of the business observed an increased number of complaints, mostly because pizzas arrive too late and too cold. To improve the service quality of his business, the owner wants to measure (i) the time from call to delivery and (ii) the pizza temperature at arrival (which can be done with a special device). Ideally, a pizza arrives within 30 min of the call; if it takes longer than 40 min, then the customers are promised a free bottle of wine (which is not always handed out though). The temperature of the pizza should be above 65 °C at the time of delivery. The analysis of the data aims to determine the factors which influence delivery time and temperature of the pizzas.

Table A.1 First few rows of the pizza delivery data

```
       day       date  time operator branch    driver temperature
1 Thursday 1 May 2014  35.1    Laura   East     Bruno        68.3
2 Thursday 1 May 2014  25.2  Melissa   East Salvatore        71.0
3 Thursday 1 May 2014  45.6  Melissa   West Salvatore        53.4
4 Thursday 1 May 2014  29.4  Melissa   East Salvatore        70.3
5 Thursday 1 May 2014  30.0  Melissa   West Salvatore        71.5
6 Thursday 1 May 2014  40.3  Melissa Centre     Bruno        60.8
...


   bill  pizzas free_wine got_wine discount_customer
1  58.4       4         0        0                 1
2  26.4       2         0        0                 0
3  58.1       3         1        0                 0
4  35.2       3         0        0                 0
5  38.4       2         0        0                 0
6  61.8       4         1        1                 0
....
```

Table A.2 First few rows of the decathlon data from the 2004 Olympic Games in Athens data

	100m	Long.jump	Shot.put	High.jump	400m
Roman Sebrle	10.85	7.84	16.36	2.12	48.36
Bryan Clay	10.44	7.96	15.23	2.06	49.19
Dmitriy Karpov	10.50	7.81	15.93	2.09	46.81
Dean Macey	10.89	7.47	15.73	2.15	48.97
Chiel Warners	10.62	7.74	14.48	1.97	47.97
Attila Zsivoczky	10.91	7.14	15.31	2.12	49.40
...					

	110m.hurdle	Discus	Pole.vault	Javelin	1500m
Roman Sebrle	14.05	48.72	5.0	70.52	280.01
Bryan Clay	14.13	50.11	4.9	69.71	282.00
Dmitriy Karpov	13.97	51.65	4.6	55.54	278.11
Dean Macey	14.56	48.34	4.4	58.46	265.42
Chiel Warners	14.01	43.73	4.9	55.39	278.05
Attila Zsivoczky	14.95	45.62	4.7	63.45	269.54
...					

Decathlon Data

This data (`decathlon.csv`, see also Table A.2) describes the results of the decathlon competition during the 2004 Olympic Games in Athens. The performance of all 30 athletes in the 100 m race (in seconds), long jump (in metres), shot-put (in metres), high jump (in metres), 400 m race (in seconds), 110 m hurdles race (in seconds), discus competition (in metres), pole vault (in metres), javelin competition (in metres), and 1500 m race (in seconds) are recorded in the data set.

Theatre Data

This data (`theatre.csv`, see also Table A.3) summarizes a survey conducted on 699 participants in a big Swiss city. The survey participants are all frequent visitors to a local theatre and were asked about their age, sex (gender, female = 1), annual income (in 1000 SFR), general expenditure on cultural activities ("Culture", in SFR per month), expenditure on theatre visits (in SFR per month), and their estimated expenditure on theatre visits in the year before the survey was done (in SFR per month).

Table A.3 First few rows of the theatre data

	Age	Sex	Income	Culture	Theatre	Theatre_ly
1	31	1	90.5	181	104	150
2	54	0	73.0	234	116	140
3	56	1	74.3	289	276	125
4	36	1	73.6	185	75	130
5	24	1	109.0	191	172	140
6	25	0	93.1	273	168	130
...						

Appendix: Solutions to Exercises

B

Solutions to Chapter 1

Solution to Exercise 1.1

(a) The population consists of all employees of the airline. This may include administration staff, pilots, stewards, cleaning personnel, and others. Each single employee relates to an observation in the survey.
(b) The population comprises all students who take part in the examination. Each student represents an observation.
(c) All people suffering high blood pressure in the study area (city, province, country, ...), are the population of interest. Each of these persons is an observation.

Solution to Exercise 1.2 The *population* in this study refers to all leopards in the park. Only a few of the leopards are equipped with the GPS devices. This is the *sample* on which the study is conducted in. Each leopard refers to an *observation*. The measurements are taken for each leopard in the sample. The GPS coordinates allow to determine the position during the entire day. Important *variables* to capture would therefore be $X_1 =$ "latitude", $X_2 =$ "longitude", and $X_3 =$ "time". Each variable would take on certain *values* for each observation; for example, the first leopard may have been observed at latitude 32° at a certain time point, and thus $x_{11} = 32°$.

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5

Solution to Exercise 1.3

Qualitative:	Preferred political party, eye colour, gender, blood type, subject line of an email.
Quantitative and discrete:	Shoe size, customer satisfaction on a scale from 1 to 10, number of goals in a hockey match.
Quantitative and continuous:	Time to travel to work, price of a canteen meal, wavelength of light, delivery time of a parcel, height of a child.

Solution to Exercise 1.4

(a) The choice of a political party is measured on a nominal scale. The names of the parties do not have a natural order.
(b) Typically the level of a computer game is measured on an ordinal scale: for example, level 10 may be more difficult than level 5, but this does not imply that level 10 is twice as difficult as level 5, or that the difference in difficulty between levels 2 and 3 is the same as the difference between levels 10 and 11.
(c) The production time of a car is measured on a continuous scale (ratio scale). In practice, it may be measured in days from the start of the production.
(d) This variable is measured on a continuous scale (ratio scale). Typically, the age is captured in years starting from the day of birth.
(e) Calender year is a continuous variable which is measured on an interval scale. Note that the year which we define as "zero" is arbitrary, and it varies from culture to culture. Because the year zero is arbitrary, and we also have dates before this year, the calender year is measured on an interval scale.
(f) The scale is continuous (ratio scale).
(g) The scale of ID numbers is nominal. The ID number may indeed consist of numbers; however, "112233" does not refer to something half as much/good as "224466". The number is descriptive.
(h) The final rank is measured on an ordinal scale. The ranks can be clearly ordered, and the participants can be ranked by using their final results. However the first winner may not have "double" the beauty of the second winner, it is merely a ranking.
(i) The intelligence quotient is a variable on a continuous scale. It is constructed in such a way that differences are interpretative—i.e. being 10 points above or 10 points below the average score of 100 points means the same deviation from the average. However, ratios cannot be interpreted, so the intelligence quotient is measured on an interval scale.

Solution to Exercise 1.5

(a) The data is provided in *.csv* format. We thus read it in with the `read.csv()` command (after we have set a working directory with `setwd()`):

```
setwd('C:/directory')
pizza <- read.csv('pizza_delivery.csv')
```

(b) The data can be viewed by means of the `fix()` or `View()` command or simply being printed:

```
fix(pizza)
pizza
```

(c) We can access the data, as for any matrix, by using squared brackets [,], see also Appendix A.1. The first entry in the brackets refers to the row and the second entry to the columns. Each entry either is empty (referring to every row/column) or consists of a vector or sequence describing the columns/rows we want to select. This means that we obtain the first 5 rows and variables via `pizza[1:5,1:5]`. If we give the new data the name "pizza2" we simply need to type:

```
pizza2 <- pizza[1:5,1:5]
pizza2
```

We can save this new data either as a *.dat* file (with `write.table()`), or as a *.csv* file (with `write.csv()`), or directly as an *R* data file (with `save()`) which gives us access to our entire *R* session.

```
write.csv(pizza2,file='pizza2.csv')
write.table(pizza2,file='pizza2.dat')
save(pizza2,file='pizza2.Rdata')
```

(d) We can access any variable by means of the $ sign. If we type pizza$new we create a new variable in the pizza data set called "new". Therefore, a simple way to add a variable to the data set is as follows:

```
pizza$NewTemperature <- 32+1.8*pizza$temperature
```

(e)

```
attach(pizza)
NewTemperature
```

```
> str(pizza)
'data.frame':   1266 obs. of  13 variables:
 $ day              : Factor w/ 7 levels "Friday","Monday",..: 5 5 5 5 
 $ date             : Factor w/ 31 levels "01-May-14","02-May-14",..: 1
 $ time             : num  35.1 25.2 45.6 29.4 30 ...
 $ operator         : Factor w/ 2 levels "Laura","Melissa": 1 2 2 2 2 2
 $ branch           : Factor w/ 3 levels "Centre","East",..: 2 2 3 2 3 
 $ driver           : Factor w/ 5 levels "Bruno","Domenico",..: 1 5 5 5
 $ temperature      : num  68.3 71 53.4 70.3 71.5 ...
 $ bill             : num  58.4 26.4 58.1 35.2 38.4 61.8 57.9 35.8 36.6
 $ pizzas           : int  4 2 3 3 2 4 3 2 2 5 ...
 $ free_wine        : int  0 0 1 0 0 1 1 0 0 0 ...
 $ got_wine         : int  0 0 0 0 0 1 1 0 0 0 ...
 $ discount_customer: int  1 0 0 0 0 0 0 0 0 0 ...
 $ NewTemperature   : num  155 160 128 159 161 ...
```

Fig. B.1 Applying `str()` to the pizza data

(f) We can apply all these commands onto the object "pizza". The command `str(pizza)` gives us an overview of the data set, see also Fig. B.1. The output shows that the data set contains 1266 observations (deliveries) and 13 variables. We can also see which of these variables are factors (categorical with defined categories) and which are numerical. We also see the first actual numbers for the each variable and the coding scheme used for the categories of the factor variables. The command `dim` summarizes the dimension of the data, i.e. the number of rows and columns of the data matrix. `Colnames` gives us the names of the variables from the data set, and so does `names`. The commands `nrow` and `ncol` give us the number of rows and columns, respectively. Applying `head` and `tail` to the data prints the first and last rows of the data, respectively.

Solution to Exercise 1.6

(a) The appropriate study design is a survey. The information would be obtained via a questionnaire given to a sample of parents. It is not a controlled experiment because we do not manipulate one particular variable, while controlling others; we rather collect data on all variables of interest.

(b) There are different options to ask for parents' attitudes: of course one could simply ask "what do you think of immunization?"; however, capturing long answers in a variable "attitude" may make it difficult to summarize and distil the information obtained. A common way to deal with such variables is to translate a concept into a score: for example, one could ask 5 "yes/no"-type questions (instead of one general question) which relate to attitudes towards immunization, such as "do you think immunization may be harmful for your child?" or "do you agree that it is a priority to immunize infants in their first year of life?" The number of answers that show a positive attitude towards immunization can be summed up. If there are 5 questions, there are up to 5 points "to earn". Thus, each parent may be asked 5 questions and his/her attitude can be summarized on a scale ranging from 0 to 5, depending on the answers given.

(c) The following variables are needed:

- Attitude: possibly several variables are needed to capture parents' information in a score, see (b) for details. The scale is ordinal because a higher score relates to a more positive attitude towards immunization, but the differences between different score levels cannot be interpreted meaningfully.
- Immunized: a binary ("yes–no" type) variable capturing whether the parent agrees to immunization against chickenpox for their youngest child or not. This is a nominal variable.
- Gender: to compare "Immunized" for male and female parents. This is a nominal variable.
- Age: to compare the age distribution in the group of parents who would immunize their child with the age distribution in the group who would not immunize their child. Age is measured on a continuous scale.

(d) A data set might look as follows:

$$\begin{pmatrix} \text{Parent} & \text{A}_1 & \dots & \text{A}_5 & \text{Attitude} & \text{Immunization} & \text{Gender} & \text{Age} \\ 1 & \text{yes} & \dots & \text{yes} & 3 & \text{yes} & \text{male} & 35 \\ 2 & \text{no} & \dots & \text{yes} & 2 & \text{no} & \text{female} & 26 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

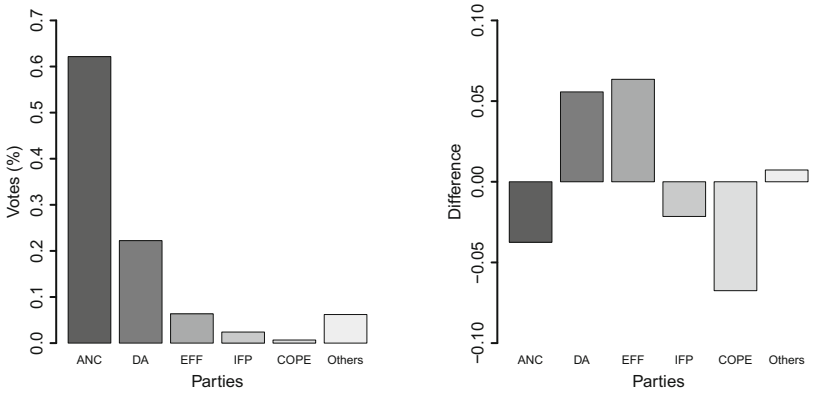
where $A_1, \dots, A_5$ refer to variables capturing attitudes towards immunization and "Attitude" is the score variable summarizing these questions. The questions may be written down as follows:

- What is the average attitude score towards immunization among parents and how much does it vary?
- What percentage of parents answer "yes" when asked whether they would immunize their youngest child against chickenpox?
- What is the difference in the proportion calculated in (b) when stratified by gender?
- What is the average age of those parents who would immunize their child compared with the average age of those parents who would not immunize their child?

Chapter 2

Solution to Exercise 2.1

(a) The table shows the relative frequencies of each party and not the absolute frequencies. We can thus draw a bar chart where the relative frequencies of votes are plotted on the y-axis and different parties are plotted on the x-axis. In *R*, we can first type in the data by defining two vectors and then use the "barplot" command to draw the bar chart (Fig. B.2a). Typing "?barplot" and "?par" shows the graphical options to improve the presentation and look of the graph:



(a) Results of the 2014 elections

(b) Difference in results between the 2014 and 2009 elections

Fig. B.2 Bar charts for national elections in South Africa

```
results2014 <- c(0.6215,0.2223,0.0635,0.0240,0.0067,0.0620)
barplot(results2014)
barplot(results2014,names.arg=c('ANC','DA','EFF','IFP','COPE',
'Others'), col=gray.colors(6),ylim=c(0,0.7),xlab='Parties',ylab =
'Votes(%)')
```

(b) There are several options to compare the results. Of course, one can simply plot the two bar charts with each bar chart representing one election. It would be important for this solution to ensure that the y-axes are identical in both the plots. However, if we want to compare the election results in one graph then we can plot the difference between the two elections, i.e. the win/loss per party. The bar chart for the new variable "difference in proportion of votes between the two elections" is shown in Fig. B.2 and is obtained as follows:

```
results2009 <- c(0.6590,0.1666,0,0.0455,0.0742,0.0547)
difference <- results2014-results2009
barplot(difference)
```

Remark Another solution would be to create subcategories on the x-axis: for example, there would be categories such as "ANC 2009 results" and "ANC 2014 results", followed by "DA 2009 results" and "DA 2014 results", and so on.

Solution to Exercise 2.2

(a) The scale of X is continuous. However, please note that the number of values X can practically take is limited (90 min plus extra time, recorded in 1 min intervals).

Table B.1 Frequency table and other information for the variable "time until first goal"

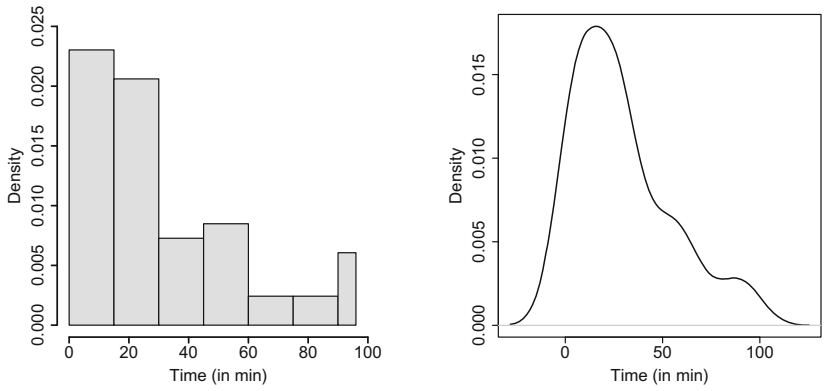
j	$[e_{j-1}, e_j)$	n_j	f_j	d_j	h_j	$F(x)$
1	[0, 15)	19	$\frac{19}{55}$	15	$\frac{19}{825}$	$\frac{19}{55}$
2	[15, 30)	17	$\frac{17}{55}$	15	$\frac{17}{825}$	$\frac{36}{55}$
3	[30, 45)	6	$\frac{6}{55}$	15	$\frac{6}{825}$	$\frac{42}{55}$
4	[45, 60)	5	$\frac{5}{55}$	15	$\frac{5}{825}$	$\frac{47}{55}$
5	[60, 75)	4	$\frac{4}{55}$	15	$\frac{4}{825}$	$\frac{51}{55}$
6	[75, 90)	2	$\frac{2}{55}$	15	$\frac{2}{825}$	$\frac{53}{55}$
7	[90, 96)	2	$\frac{2}{55}$	6	$\frac{2}{825}$	1
Total		56	1			

(b) It is straightforward to obtain the frequency table, as well as the other information needed to obtain the histogram and the ECDF, see Table B.1.

(c) We need to obtain the heights for each category to obtain the histogram using $h_j = f_j/d_j$, see Table B.1.

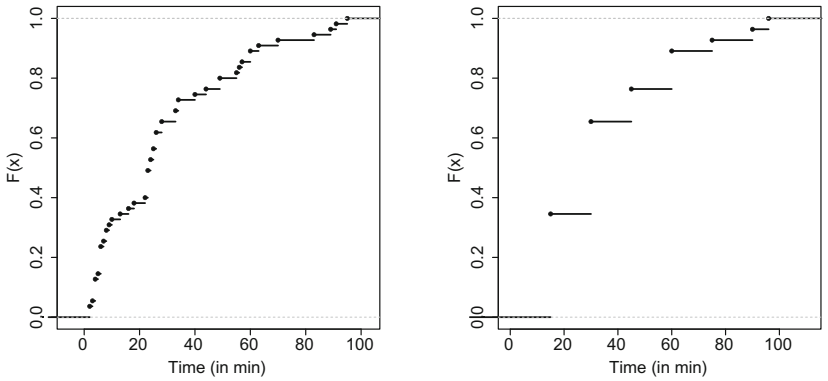
(d) We obtain the histogram and kernel density plot in *R* (Fig. B.3a) using the commands

```
goals <- c(6,24,91,...,7)
hist(goals, breaks=c(0,15,30,45,60,75,90,96))
plot(density(goals, adjust=1,kernel='gaussian'))
```



(a) Histogram (b) Kernel density plot

Fig. B.3 Distribution of time to goal



(a) Original data (b) Grouped data

Fig. B.4 Empirical cumulative distribution function for the variable "time until first goal"

(e) The ECDF values for $F(x)$ are calculated using the relative frequencies $f(x)$, see Table B.1.

(f) (i) We can easily plot the ECDF (Fig. B.4a) for the original data using the *R* command

```
plot.ecdf(goals)
```

(ii) Generating the ECDF for the grouped data requires more effort and is not necessarily intuitive: first we categorize the continuous variable using the function `cut`. We use the label option to indicate that the name of each category corresponds to the upper limit of the respective interval. This new variable is a "factor" variable and the `plot.ecdf` function does not work with this type of variable. We need to first change the "factor" variable into a "character" variable with strings corresponding to the labels and coerce this into numeric values. Then we use `plot.ecdf`, see also Fig. B.4b. Alternatively, we can directly replace the raw values with numbers corresponding to the upper interval limits.

```
goals_cat <- cut(goals, breaks=c(0,15,30,45,60,75,90,96),
labels=c(15,30,45,60,75,90,96))
plot.ecdf(as.numeric(as.character(goals_cat))
```

(g) To solve the exercises, we simply use Formula (2.11) and Rules (2.3ff.)

(i) $H(X \leq 45) = F(45) = \frac{42}{55} \approx 0.76$.

(ii) $H(X > 80) = 1 - F(80) = 1 - \left(\frac{51}{55} + \frac{2/55}{15}(80 - 75)\right) \approx 0.085$.

(iii) $H(20 \leq X \leq 65) = F(65) - F(20) = \frac{47}{55} + \frac{4/55}{15} \cdot (65 - 60) - \left(\frac{19}{55} + \frac{17/55}{15} \cdot (20 - 15)\right) \approx 0.43.$

(h) We know from (2.11) that

$$F(x_p) = \; p = \; F(e_{j-1}) + h_j(x_p - e_{j-1})$$

with $h_j = f_j/d_j$ which relates to

$$x_p = e_{j-1} + \frac{p \; - \; F(e_{j-1})}{h_j}.$$

We are interested in $x_{0.8}$ because 80 % of the matches have seen a goal at this time point:

$$x_{0.8} = \; 45 + \frac{0.8 \; - \; \frac{43}{56}}{\frac{1}{168}} = 50.4.$$

We conclude that 80 % of the "first goals" happened up to 50.4 min.

Solution to Exercise 2.3

(a) We obtain the relative frequencies for the first and fourth intervals as 0.125 ($h_j \cdot d_j = 0.125 \cdot 1$). Accordingly, for the other two intervals, we obtain frequencies of $f_j = 3 \cdot 0.125 = 0.375$.

(b) We obtain the absolute frequencies for the first and fourth intervals as 250 ($2000 \cdot 0.125$). For the other intervals, we obtain 750 ($2000 \cdot 0.375$).

Solution to Exercise 2.4

(a) The absolute frequencies n_j are evident from the following table:

j	e_{j-1}	e_j	$F(e_j)$	f_j	$n_j (= f_j n)$	d_j	a_j
1	8	14	0.25	0.25	$0.25 \cdot 500 = 125$	6	11
2	14	22	0.40	0.15	75	8	18
3	22	34	0.75	0.35	175	12	28
4	34	50	0.97	0.22	110	16	42
5	50	82	1.00	0.03	15	32	66

(b) We obtain $F(X > 34) = 1 - F(34) = 1 - 0.75 = 0.25$.

Table B.2 Information needed to calculate the ECDF

Score	1	2	3	4	5	6	7	8	9	10
Results	1	3	8	8	27	30	11	6	4	2
f_j	$\frac{1}{100}$	$\frac{3}{100}$	$\frac{8}{100}$	$\frac{8}{100}$	$\frac{27}{100}$	$\frac{30}{100}$	$\frac{11}{100}$	$\frac{6}{100}$	$\frac{4}{100}$	$\frac{2}{100}$
F_j	$\frac{1}{100}$	$\frac{4}{100}$	$\frac{12}{100}$	$\frac{20}{100}$	$\frac{47}{100}$	$\frac{77}{100}$	$\frac{88}{100}$	$\frac{94}{100}$	$\frac{98}{100}$	1

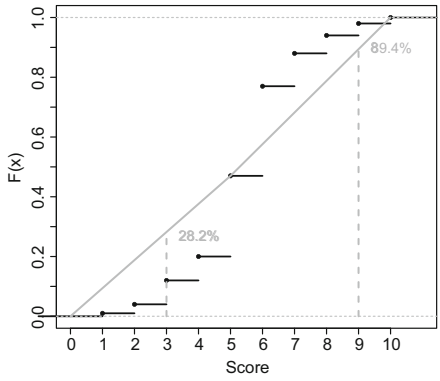


Fig. B.5 Empirical cumulative distribution function for the variable "score"

Solution to Exercise 2.5

(a) The data needed to calculate and draw the ECDF is summarized in Table B.2; the ECDF is plotted in Fig. B.5.

(b) It follows from Table B.2 that $F(3) = 12\,\%$ and $F(9) = 98\,\%$.

(c) The grey solid line in Fig. B.5 summarizes the ECDF for the grouped data. It goes from $(0, 0)$ to $(1, 1)$ with a breakpoint at $(5, 0.47)$ since $F(5) = 0.47$ summarizes the information for the group "disagree". Using (2.11) we calculate:

$$\begin{aligned} F(3) &= F(e_{j-1}) + \frac{f_j}{d_j}(x - e_{j-1}) \\ &= F(0) + \frac{0.47}{5} \cdot (3 - 0) = 28.2\,\% \\ F(9) &= F(5) + \frac{0.53}{5} \cdot (9 - 5) = 89.4\,\%. \end{aligned}$$

(d) The results of (b) and (c) are very different. The formula applied in (c) assumes that the values in each category are uniformly distributed, i.e. that within each category, each value occurs as often as each other value. However, we know from (a) that this is not true: there are more values towards the central score numbers. The assumption used in (c) is therefore inappropriate as also highlighted in Fig. B.5.

Solution to Exercise 2.6 We read in and attach the data as follows:

```
setwd('C:/directory')
pizza <- read.csv('pizza_delivery.csv')
attach(pizza)
```

(a) We need the options `ylim`, `xlim`, `ylab`, `xlab`, `col` to adjust the histogram produced with `hist()`. We then add a dashed (`lty=2`) line (`type='l'`), which is thick (`lwd=3`), from (65, 0) to (65, 400) using the `lines()` command. See also Fig. B.6a.

```
hist(temperature,xlab='Temperature',xlim=c(40,90),
ylim=c(0,400),col='lightgrey',ylab='Number of deliveries')
lines(c(65,65),c(0,400),type='l',lty=2,lwd=3)
```

(b) We can create the histogram as follows, see also Fig. B.6b:

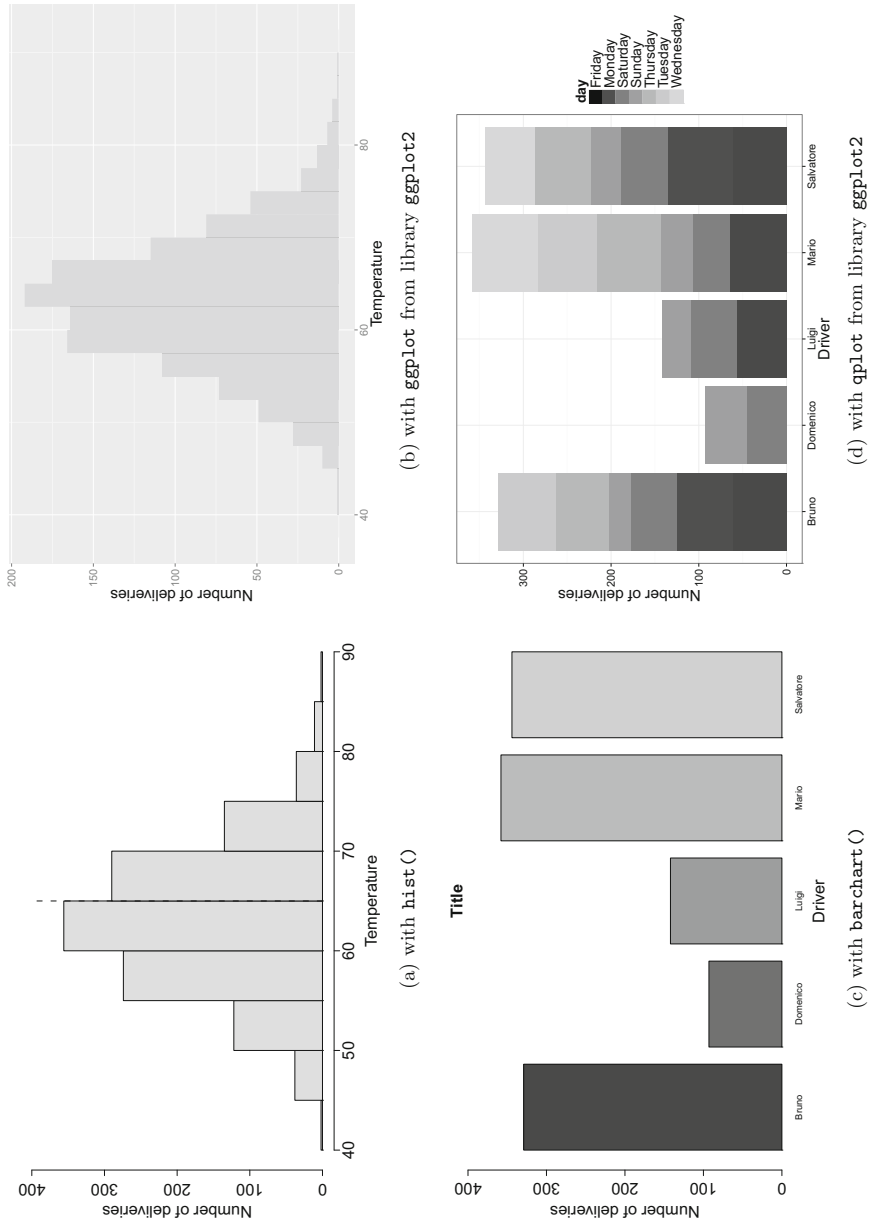
```
library(ggplot2)
p1 <- ggplot(data=pizza,aes(x=temperature))
p2 <- p1 + geom_histogram(fill='darkgrey',alpha=0.5,binwidth=2.5)
+ scale_y_continuous('Number of deliveries')
plot(p2)
```

(c) A possible solution is as follows, see also Fig. B.6c:

```
barplot(table(driver),ylim=c(0,200),col=gray.colors(7),
ylab='Number of deliveries', xlab='Driver',main='Title'))
```

(d) We can produce the graph (Fig. B.6d) as follows:

```
p3 <- qplot(driver,data=pizza,aes=('bar'),fill=day)
p4 <- p3 + scale_fill_grey() +theme_bw()
plot(p4)
```



(a) with `hist()` (b) with `ggplot` from library `ggplot2`

(c) with `barchart()` (d) with `qplot` from library `ggplot2`

Fig. B.6 Creating figures with *R*

Chapter 3

Solution to Exercise 3.1

(a) The arithmetic means can be calculated as follows:

$$\bar{x}_D = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}(12.5 + \cdots + 17.5) = 17.32,$$

$$\bar{x}_A = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}(342 + \cdots + 466) = 612.4.$$

The ordered values of the two variables are:

```
i: 1    2    3    4    5    6    7    8    9    10
-------------------------------------------------------
D: 7.6  12.1 12.5 14.8 16.2 16.5 17.5 18.5 27.4 29.9
A: 238  342  398  466  502  555  670  796  912  1245
```

$$\tilde{x}_{0.5,D} = \frac{1}{2}(\tilde{x}_{(5)} + \tilde{x}_{(6)}) = \frac{1}{2}(16.2 + 16.5) = 16.35,$$

$$\tilde{x}_{0.5,A} = \frac{1}{2}(\tilde{x}_{(5)} + \tilde{x}_{(6)}) = \frac{1}{2}(502 + 555) = 528.5.$$

(b) We have $n\alpha = 10 \cdot 0.25 = 2.5$ which is not an integer. We can therefore calculate the 25 % quantiles, i.e. the first quartiles, as

$$\tilde{x}_{0.25,D} = \tilde{x}_{(3)} = 12.5; \quad \tilde{x}_{0.25,A} = \tilde{x}_{(3)} = 398.$$

Similarly, $n\alpha = 10 \cdot 0.75 = 7.5$ is not an integer and thus

$$\tilde{x}_{0.75,D} = \tilde{x}_{(8)} = 18.5; \quad \tilde{x}_{0.75,A} = \tilde{x}_{(8)} = 796.$$

One can see that the distributions for both variables are not symmetric. For example, when looking at the distance hiked, the difference between the median and the first quartile ($16.35 - 12.5$) is much larger than the difference between the median and the third quartile ($18.5 - 16.35$); this indicates a distribution that is skewed towards the left.

(c) We can calculate the interquartile ranges as follows:

$$d_{Q,A} = 796 - 398 = 398; \quad d_{Q,D} = 18.5 - 12.5 = 6.$$

The mean absolute deviations are:

$$D_D(\tilde{x}_{0.5}) = = \frac{1}{10}(|7.6 - 16.35| + \cdots + |29.9 - 16.35|) = 4.68,$$

$$D_A(\tilde{x}_{0.5}) = = \frac{1}{10}(|238 - 528.5| + \cdots + |1245 - 528.5|) = 223.2.$$

The variances of both the variables are

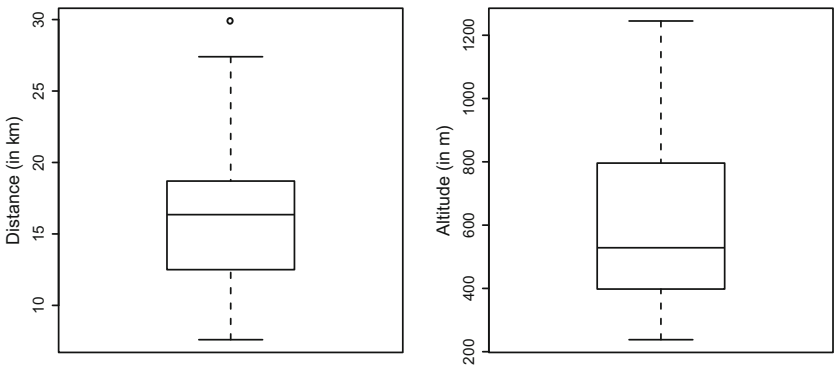
$$\tilde{s}_D^2 = \frac{1}{10}([7.6 - 16.35]^2 + \cdots + [29.9 - 16.35]^2) \approx 41.5,$$
$$\tilde{s}_A^2 = \frac{1}{10}([238 - 528.5]^2 + \cdots + [1245 - 528.5]^2) \approx 82,314.$$

The standard deviations are therefore $\tilde{s}_D = \sqrt{41.5}$ and $\tilde{s}_A = \sqrt{82,314}$.

(d) To answer this question, we can use the rules of linear transformations.

$$\bar{y} \overset{(3.4)}{=} 0 + 3.28\bar{x} = 3.28 \cdot 528.5 = 1722.48,$$
$$\tilde{s}_y^2 \overset{(3.29)}{=} b^2\tilde{s}_x^2 = 3.28^2 \cdot 272.2 \approx 2928.$$

(e) To draw the box plots, we can use the results from (a), (b), and (c) in this solution. The median, first quartile, third quartile, and the interquartile range have already been calculated. It is also easy to determine the minimum and maximum values from the table of ordered values in (a). What is missing is the knowledge of whether to treat some of the values as extreme values or not. For the distance hiked, an extreme value would be any value $>18.5 + 1.5 \times 6 = 27.5$ or $<12.5 - 1.5 \times 6 = 3.5$. It follows that there is only one extreme value: 29.9 km. For the maximum altitude, there is no extreme value because there is no value $>796 + 1.5 \times 398 = 1292$ or $<398 - 1.5 \times 398 = -199$. The box plots are shown in Fig. B.7a, b.



(a) Box plot for distance hiked (b) Box plot for maximum altitude

Fig. B.7 Box plots for Exercise 3.1

(f) The data can be summarized as follows:

Class intervals	(5; 15]	(15; 20]	(20; 30]
n_j	4	4	2
f_j	4/10	4/10	2/10
$\sum f_j$	4/10	8/10	1

We can calculate the weighted arithmetic mean by using the relative frequencies f_j and the middle of the intervals m_j:

$$\bar{x} = \sum_j f_j m_j = \frac{4}{10} \cdot 10 + \frac{4}{10} \cdot 17.5 + \frac{2}{10} \cdot 25 = 16.$$

To estimate the weighted median, we need to determine the class for which

$$\sum_{j=1}^{m-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j \geq 0.5$$

holds. This is clearly the case for the second class $K_2 = (15; 20]$. Thus

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m = 15 + \frac{0.5 - 0.4}{0.4} \cdot 5 = 16.25.$$

(g) If the raw data is known, then the variance for the grouped data will be identical to the variance calculated in (c). For educational purposes, we show the identity here. The variance for grouped data can be calculated as:

$$\tilde{s}^2 = \underbrace{\frac{1}{n} \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2}_{\text{between}} + \underbrace{\frac{1}{n} \sum_{j=1}^{k} n_j \tilde{s}_j^2}_{\text{within}}$$

Using the arithmetic mean $\bar{x} = 17.32$ as well as the means in each class, $\bar{x}_1 = 11.75$, $\bar{x}_2 = 17.225$, and $\bar{x}_3 = 28.65$, we can calculate the variance between the classes:

$$\begin{aligned} \tilde{s}_b^2 = \frac{1}{10}([4 \cdot (11.75 - 17.32)^2] + [4 \cdot (17.225 - 17.32)^2] \\ + [2 \cdot (28.65 - 17.32)^2]) = 38.08735. \end{aligned}$$

The variances within each class are:

$$\begin{aligned} \tilde{s}_1^2 &= \frac{1}{4}[(7.6 - 11.75)^2 + \cdots + (14.8 - 11.75)^2] = 6.8025, \\ \tilde{s}_2^2 &= \frac{1}{4}[(16.2 - 17.225)^2 + \cdots + (18.5 - 17.225)^2] = 0.956875, \\ \tilde{s}_3^2 &= \frac{1}{2}[(27.4 - 28.65)^2 + (29.9 - 28.65)^2] = 1.5625. \end{aligned}$$

We thus get

$$\tilde{s}_w^2 = \frac{1}{10}(4 \cdot 6.8025 + 4 \cdot 0.956875 + 2 \cdot 1.5625) = 3.41625.$$

The total variance is therefore $\tilde{s}^2 = \tilde{s}_w^2 + \tilde{s}_b^2 = 3.41625 + 38.08735 \approx 41.5$. The results will typically differ if we do not know the raw data: we have to replace the arithmetic means within each class, $\bar{x}_j$, with the middle of each class a_j, i.e. $a_1 = 10, a_2 = 17.5, a_3 = 25$:

$$\begin{aligned}\tilde{s}_b^2 = \frac{1}{10}([4 \cdot (10 - 17.32)^2] + [4 \cdot (17.5 - 17.32)^2] \\ +[2 \cdot (25 - 17.32)^2]) = 33.2424.\end{aligned}$$

We further get

$$\begin{aligned}\tilde{s}_1^2 &= \frac{1}{4}[(7.6 - 10)^2 + \cdots + (14.8 - 10)^2] = 9.865, \\ \tilde{s}_2^2 &= \frac{1}{4}[(16.2 - 17.5)^2 + \cdots + (18.5 - 17.5)^2] = 0.9225, \\ \tilde{s}_3^2 &= \frac{1}{2}[(27.4 - 25)^2 + (29.9 - 25)^2] = 1.5625,\end{aligned}$$

and

$$\tilde{s}_w^2 = \frac{1}{10}(4 \cdot 9.865 + 4 \cdot 0.9225 + 2 \cdot 14.885) = 7.292.$$

The variance is $\tilde{s}^2 = \tilde{s}_w^2 + \tilde{s}_b^2 = 7.292 + 33.2424 \approx 40.5$. The approximation is therefore good. However, please note that the between-class variance was estimated too low, but the within-class variance was estimated too high; only the combination of the two variance components led to reasonable results in this example. It is evident that the approximation in the third class was not ideal. The middle of the interval, 25, was not a good proxy for the true mean in this class, 28.65.

(h) It is easy to calculate the mean and the median:

```
distance <- c(12.5,29.9,...,17.5)
altitude <- c(342,1245,...,466)
mean(distance)
mean(altitude)
median(distance)
median(altitude)
```

We can use the `quantile` function, together with the `probs` option, to get the quantiles:

```
quantile(distance,probs=0.75)
quantile(distance,probs=0.25)
quantile(altitude,probs=0.75)
quantile(altitude,probs=0.25)
```

However, the reader will see that the results differ slightly from our results obtained in (b). As noted in Example 3.1.5, *R* offers nine different ways to obtain quantiles, each of which can be chosen by the `type` argument. The difference between these options cannot be understood easily without a background in probability theory. It may, however, be worth highlighting that we get the same results as in (b) if we choose the `type=2` option in this example. The interquartile ranges can be calculated by means of the difference of the quantiles obtained above. To determine the mean absolute deviation, we have to program our own function:

```
amd <- function(mv){1/length(mv)*sum(abs(mv-median(mv)))}
amd(distance)
amd(altitude)
```

We can calculate the variance using the `var` command. However, as noted in Example 3.2.4, on p. 52, *R* uses $1/(n-1)$ rather than $1/n$ when calculating the variance. This important alternative formula for variance estimation is explained in Chap. 9, Theorem 9.2.1. To obtain the results from (c), we hence need to multiply the output from *R* by $(n-1)/n$:

```
var(altitude)*9/10
var(distance)*9/10
```

The box plots can be drawn by using the `boxplot` command:

```
boxplot(altitude)
boxplot(distance)
```

The weighted mean is obtained as follows:

```
weighted.mean(c(10,17.5,25),c(4/10,4/10,2/10))
```

Solution to Exercise 3.2

(a) We need to solve the equation that defines the arithmetic mean:

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_i x_i \\ -90 &= \frac{1}{10}(200 + 600 - 200 - 200 - 200 - 100 - 100 - 400 + 0 + R) \\ -90 &= \frac{1}{10}(-400 + R) \\ &\Rightarrow R = -500.\end{aligned}$$

(b) The mode is $\bar{x}_M = -200$. Using $n\alpha = 2.5$ and $n\alpha = 7.5$, respectively, we can determine the quartiles as follows:

$$\begin{aligned}\tilde{x}_{0.25} &= x_{(3)} = -200, \\ \tilde{x}_{0.75} &= x_{(8)} = 0.\end{aligned}$$

The interquartile range is $d_Q = 0 - (-200) = 200$.

(c) It is not possible to use the coefficient of variation because some of the values are negative.

Solution to Exercise 3.3 We calculate

$$n_m = n - n_w - n_c = 100 - 45 - 20 = 35.$$

Using the formula for the arithmetic mean for grouped data,

$$\bar{x} = \frac{1}{n}(n_w\bar{x}_w + n_m\bar{x}_m + n_c\bar{x}_c),$$

we further get

$$\begin{aligned}\bar{x}_m &= \frac{1}{n_m}(n\bar{x} - n_w\bar{x}_w - n_c\bar{x}_c) \\ &= \frac{1}{35}(100 \cdot 15 - 45 \cdot 16 - 20 \cdot 7.5) = 18.\end{aligned}$$

Similarly, we can use the Theorem of Variance Decomposition, Eq. (3.27), to calculate the variance for the grouped data:

$$\begin{aligned}s^2 = s_w^2 + s_b^2 &= \frac{1}{n}(n_w s_w^2 + n_m s_m^2 + n_c s_c^2) \\ &+ \frac{1}{n}\left(n_w(\bar{x}_w - \bar{x})^2 + n_m(\bar{x}_m - \bar{x})^2 + n_c(\bar{x}_c - \bar{x})^2\right).\end{aligned}$$

This yields

$$\begin{aligned}s_m^2 &= \frac{1}{n_m}\left[ns^2 - n_w s_w^2 - n_c s_c^2 - n_w(\bar{x}_w - \bar{x})^2 - n_m(\bar{x}_m - \bar{x})^2 - n_c(\bar{x}_c - \bar{x})^2\right] \\ &= \frac{1}{35}(100 \cdot 19.55 - 45 \cdot 6 - 20 \cdot 3 - 45 \cdot 1^2 - 35 \cdot 3^2 - 20 \cdot 7.5^2) = 4.\end{aligned}$$

Solution to Exercise 3.4

(a) We evaluate a period of 6 years which means that $T = 0, 1, 2, 3, 4, 5$. To determine the average growth rate, we need to first calculate the geometric mean. There are two options to facilitate this:

(i) Solution:

$$B_t/B_{t-1} = (-, 1.04, 1.125, 0.925, 1.2, 0.933)$$
$$\bar{x}_G = (1.04 \cdot 1.125 \cdot 0.925 \cdot 1.2 \cdot 0.933)^{1/5} = 1.04.$$

(ii) Easier solution:

$$\bar{x}_G = (28/23)^{1/5} = 1.04.$$

Since $\bar{x}_G = 1.04$, the average growth rate is $r = 1.04 - 1 = 4\ \%$.

(b) To predict the number of members in 2018, we could apply the average growth rate to the number of members in 2016 for two consecutive years:

$$B_{2018} = \bar{x}_G B_{2017}, \quad B_{2017} = \bar{x}_G B_{2016}, \quad \Rightarrow B_{2018} = \bar{x}_G^2 B_{2016}$$
$$B_{2018} = 1.04^2 \cdot 28 = 30.28 \approx 31.$$

(c) We could use the approach outlined in (b) to predict the number of members in 2025. However, this assumes that the average growth rate between 2011 and 2016 remains valid until 2025. This is rather unrealistic. The number of members of the club increases in some years, but decreases in other years. There is no guarantee that the pattern observed until 2016 can be generalized to the future. This highlights that statistical methodology should in general be used with care when making long-term future predictions.

(d) The invested money is $\sum_i x_i =$ € 250 million. We deal with partially grouped data because the club's members invest in groups, but the invested sum is clearly defined and not part of a group or interval. We can thus summarize the data as follows:

(i)	1	2	3	4
Number of members	10	8	8	4
Rel. number of members f_j	10/30	8/30	8/30	4/30
$\tilde{u}_i = \sum_j f_j$	10/30	18/30	26/30	1
Money invested x_i	40	60	70	80
Rel. amount per group	40/250	60/250	70/250	80/250
v_i	40/250	100/250	170/250	1

The Lorenz curve is plotted in Fig. B.8.

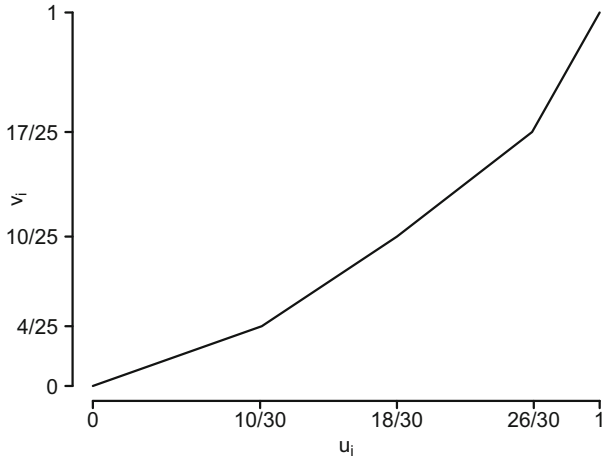


Fig. B.8 Lorenz curve

(e) The Gini coefficient can be calculated using formula (3.37) as follows:

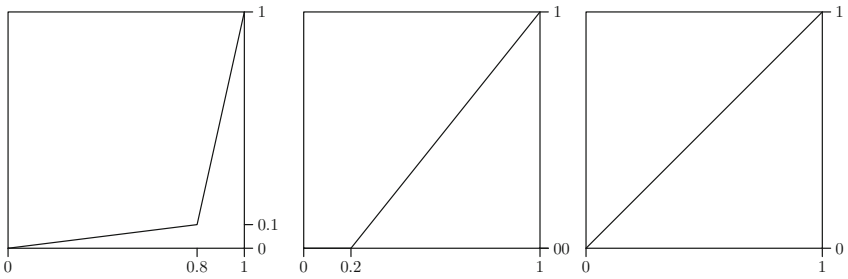
$$G = 1 - \frac{1}{30}(10(0 + 4/25) + 8(4/25 + 2/5) + 8(2/5 + 17/25) + 4(17/25 + 1)) = 1 - 268/375 = 107/375 = 0.2853.$$

$$G^+ \overset{(3.39)}{=} 30/29 \cdot 107/375 = 214/725 = 0.2952.$$

The concentration of investment is thus rather weak.

Solution to Exercise 3.5 Let us look at Fig. 3.8b first. The quantiles coincide approximately with the bisection line. This means that the distribution of "length of service" is similar for men and women. They have worked for about the same amount of time for the company. For example, the median service time should be approximately the same for men and women, the first quartile should be very similar too, the third quartile should be similar too, and so on. However, Fig. 3.8a shows a somewhat different pattern: the quantiles for men are consistently higher than those for women. For example, the median salary will be higher for men. In conclusion, we see that men and women have a similar length of service in the company, but earn less.

Solution to Exercise 3.6 There are many ways in which a "mode" function can be programmed. We present a simple and understandable solution, not the most efficient one. Recall that `table` constructs a frequency table. This table shows immediately which value(s) occur(s) most often in the data. How can we extract it? Applying the



(a) For Exercise 3.7 a) (b) For Exercise 3.7 b) (c) For Exercise 3.7 c)

Fig. B.9 Lorenz curves

`names` function on the table returns the names of the values represented in the table. The only thing we need to do is to choose the name which corresponds to the value which has been counted most. This yields the following function:

```
mymode <- function(vec){
mt <- table(vec)
names(mt)[mt == max(mt)]
}
```

This function will work in general, though it returns a character vector. Using `as.numeric` is one option to make the character string numeric, if necessary.

Solution to Exercise 3.7

(a) In this exercise, we do not have individual data; i.e. we do not know how much each inhabitant earns. The summarized data simply tells us about the wealth of two groups. For simplicity and for illustrative purposes, we assume that the wealth is equally distributed in each group. We determine $(\tilde{u}_i, \tilde{v}_i)$ as $(0.8, 0.1)$ and $(1, 1)$ because 80 % of the population earn 10 % of the wealth and 100 % of the population earn everything. The respective Lorenz curve is illustrated in Fig. B.9a.

(b) The upper class lost its wealth. This means that 20 % of the population do not own anything at all. However, the remaining 80 % owns the rest. This yields $(\tilde{u}_i, \tilde{v}_i)$ of $(0.2, 0)$ and $(1, 1)$, see also Fig. B.9b.

(c) In this scenario, 20 % of the population leave the country. However, the remaining 80 %—which are now 100 % of the population—earn the rest. The money is equally distributed between the population. Figure B.9c shows this situation.

Solution to Exercise 3.8

(a) It is necessary to use the harmonic mean to calculate the average speed. Using $w_1 = n_1/n = 180/418 \approx 0.43$, $w_2 = 117/418 \approx 0.28$, and $w_3 = 121/418 \approx 0.29$ we get

$$\begin{aligned} \bar{x}_H &= \frac{1}{\sum_{i=1}^{k} \frac{w_i}{x_i}} \\ &= \frac{1}{0.43/48 + 0.28/37 + 0.29/52} \approx 45.2\,\text{km/h}. \end{aligned}$$

(b) Travelling at 45.2 km/h means travelling about 361 km in 8 h. The bus will not be in time.

Solution to Exercise 3.9

(a) The sum of investment is $\sum_i x_i = €18,020$. To calculate and draw the Lorenz curve, we need the following table:

(i)	1	2	3	4
investment x_i	800	2220	4700	10300
f_j	1/4	1/4	1/4	1/4
u_i	1/4	2/4	3/4	1
relative investment	0.044	0.123	0.261	0.572
v_i	0.044	0.168	0.428	1

The curve is presented in Fig. B.10.

(b) We can calculate the Gini coefficient as follows:

$$\begin{aligned} G &\overset{(3.37)}{=} 1 - \frac{1}{4}[(0 + 0.044) + (0.044 + 0.168) + (0.168 + 0.428) \\ &\quad + (0.428 + 1)] = 1 - \frac{1}{4} \cdot 2.28 = 0.43. \\ G^+ &\overset{(3.39)}{=} \frac{n}{n-1} G = \frac{4}{3} \cdot 0.43 = 0.57. \end{aligned}$$

(c) The Gini coefficient remains the same as the relative investment stays the same.

(d) Using the library `ineq` we can easily reproduce the results in *R*:

```
library(ineq)
investment <- c(800,10300,4700,2200)
plot(Lc(investment))
ineq(investment)
```

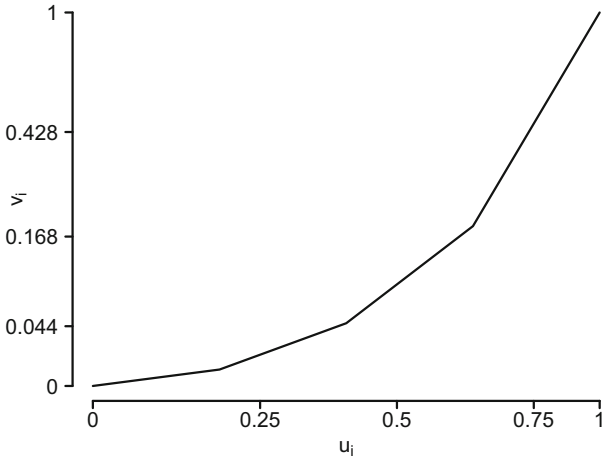


Fig. B.10 Lorenz curve for investment sum

However, please note that the `ineq` command calculates the unstandardized Gini coefficient.

Solution to Exercise 3.10

(a) The easiest way to get all these measures is to use the `summary` function and apply it to the data columns which relate to quantitative variables:

```
setwd('C:/yourpath')
pizza <- read.csv('pizza_delivery.csv')
attach(pizza)
summary(pizza[,c('time','temperature','bill','pizzas')])
```

We then get the following output:

```
      time        temperature         bill           pizzas
 Min.   :12.27   Min.   :41.76   Min.   : 9.10   Min.   : 1.000
 1st Qu.:30.06   1st Qu.:58.24   1st Qu.:35.50   1st Qu.: 2.000
 Median :34.38   Median :62.93   Median :42.90   Median : 3.000
 Mean   :34.23   Mean   :62.86   Mean   :42.76   Mean   : 3.013
 3rd Qu.:38.58   3rd Qu.:67.23   3rd Qu.:50.50   3rd Qu.: 4.000
 Max.   :53.10   Max.   :87.58   Max.   :75.00   Max.   :11.000
```

(b) We can use the quantile function:

```
quantile(time,probs=0.99)
quantile(temperature,probs=0.99)
```

The results are 48.62 min for delivery time and 79.87 °C for temperature. This means 99 % of the delivery times are less than or equal to 48.62 min and 1 % of deliveries are greater than or equal to 48.62 min. Similarly, only 1 % of pizzas were delivered with a temperature greater than 79.87 °C.

(c) The following simple function calculates the absolute mean deviation:

```
amdev <- function(mv){1/length(mv)*sum(abs(mv-mean(mv)))}
amdev(temperature)
```

(d) We can use the `scale`, `mean`, and `var` commands, respectively.

```
sc.time <- scale(time)
mean(sc.time)
var(sc.time)
```

As one would expect, the mean is zero and the variance is 1 for the scaled variable.

(e) The `boxplot` command draws a box plot; the `range` option specifies the range for which extreme values are defined. As specified in the help files, `range=0` produces a box plot with no extreme values.

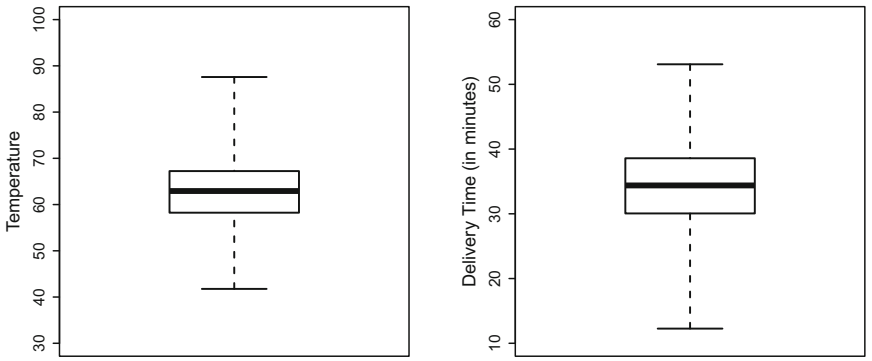
```
boxplot(temperature,range=0)
boxplot(time,range=0)
```

The box plots are displayed in Fig. B.11.

(f) We use the `cut` command to create a variable which has the categories (10, 20], (20, 30], (30, 40], (40, 50], (50, 60], respectively. Using the interval mid-points, as well as the relative frequencies in each class (obtained via the `table` command), we get:

```
tc <- cut(time,breaks=seq(10,60,10))
weighted.mean(c(15,25,35,45,55),table(tc)/sum(table(tc)))
[1] 34.18641
mean(time)
[1] 34.22955
```

The weighted mean is very similar to the mean from the raw data, see output above.



(a) For temperature (b) For delivery time

Fig. B.11 Box plots

(g) The plots can be reproduced by using the `qqplot` command:

```
qqplot(time[driver=='Luigi'],time[driver=='Domenico'])
qqplot(time[driver=='Mario'],time[driver=='Salvatore'])
```

Chapter 4

Solution to Exercise 4.1

(a) We need the following table to calculate the correlation coefficient R:

Café (i)	x_i	$R(x_i)$	y_i	$R(y_i)$	d_i	d_i^2
1	3	1	6	2	−1	1
2	8	4	7	3	1	1
3	7	3	10	5	−2	4
4	9	5	8	4	1	1
5	5	2	4	1	1	1

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6(1+1+4+1+1)}{5(25-1)} = 1 - 0.4 = 0.6.$$

There is a moderate-to-strong positive correlation between the ratings of the two coffee enthusiasts. In general, a high rating from one staff member implies a rather high rating from the other staff member.

(b) Above we have assigned ranks in an increasing order; i.e. the lowest x_i/y_i gets the lowest rank (1) and the highest x_i/y_i gets the highest rank (5). If we use

decreasing order and assign the lowest rank to the highest values, we get the following results:

Café (i)	x_i	$R(x_i)$	y_i	$R(y_i)$	d_i	d_i^2
1	3	5	6	4	1	1
2	8	2	7	3	−1	1
3	7	3	10	1	2	4
4	9	1	8	2	−1	1
5	5	4	4	5	−1	1

As in (a), we have $\sum_i d_i^2 = 8$ and therefore, the results are identical: $R = 0.6$. Depending on whether ranks are assigned in an increasing order or a decreasing order, the sign of d_i differs, but the calculation of R is not affected since the squared values of d_i are used for its calculation and the sign of d_i is thus not important.

(c) We can summarize the ratings in a 2×2 table:

		X	Y
Coffee Quality	Bad	2	1
	Good	3	4

The odds ratio is $OR = (2 \times 4)/(3 \times 1) = 2$. The chance of rating a coffee as good is twice as likely for person X compared to person Y.

Solution to Exercise 4.2

(a) The expected frequencies (under independence) are:

	Satisfied	Unsatisfied
Car	$\frac{74 \cdot 58}{150} = 28.61$	$\frac{76 \cdot 58 \cdot}{150} = 29.39$
Car (diesel engine)	$\frac{74 \cdot 60}{150} = 29.6$	$\frac{76 \cdot 60}{150} = 30.4$
Motorbike	$\frac{74 \cdot 32}{150} = 15.79$	$\frac{76 \cdot 32}{150} = 16.21$

We therefore have

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$$
$$= \frac{(33-28.61)^2}{28.61} + \frac{(25-29.39)^2}{29.39} + \frac{(29-29.6)^2}{29.6}$$
$$+ \frac{(31-30.4)^2}{30.4} + \frac{(12-15.79)^2}{15.79} + \frac{(20-16.21)^2}{16.21}$$
$$= 0.6736 + 0.6557 + 0.0122 + 0.0112 + 0.9097 + 0.8861 = 3.1485.$$

The maximum value χ^2 can take is $150(2-1) = 150$ which indicates that there is almost no association between type of vehicle and satisfaction with the insurance. The other measures can be calculated as follows:

$$V = \sqrt{\frac{\chi^2}{n(\min(k,l)-1)}} = \sqrt{\frac{3.1485}{150(2-1)}} = 0.14.$$

C_{corr}:

$$C_{\text{corr}} = \sqrt{\frac{\min(k,l)}{\min(k,l)-1}}\sqrt{\frac{\chi^2}{\chi^2+n}} =$$
$$= \sqrt{\frac{2}{1}}\sqrt{\frac{3.1485}{3.1485+150}} = \sqrt{2}\sqrt{0.02056} \approx 0.20.$$

The small values of V and C_{corr} confirm that the association is rather weak.

(b) The summarized table looks as follows:

	Satisfied	Unsatisfied
Car	62	56
Motorbike	12	20

Using (4.7), we obtain

$$\chi^2 = \frac{n(ad-bc)^2}{(a+d)(c+d)(a+c)(b+d)}$$
$$= \frac{150(1240-672)^2}{118\cdot 32\cdot 74\cdot 76} = \frac{48,393,600}{21,236,224} \approx 2.2788.$$

The maximum value χ^2 can take is $150(2-1)$. The association is therefore weak. The odds ratio is

$$OR = \frac{ad}{bc} = \frac{62\cdot 20}{12\cdot 56} = \frac{1240}{672} \approx 1.845.$$

The chances of being satisfied with the insurance are 1.845 times higher among those who drive a car.

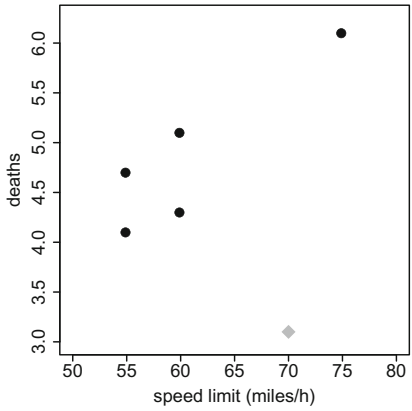


Fig. B.12 Scatter diagram for speed limit and number of deaths

(c) All χ^2-based statistics suggest that there is only a small association between the two variables, for both the 2×3 and the 2×2 tables. However, the odds ratio gives us a more nuanced interpretation, showing that customers driving a car are somewhat more satisfied with their insurance. The question is whether the additional information from the odds ratio is stable and trustworthy. Confidence intervals for the odds ratio can provide guidance under such circumstances, see Sect. 9.4.4 for more details.

Solution to Exercise 4.3

(a) The scatter plot is given in Fig. B.12. The black circles show the five observations. A positive relationship can be discovered: the higher the speed limit, the higher the number of deaths. However, "Italy" (the observation on the top right) is the observation which gives the graph a visible pattern and drives our impression about the potential relationship.

(b) Using $\bar{x} = 61$ and $\bar{y} = 4.86$ we obtain

$$S_{xx} = (55 - 61)^2 + (55 - 61)^2 + \cdots + (75 - 61)^2 = 270$$
$$S_{yy} = (4.1 - 4.86)^2 + (4.7 - 4.86)^2 + \cdots + (6.1 - 4.86)^2 = 2.512$$
$$S_{xy} = (55 - 61)(4.1 - 4.86) + \cdots + (75 - 61)(6.1 - 4.86) = 23.2$$

and therefore

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \frac{23.2}{\sqrt{270 \cdot 2.512}} = 0.891.$$

The correlation coefficient of Spearman is calculated using the following table:

Country (i)	x_i	$R(x_i)$	y_i	$R(y_i)$	d_i	d_i^2
Denmark	55	4.5	4.1	5	−0.5	0.25
Japan	55	4.5	4.7	3	1.5	2.25
Canada	60	2.5	4.3	4	−1.5	2.25
Netherlands	60	2.5	5.1	2	0.5	0.25
Italy	75	1	6.1	1	0	0

This yields

$$R = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 5}{5(25 - 1)} = 0.75.$$

Please note that above we averaged the ranks for ties. The *R* function `cor` uses a more complicated approach; this is why the results differ slightly when using *R*.

(c) The results stay the same. Pearson's correlation coefficient is invariant with respect to linear transformations which means that it does not matter whether we use miles/h or km/h.

(d) (i) The grey square in Fig. B.12 represents the additional observation. The overall pattern of the scatter plot changes with this new observation pair: the positive relationship between speed limit and number of traffic deaths is not so clear anymore. This emphasizes how individual observations may affect our impression of a scatter plot.

(ii) Based on the same approach as in (b), we can calculate $\bar{x} = 62.5$, $\bar{y} = 4.6333$, $S_{xx} = 337.5$, $S_{yy} = 4.0533$, $S_{xy} = 13$, and $r = 0.3515$. A single observation changes our conclusions from a strong positive relationship to a moderate-to-weak relationship. It is evident that Pearson's correlation coefficient is volatile and may be affected heavily by outliers or extreme observations.

Solution to Exercise 4.4

(a) The contingency table for the absolute frequencies is as follows:

	1. Class	2. Class	3. Class	Staff	Total
Rescued	202	125	180	211	718
Not rescued	135	160	541	674	1510
Total	337	285	721	885	2228

(b) To obtain the conditional relative frequency distributions, we calculate the proportion of passengers rescued (X) for each travel class (Y). In the notation of Definition 4.1.1, we determine $f_{i|j}^{X|Y} = f_{ij}/f_{+j} = n_{ij}/n_{+j}$ for all i and j. For example, $f_{\text{rescued}|\text{1. class}} = 202/337 = 0.5994$. This yields

	1. Class	2. Class	3. Class	Staff
Rescued	0.5994	0.4386	0.2497	0.2384
Not rescued	0.4006	0.5614	0.7503	0.7616

It is evident that the proportion of passengers being rescued differs by travel class. It can be speculated that there is an association between the two variables pointing towards better chances of being rescued among passengers from higher travel classes.

(c) Using (4.3), we get

	1. Class	2. Class	3. Class	Staff	Total
Rescued	108.6	91.8	232.4	285.2	718
Not rescued	228.4	193.2	488.6	599.8	1510
Total	337	285	721	885	2228

which can be used to calculate χ^2 and V as follows:

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} = \frac{(202-108.6)^2}{108.6} + \frac{(125-91.8)^2}{91.8} \\
&\quad + \frac{(180-232.4)^2}{232.4} + \frac{(211-285.2)^2}{285.2} + \frac{(135-228.4)^2}{228.4} \\
&\quad + \frac{(160-193.2)^2}{193.2} + \frac{(541-488.6)^2}{488.6} + \frac{(674-599.8)^2}{599.8} \\
&= 80.33 + 12.01 + 11.82 + 19.30 + 38.19 + 5.71 + 5.62 + 9.18 \\
&= 182.16. \\
V &= \sqrt{\frac{\chi^2}{n(\min(k,l)-1)}} = \sqrt{\frac{182.16}{2228(2-1)}} = 0.286.
\end{aligned}$$

The value of V indicates a moderate or weak relationship between the two variables. This is in contradiction to the hypothesis derived from the conditional distributions in (b).

(d) The table is as follows:

	1. Class/2. Class	3. Class/Staff	Total
Rescued	327	391	718
Not rescued	295	1215	1510
Total	622	1606	2228

Using (4.7) we get

$$\chi^2 = \frac{2228(327 \cdot 1215 - 295 \cdot 391)^2}{718 \cdot 1510 \cdot 622 \cdot 1606} = 163.55.$$

and therefore $V = \sqrt{\frac{163.55}{2228}} = 0.271$.
There are several relative risks that can be calculated, for example:

$$\frac{f_{1|1}}{f_{1|2}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{327/622}{391/1606} \approx 2.16,$$
$$\frac{f_{2|1}}{f_{2|2}} = \frac{n_{21}/n_{+1}}{n_{22}/n_{+2}} = \frac{295/622}{1215/1606} \approx 0.63.$$

The proportion of passengers who were rescued was 2.16 times higher in the 1./2. class compared to the 3. class and staff. Similarly, the proportion of passengers who were not rescued was 0.62 times lower in the 1./2. class compared to the 3. class and staff. The odds ratio is $OR = \frac{a \cdot d}{b \cdot c} = \frac{397,305}{115,345} = 3.444$. This is nothing but the ratio of the relative risks, i.e. 2.16/0.63. The chance of being rescued (i.e. the ratio rescued/not rescued) was almost 3.5 times higher for the 1./2. class compared to the 3. class and staff.

(e) While Cramer's V led to rather conservative conclusions regarding a possible relationship of travel class and rescue status, the conditional relative frequency distribution, the relative risks, and the odds ratio support the hypothesis that, at least to some degree, the chances of being rescued were higher in better travel classes. This makes sense because better travel classes were located towards the top of the ship with best access to the lifeboats while both third-class passengers and large numbers of the staff were located and working in the bottom of the ship, where the water came in first.

Solution to Exercise 4.5

(a) Using (4.17) and (4.18), we get

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum_{i=1}^{36} x_i y_i - 36\bar{x}\bar{y}}{\sqrt{n\tilde{s}_x^2 n\tilde{s}_y^2}} = \frac{22776 - 36 \cdot 12.22 \cdot 51.28}{n\sqrt{\tilde{s}_x^2 \tilde{s}_y^2}}$$
$$= \frac{216.9}{36\sqrt{76.95 \cdot 706.98}} \approx 0.026.$$

This indicates that there is no linear relationship between temperature and hotel occupancy.

(b) The scatter plot shows no clear pattern. This explains why the correlation coefficient is close to 0. However, if we look only at the points for each city separately, we see different structures for different cities: a possible negative relationship for Davos (D), a rather positive relationship for Polenca (P) and no visible relationship for Basel (B). This makes sense because for winter holiday destinations hotel occupancy should be higher when the temperature is low and for summer holiday destinations occupancy should be high in the summer months.

(c) We type in the data by specifying three variables: temperature (X), occupancy (Y) and city (Z). We then simply use the `cor` command to calculate the correlation—and condition on the values of Z which we are interested in:

```
X <- c(-6,-5,2,...,9,4)
Y <- c(91,89,76,...,9,12)
Z <- c(rep('Davos',12),rep('Polenca',12),rep('Basel',12))
cor(X[Z=='Davos'],Y[Z=='Davos'])
cor(X[Z=='Basel'],Y[Z=='Basel'])
cor(X[Z=='Polenca'],Y[Z=='Polenca'])
```

This yields correlation coefficients of -0.87 for Davos, 0.42 for Basel and 0.82 for Polenca. It is obvious that looking at X and Y only indicates no correlation, but the information from Z shows strong linear relationships in subgroups. This example shows the limitations of using correlation coefficients and the necessity to think in a multivariate way. Regression models offer solutions. We refer the reader to Chap. 11, in particular Sect. 11.7.3 for more details.

Solution to Exercise 4.6

(a) We use the visual rule of thumb and work from the top left to the bottom right for the concordant pairs and from the top right to the bottom left for the discordant pairs:

$$\begin{aligned} K &= 82 \cdot 43 + 82 \cdot 9 + 82 \cdot 2 + 82 \cdot 10 + 8 \cdot 2 + 8 \cdot 10 + 4 \cdot 10 + 4 \cdot 9 \\ &\quad +43 \cdot 10 = 5850 \\ D &= 4 \cdot 8 + 9 \cdot 2 = 50 \\ \gamma &= \frac{K - D}{K + D} = \frac{5800}{5900} = 0.98. \end{aligned}$$

(b)

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} = \frac{(82 - \frac{86 \cdot 90}{158})^2}{\frac{86 \cdot 90}{158}} + \frac{(4 - \frac{49 \cdot 86}{158})^2}{\frac{49 \cdot 86}{158}}$$

$$+\frac{(0-\frac{19\cdot 86}{158})^2}{\frac{19\cdot 86}{158}}+\frac{(8-\frac{90\cdot 60}{158})^2}{\frac{90\cdot 60}{158}}+\frac{(43-\frac{49\cdot 60}{158})^2}{\frac{49\cdot 60}{158}}+\frac{(9-\frac{60\cdot 19}{158})^2}{\frac{60\cdot 19}{158}}$$
$$+\frac{(0-\frac{12\cdot 90}{158})^2}{\frac{12\cdot 90}{158}}+\frac{(2-\frac{12\cdot 49}{158})^2}{\frac{12\cdot 49}{158}}+\frac{(10-\frac{12\cdot 19}{158})^2}{\frac{12\cdot 19}{158}}$$
$$= 22.25.19.27 + 10.34 + 20.05 + 31.98 + 0.47 + 6.84 + 0.80 + 50.74$$
$$= 162.74.$$
$$V = \sqrt{\frac{\chi^2}{n(\min(k,l)-1)}} = \sqrt{\frac{162.74}{158\cdot 2}} \approx 0.72.$$

(c) The table is as follows:

		Use a leash		
		Agree or no.	Disagree	Total
Use for concerts	Agree or no.	137	9	146
	Disagree	2	10	12
	Total	139	19	158

(d) The relative risk can either be summarized as:

$$\frac{2/139}{10/19} \approx 0.03 \quad \text{or} \quad \frac{10/19}{2/139} \approx 36.6.$$

The proportion of those who disagree with using the park for summer concerts is 0.03 times lower in the group who agree or have no opinion about using leashes for dogs compared to those who disagree. Similarly, the proportion of those who disagree with using the park for summer concerts is 36.6 times higher in the group who also disagree with using leashes for dogs compared to those who do not disagree.

(e) The odds ratio is $\text{OR} = (137 \cdot 10)/(2 \cdot 9) \approx 36.1$.

- The chance of not disagreeing with the concert proposal is 36.1 times higher for those who also do not disagree with the leash proposal.
- The chance of not disagreeing with the leash proposal is 36.1 times higher for those who also do not disagree with the concert proposal.
- In simpler words: The chance of agreeing or having no opinion for one of the questions is 36.1 times higher if the person also has no opinion or agrees with the other question.

(f)

$$\gamma = \frac{137 \cdot 10 - 9 \cdot 2}{137 \cdot 10 + 9 \cdot 2} = \frac{1352}{1388} = 0.974.$$

(g) In general, it makes sense to use all the information available, i.e. to use the ordinal structure of the data and all three categories. While it is clear that γ is superior to V in our example, one may argue that the relative risks or the odds ratio could be more useful because they provide an intuitive quantification on how the two variables relate to each other rather than just giving a summary of strength and direction of association. However, as we have seen earlier, the interpretations of the relative risks and the odds ratio are quite clumsy in this example. It can be difficult to follow the somewhat complicated interpretation. A simple summary would be to say that agreement with both questions was strongly associated ($\gamma = 0.98$).

Solution to Exercise 4.7 We use the definition of the correlation coefficient, replace y_i with $a + bx_i$ and replace $\bar{y}$ with $a + b\bar{x}$ to obtain

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a + bx_i - (a + b\bar{x}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(a + bx_i - (a + b\bar{x}))^2}}.$$

This equates to:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(b(x_i - \bar{x}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(b(x_i - \bar{x}))^2}} = \frac{b\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sqrt{b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} = 1.$$

Solution to Exercise 4.8

(a) We read in the data, make sure the first column is recognized as containing the row names, attach the data, and obtain the correlation using the `cor` command:

```
decathlon <- read.csv('decathlon.csv', row.names=1)
attach(decathlon)
cor(X.Discus,X.High.jump)
```

The correlation is 0.52984. This means there is a moderate-to-strong positive correlation; i.e. the longer the distance the discus is thrown, the higher the height in the high jump competition.

(b) There are 10 variables. For the first variable, we can calculate the correlation with 9 other variables. For the second variable, we can also calculate the correlation with 9 other variables. However, we have already calculated one out of the 9 correlations, i.e. when analysing variable number one. So it is sufficient to calculate 8 correlations for the second variable. Similarly, we need another 7 correlations for the third variable, 6 correlations for the fourth variable, and so on. In total, we therefore need to have $9 + 8 + 7 + \cdots + 1 = 45$ correlations. Since the correlation coefficient describes the relationship between two variables, it

Fig. B.13 Correlation matrix for the decathlon data

	X.100m	X.Long.jump	X.Shot.put	X.High.jump	X.400m	X.110m.hurdle	X.Discus	X.Pole.vault	X.Javeline	X.1500m
X.100m	1	−0.7	−0.37	−0.31	0.63	0.54	−0.23	−0.26	−0.01	0.06
X.Long.jump	−0.7	1	0.2	0.35	−0.67	−0.54	0.25	0.29	0.09	−0.15
X.Shot.put	−0.37	0.2	1	0.61	−0.2	−0.25	0.67	0.02	0.38	0.13
X.High.jump	−0.31	0.35	0.61	1	−0.17	−0.33	0.52	−0.04	0.2	0
X.400m	0.63	−0.67	−0.2	−0.17	1	0.52	−0.14	−0.12	−0.05	0.55
X.110m.hurdle	0.54	−0.54	−0.25	−0.33	0.52	1	−0.22	−0.15	−0.08	0.18
X.Discus	−0.23	0.25	0.67	0.52	−0.14	−0.22	1	−0.18	0.25	0.22
X.Pole.vault	−0.26	0.29	0.02	−0.04	−0.12	−0.15	−0.18	1	−0.07	0.18
X.Javeline	−0.01	0.09	0.38	0.2	−0.05	−0.08	0.25	−0.07	1	−0.25
X.1500m	0.06	−0.15	0.13	0	0.55	0.18	0.22	0.18	−0.25	1

makes sense to summarize the results in a contingency table, similar to a matrix, see Fig. B.13.

(c) Using `cor(decathlon)` yields the correlation coefficients between all variable pairs. This is much simpler than calculating the correlation coefficient for each of the 45 comparisons. Note that the correlation matrix provides the 45 comparisons both in the upper triangle and in the lower triangle of the table. We know that $r(X, Y) = r(Y, X)$, but R still provides us with both, although they are identical. Note that the diagonal elements are 1 because $r(X, X) = 1$.

(d) One way to omit rows with missing data automatically is to use the `na.omit` command:

```
cor(na.omit(decathlon))
```

The results are displayed in Fig. B.13. We see moderate-to-strong correlations between the 100 m race, 400 m race, 110 m hurdle race and long jump. This may reflect the speed-and-athletic component of the decathlon contest. We also see moderate-to-strong correlations between the shot-put, high jump, and discus events. This may reflect the strength-and-technique component of the contest. The other disciplines also show correlations which make sense, but they are rather weak and may reflect the uniqueness of these disciplines.

Solution to Exercise 4.9

(a) A possible code is listed below:

```
pizza <- read.csv('pizza_delivery.csv')
pizza$tempcat <- cut(pizza$temperature, breaks=c(0,65,100))
pizza$timecat <- cut(pizza$time, breaks=c(0,30,100))
attach(pizza)
addmargins(table(tempcat,timecat))
```

```
          timecat
tempcat    (0,30] (30,100]  Sum
  (0,65]      101      691  792
  (65,100]    213      261  474
  Sum         314      952 1266
```

We can see that there is a higher proportion of high temperature ((65, 100]) in the category of short delivery times ((0, 30]) compared to long delivery times ((30, 100]).

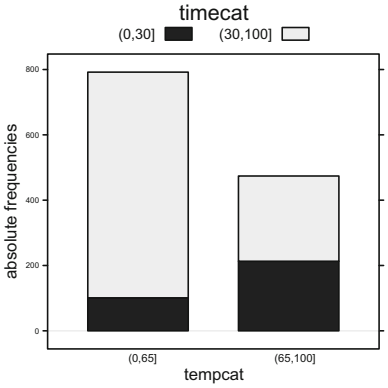
(b) Using the data from (a), we can calculate the odds ratio:

```
(101*261)/(213*691)
```

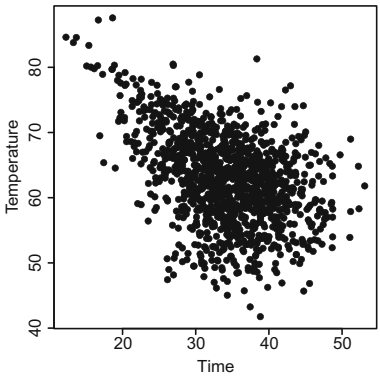
Thus, the chances of receiving a cold pizza are 0.18 lower if the delivery time is short.

(c) We use the `vcd` and `ryouready` packages to determine the desired measures of association:

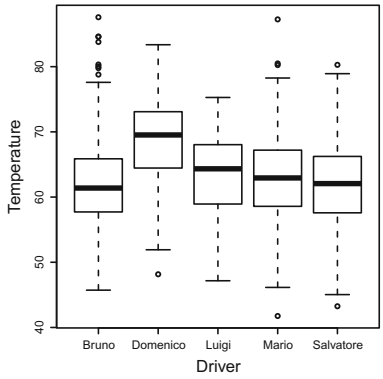
```
library(vcd)
library(ryouready)
library(lattice)
assocstats(xtabs(~tempcat+timecat))
ord.gamma(table(tempcat,timecat))
ord.tau(table(tempcat,timecat))
barchart(table(tempcat,timecat),horizontal=F,stack=T)
```



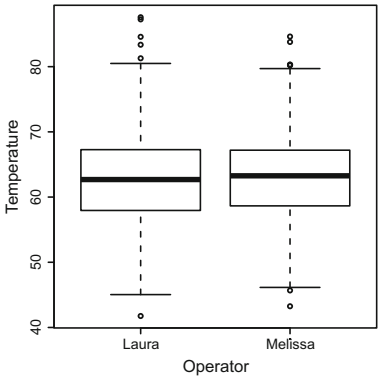
(a) Stacked bar chart for the categorical time and temperature variables



(b) Scatter plot for temperature and time



(c) Temperature by driver



(d) Temperature by operator

Fig. B.14 Plots for Exercise 4.9

Cramer's V is 0.361, Stuart's τ_c is -0.302, and γ is -0.696. The first two measures indicate a moderate-to-weak relationship between temperature and delivery time. It is clear from the last two measures that this relationship is negative, i.e. that shorter delivery times imply higher temperatures. However, interestingly, τ_c and γ provide us with a different strengths of association. In any case, it is clear that for shorter delivery times the customers receive warmer pizzas, as evident from the stacked bar chart (Fig. B.14a).

(d) The scatter plot (Fig. B.14b) shows a decreasing temperature for an increasing delivery time. This is also highlighted in the correlation coefficients which are -0.43 and -0.39 for Bravais–Pearson and Spearman, respectively.

```
plot(time,temperature)
cor(time,temperature)
cor(time,temperature,method='spearman')
```

(e) It makes sense to compare continuous variables (temperature, number of pizzas, bill) using the correlation coefficient from Bravais–Pearson. The relationships between temperature and driver and operator could be visualized using stratified box plots.

```
boxplot(temperature~driver)
boxplot(temperature~operator)
cor(temperature,pizzas)
cor(temperature,bill)
```

The correlation coefficients are -0.37 and -0.42 for number of pizzas and bill, respectively. More pizzas and a higher bill are associated with a lower temperature. The box plots (Fig. B.14c, d) show variation in the delivery times of the different drivers, but almost identical delivery times for the two operators. These results give us a first idea about the relationship in the data. However, they do not tell us the full story. For example: is the pizza temperature for higher bills low because a higher bill means more pizzas, and therefore, a less efficient preparation of the food? Is one driver faster because he mainly delivers for a particular branch? Could it be that the operators have a different performance but because they deal with different branches and drivers, these differences are not visible? To address these questions, a multivariate perspective is needed. Regression models, as introduced in Chap. 11, provide some answers.

Chapter 5

Solution to Exercise 5.1 There are $n = 10$ guests and $m = 2$ guests shake hands with each other. The order is not important: two guests shake hands no matter who is "drawn first". It is also not possible to shake hands with oneself which means that in terms of the urn model, we have the "no replacement" case. Therefore, we obtain the solution as

$$\binom{n}{m} = \binom{10}{2} = \frac{10 \cdot 9}{2 \cdot 1} = 45$$

handshakes in total.

Solution to Exercise 5.2 We assume that it is not relevant to know which student gets tested at which time point. We are thus dealing with combinations without considering the order. We have $n = 25$ and $m = 5$ and therefore obtain:

(a) a total number of $\binom{25}{5} = 53{,}130$ possibilities.

(b) a total number of $\binom{25+5-1}{5} = \binom{29}{5} = 118{,}755$ possibilities.

In *R*, the commands `choose(25,5)` and `choose(29,5)`, respectively provide the required results.

Solution to Exercise 5.3 The board consists of $n = 381$ knots. Each knot is either occupied or not. We may thus assume a "drawing" without replacement in the sense that each knot can be drawn (occupied) only once. If we place $m = 64$ counters on the board, then we can simultaneously think of "drawing" 64 occupied knots out of a total of 381 knots. The order in which we draw is not relevant—either a knot is occupied or not. The total number of combinations is $\binom{n}{m} = \binom{381}{64} \approx 4.35 \cdot 10^{73}$. We obtain the final number in *R* using the command `choose(381,64)`.

Solution to Exercise 5.4 We obtain the results (again using the command `choose(n,m)` in *R*) as follows:

(a) The customer takes the beers "with replacement" because the customer can choose among any type of beer for each position in the tray. One can also think of an urn model with 6 balls relating to the six different beers, where beers are drawn with replacement and placed on the tray. The order in which the beers are placed on the tray is not relevant. We thus have

$$\binom{n+m-1}{m} = \binom{6+20-1}{20} = \binom{25}{20} = 53,130$$

combinations.

(b) If the customer insists on having at least one beer per brewery on his tray, then 6 out of the 20 positions of the tray are already occupied. Based on the same thoughts as in (a), we calculate the total number of combinations as

$$\binom{n+m-1}{m} = \binom{6+14-1}{14} = \binom{19}{14} = 11,628.$$

Solution to Exercise 5.5 Since each team has exactly one final place, we have a "without replacement" situation. Using $n = 32$ and $m = 3$ (and `choose(n,m)` in *R*) yields

(a) $\frac{32!}{(32-3)!} = \binom{32}{3}3! = 29,760$ and

(b) $\binom{32}{3} = 4960$.

Solution to Exercise 5.6 There are $n = 12$ different letters for $m = 4$ positions of the membership code. Each letter can be used more than once if desired and we thus obtain $n^m = 12^4 = 20{,}736$ possible combinations. We therefore conclude that sufficient membership codes are left. However, this may not last long and the book store may still wish to create another solution for its membership codes.

Solution to Exercise 5.7 For each member of the jury, there are 61 scoring options:

$$\begin{array}{cccccccccc}
0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 & 0.7 & 0.8 & 0.9 \\
1 & 1.1 & 1.2 & 1.3 & 1.4 & 1.5 & 1.6 & 1.7 & 1.8 & 1.9 \\
. & & & & & . & & & & . \\
. & & & & & . & & & & . \\
5 & 5.1 & 5.2 & 5.3 & 5.4 & 5.5 & 5.6 & 5.7 & 5.8 & 5.9 \\
6 & & & & & & & & &
\end{array}$$

Different jury members are allowed to assign the same scores. We thus deal with combinations "with replacement". To verify this, just think of an urn with 61 balls where each ball refers to one possible score. Now one ball is drawn, assigned to a specific jury member and then put back into the urn. Since each score is "attached" to a particular jury member, we have combinations with consideration of the order and therefore obtain a total of $n^m = 61^9 \approx 1.17 \cdot 10^{16}$ possibilities. If you have difficulties in understanding the role of "replacement" and "order" in this example, recall that each member has 61 scoring options: thus, $61 \times 61 \times \cdots \times 61$ (9 times) combinations are possible.

Solution to Exercise 5.8

(a) We obtain:

$$\binom{2}{0} = 1 \leftrightarrow \binom{n}{2} = \binom{2}{2} = 1; \qquad \binom{3}{1} = 3 \leftrightarrow \binom{n}{2} = \binom{3}{2} = 3;$$
$$\binom{4}{2} = 6 \leftrightarrow \binom{n}{2} = \binom{4}{2} = 6; \qquad \binom{5}{3} = 10 \leftrightarrow \binom{n}{2} = \binom{5}{2} = 10.$$

(b) Based on the observations from (a) we conclude that each entry on the diagonal line can be represented by $\binom{n}{2}$. The sum of two consecutive entries is thus $\binom{n}{2} +$

$\binom{n+1}{2}$. Using the fourth relation in (5.5), it follows that:

$$\binom{n}{2} + \binom{n+1}{2} \overset{(5.5)}{=} \frac{n(n-1)}{2} + \frac{(n+1)n}{2}$$
$$= \frac{n(n-1+n+1)}{2} = \frac{n \cdot 2n}{2} = n^2.$$

Chapter 6

Solution to Exercise 6.1

(a) We obtain

- $A \cap B = \{8\}$.
- $B \cap C = \{\emptyset\}$.
- $A \cap C = \{0, 4, 8, 9, 15\}$.
- $C \setminus A = \{4, 9, 15\}$.
- $\Omega \setminus (B \cup A \cup C) = \{6, 7, 11, 13, 14\}$.

(b) The solutions are:

- $P(\bar{F}) = 1 - P(F) = 0.5$.
- $P(G) = 1 - P(E) - P(F) = 0.3$.
- $P(E \cap G) = 0$ because the events are pairwise disjoint.
- $P(E \setminus E) = 0$.
- $P(E \cup F) = P(E) + P(F) = 0.7$.

Solution to Exercise 6.2 We know that the probability of failing the practical examination is $P(PE) = 0.25$, of failing the theoretical examination is $P(TE) = 0.15$, and of failing both is $P(PE \cap TE) = 0.1$.

(a) If we ask for the probability of failing in *at least* one examination, we imply that either the theoretical examination, or the practical examination, or both are not passed. We can therefore calculate the union of events $P(PE \cup TE) = P(PE) + P(TE) - P(TE \cap PE) = 0.25 + 0.15 - 0.1 = 0.3$.

(b) $P(PE \setminus TE) = P(PE) - P(PE \cap TE) = 0.25 - 0.1 = 0.15$.

(c) $P(\overline{PE \cup TE}) = 1 - P(PE \cup TE) = 1 - 0.3 = 0.7$.

(d) We are interested in the probability of the person failing exactly in one exam. This corresponds to $P(M \backslash C \cup C \backslash M) = P(M \cup C) - P(C \cap M) = 0.3 - 0.1 = 0.2$.

Solution to Exercise 6.3 The total number of possible simple events is $|\Omega| = 12$. The number of favourable simple events is

(a) $|A| = 6$ (i.e. the numbers 2, 4, 6, 8, 10, 12). Hence, $P(A) = \frac{6}{12} = \frac{1}{2}$.

(b) $|B| = 3$ (i.e. the numbers 10, 11, 12). Hence, $P(B) = \frac{3}{12} = \frac{1}{4}$.

(c) $|C| = 2$ (i.e. the numbers 10, 12). Hence, $P(A \cap B) = \frac{2}{12} = \frac{1}{6}$.

(d) $|D| = 7$ (i.e. the numbers 2, 4, 6, 8, 10, 11, 12). Hence, $P(A \cup B) = \frac{7}{12}$.

Solution to Exercise 6.4 The total number of simple events is $\binom{12}{2}$.

(a) The number of favourable simple events is one and therefore $P(\text{right two presents}) = \frac{|A|}{|\Omega|} = \frac{1}{\binom{12}{2}} \approx 0.015$.

(b) The number of favourable simple events is $\binom{10}{2}$ because the person draws two presents out of the ten "wrong" presents:
$P(\text{wrong two presents}) = \frac{|A|}{|\Omega|} = \binom{10}{2} / \binom{12}{2} \approx 0.682$. In Sect. 8.1.8, we explain the hypergeometric distribution which will shed further light on this exercise.

Solution to Exercise 6.5

(a) Let V denote the event that there is too much salt in the soup and let L denote the event that the chef is in love. We know that

$$P(V) = 0.2 \Rightarrow P(\bar{V}) = 0.8.$$

Similarly, we have

$$P(L) = 0.3 \Rightarrow P(\bar{L}) = 0.7.$$

We therefore get:

$$\begin{aligned} P(V \cap L) &= P(V|L) \cdot P(L) = 0.6 \cdot 0.3 = 0.18. \\ P(\bar{V} \cap L) &= P(L) - P(V \cap L) = 0.3 - 0.18 = 0.12. \\ P(V \cap \bar{L}) &= P(V) - P(V \cap L) = 0.2 - 0.18 = 0.02. \\ P(\bar{V} \cap \bar{L}) &= P(\bar{V}) - P(\bar{V} \cap L) = 0.8 - 0.12 = 0.68. \end{aligned}$$

This yields the following contingency table:

	V	$\bar{V}$	Total
L	0.18	0.12	0.3
$\bar{L}$	0.02	0.68	0.7
Total	0.2	0.8	1

(b) The variables are not stochastically independent since, for example, $P(V) \cdot P(L) = 0.3 \cdot 0.2 = 0.06 \neq 0.18 = P(V \cap L)$.

Solution to Exercise 6.6

(a) We define the following events: G = Basil is treated well, $\bar{G}$ = Basil is not treated well; E = Basil survives, $\bar{E}$ = Basil dies. We know that

$$P(\bar{G}) = \frac{1}{3} \Rightarrow P(G) = \frac{2}{3}; \quad P(E|G) = \frac{1}{2}; \quad P(E|\bar{G}) = \frac{3}{4}.$$

Using the Law of Total Probability, we get

$$\begin{aligned} P(E) &= P(E|G) \cdot P(G) + P(E|\bar{G}) \cdot P(\bar{G}) \\ &= \frac{1}{2} \cdot \frac{2}{3} + \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{3} + \frac{1}{4} = \frac{7}{12} \approx 0.58. \end{aligned}$$

(b) We can use Bayes' Theorem to answer the question:

$$P(\bar{G}|E) = \frac{P(E|\bar{G}) \cdot P(\bar{G})}{P(E|\bar{G}) \cdot P(\bar{G}) + P(E|G) \cdot P(G)} = \frac{\frac{3}{4} \cdot \frac{1}{3}}{\frac{7}{12}} = \frac{3}{7} \approx 0.43.$$

Solution to Exercise 6.7 We define the following events and probabilities

- A: Bill never paid, $P(A) = 0.05 \Rightarrow P(\bar{A}) = 0.95$.
- M: Bill paid too late, $P(M) = ?$
- $P(M|\bar{A}) = 0.2$.
- $P(M|A) = 1$ because someone who never pays will always pay too late.

(a) We are interested in $P(M)$, the probability that someone does not pay his bill in a particular month, either because he is not able to or he pays too late. We can use the Law of Total Probability to obtain the results:

$$\begin{aligned} P(M) &= P(M|A)P(A) + P(M|\bar{A})P(\bar{A}) \\ &= 0.05 \cdot 1 + 0.2 \cdot 0.95 = 0.24. \end{aligned}$$

(b) We can use Bayes' Theorem to obtain the results:

$$P(A \mid M) = \frac{P(A)P(M \mid A)}{P(M)} = \frac{0.05}{0.24} = 0.208.$$

(c) If the bill was not paid in a particular month, the probability is 20.8 % that it will never be paid, and 78.2 % that it will be paid. One could argue that a preventive measure that affects almost 79 % of trustworthy customers are not ideal and the bank should therefore not block a credit card if a bill is not paid on time.

Solution to Exercise 6.8

(a) The "and" operator refers to the joint distribution of two variables. In our example, we want to know the probability of being infected *and* having been transported by the truck. This probability can be directly obtained from the respective entry in the contingency table: 40 out of 200 cows fulfil both criteria and thus

$$P(B \cap A) = \frac{40}{200}.$$

(b) We can use $P(A) = \frac{100}{200} = P(\bar{A})$ to obtain:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{40/200}{100/200} = \frac{40}{100}.$$

(c) Using these results and $P(B) = \frac{60}{200}$, and $P(\bar{B}) = \frac{140}{200} = 1 - P(B)$, we obtain

$$P(B|\bar{A}) = \frac{P(B \cap \bar{A})}{P(\bar{A})} = \frac{20/200}{100/200} = \frac{20}{100}$$

by using the Law of Total Probability. We can thus calculate

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \\ &= 0.40 \cdot 0.50 + 0.20 \cdot 0.50 = 0.30. \end{aligned}$$

This means that the probability of a cow being infected is 30 %. Alternatively, we could have simply looked at the marginal distribution of the contingency table to get $P(B) = 60/200 = 0.3$.

Solution to Exercise 6.9

(a) The two shots are independent of each other and thus

$$P(A \cap B) = 0.4 \cdot 0.5 = 0.2.$$
$$P(A \cup B) = 0.4 + 0.5 - 0.2 = 0.7.$$

(b) We need to calculate

$$P(A \backslash B \cup B \backslash A) = 0.4 - 0.2 + 0.5 - 0.2 = 0.5.$$

(c)

$$P(B \backslash A) = 0.5 - 0.2 = 0.3.$$

Chapter 7

Solution to Exercise 7.1

(a) The first derivative of the CDF yields the PDF, $F'(x) = f(x)$:

$$f(x) = \begin{cases} 0 & \text{if } x < 2 \\ -\frac{1}{2}x + 2 & \text{if } 2 \le x \le 4 \\ 0 & \text{if } x > 4. \end{cases}$$

(b) We know from Theorem 7.2.3 that for any continuous variable $P(X = x_0) = 0$ and therefore $P(X = 4) = 0$. We calculate $P(X < 3) = P(X \le 3) - P(X = 3) = F(3) - 0 = -\frac{9}{4} + 6 - 3 = 0.75$.

(c) Using (7.15), we obtain the expectation as

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x\ f(x)\ \mathrm{d}x = \int_{-\infty}^{2} x0\ \mathrm{d}x + \int_{2}^{4} x\ \left(-\frac{1}{2}x + 2\right)\ \mathrm{d}x + \int_{4}^{\infty} x0\ \mathrm{d}x \\ &= 0 + \int_{2}^{4} \left(-\frac{1}{2}x^2 + 2x\right) \mathrm{d}x + 0 \\ &= \left[-\frac{1}{6}x^3 + x^2\right]_2^4 = \left(-\frac{64}{6} + 16\right) - \left(-\frac{8}{6} + 4\right) = \frac{8}{3}. \end{aligned}$$

Given that we have already calculated $\mathrm{E}(X)$, we can use Theorem 7.3.1 to calculate the variance as $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$. The expectation of X^2 is

$$\begin{aligned} \mathrm{E}(X^2) &= \int_{2}^{4} x^2 \left(-\frac{1}{2}x + 2\right) \mathrm{d}x = \int_{2}^{4} \left(-\frac{1}{2}x^3 + 2x^2\right) \mathrm{d}x \\ &= \left[-\frac{1}{8}x^4 + \frac{2}{3}x^3\right]_2^4 = \left(-32 + \frac{128}{3}\right) - \left(-2 + \frac{16}{3}\right) = \frac{22}{3}. \end{aligned}$$

We thus obtain $\mathrm{Var}(X) = \frac{22}{3} - \left(\frac{8}{3}\right)^2 = \frac{66-64}{9} = \frac{2}{9}$.

Solution to Exercise 7.2

(a) The probability mass function of X is

x_i	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{3}{9}$

Using (7.16), we calculate the expectation as

$$\begin{aligned} \mathrm{E}(X) &= 1 \cdot \frac{1}{9} + 2 \cdot \frac{1}{9} + 3 \cdot \frac{1}{9} + 4 \cdot \frac{2}{9} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{3}{9} \\ &= \frac{1 + 2 + 3 + 8 + 5 + 18}{9} = \frac{37}{9} \approx 4.1. \end{aligned}$$

To obtain the variance, we need

$$\begin{aligned}\mathrm{E}(X^2) &= 1 \cdot \frac{1}{9} + 4 \cdot \frac{1}{9} + 9 \cdot \frac{1}{9} + 16 \cdot \frac{2}{9} + 25 \cdot \frac{1}{9} + 36 \cdot \frac{3}{9} \\ &= \frac{1+4+9+32+25+108}{9} = \frac{179}{9}.\end{aligned}$$

Therefore, using $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$, we get

$$\mathrm{Var}(X) = \frac{179}{9} - \left(\frac{37}{9}\right)^2 = \frac{1611 - 1369}{81} = \frac{242}{81} \approx 2.98.$$

The manipulated die yields on average higher values than a fair die because its expectation is $4.1 > 3.5$. The variability of is, however, similar because $2.98 \approx 2.92$.

(b) The probability mass function of $Y = \frac{1}{X}$ is:

$y_i = \frac{1}{x_i}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$
$P(\frac{1}{X} = y)$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{3}{9}$

The expectation can hence be calculated as

$$\begin{aligned}\mathrm{E}(Y) = \mathrm{E}\left(\frac{1}{X}\right) &= 1 \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9} + \frac{1}{3} \cdot \frac{1}{9} + \frac{1}{4} \cdot \frac{2}{9} + \frac{1}{5} \cdot \frac{1}{9} + \frac{1}{6} \cdot \frac{3}{9} \\ &= \frac{1}{9} + \frac{1}{18} + \frac{1}{27} + \frac{1}{18} + \frac{1}{45} + \frac{1}{18} = \frac{91}{270}.\end{aligned}$$

Comparing the results from (a) and (b) shows clearly that $\mathrm{E}(\frac{1}{X}) \neq \frac{1}{\mathrm{E}(X)}$. Recall that $\mathrm{E}(bX) = b\mathrm{E}(X)$. It is interesting to see that for some transformations $T(X)$ it holds that $\mathrm{E}(T(X)) = T(\mathrm{E}(X))$, but for some it does not. This reminds us to be careful when thinking of the implications of transformations.

Solution to Exercise 7.3

(a) There are several ways to plot the CDF. One possibility is to define the function and plot it with the `curve` command. Since the function has different definitions for the intervals $[\infty, 0)$, $[0, 1]$, $(1, \infty]$, we need to take this into account. Remember that a logical statement in *R* corresponds to a number, i.e. `TRUE` = 1 and `FALSE` = 0; we can thus simply add the different pieces of the function and multiply them with a condition which specifies if X is contained in the interval or not (Fig. B.15):

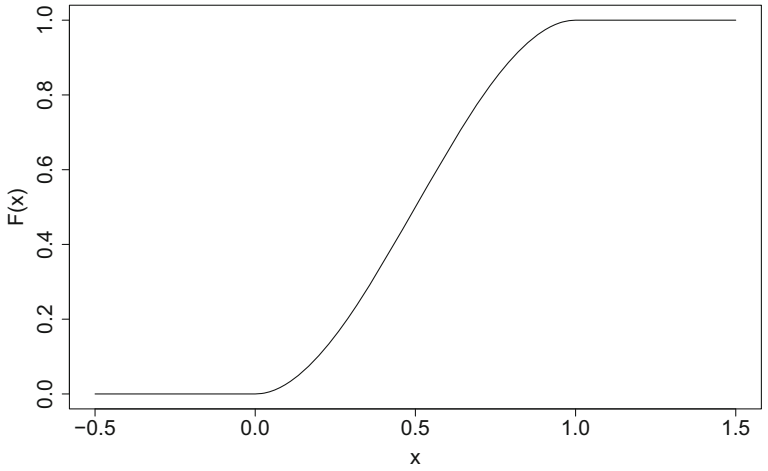


Fig. B.15 Cumulative distribution function for the proportion of wine sold

```
cdf <−function(x){
(3 ∗ x^2 − 2 ∗ x^3) ∗ (x <= 1 & x >= 0) + 1 ∗ (x > 1) + 0 ∗ (x < 0)
}
curve(cdf,from=-0.5,to=1.5)
```

(b) The PDF is

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = F'(x) = f(x) = \begin{cases} 6(x - x^2) & \text{if } 0 \le x \le 1 \\ 0 & \text{elsewhere.} \end{cases}$$

(c)

$$\begin{aligned} P\left(\frac{1}{3} \le X \le \frac{2}{3}\right) &= \int_{\frac{1}{3}}^{\frac{2}{3}} f(x)\mathrm{d}x = F\left(\frac{2}{3}\right) - F\left(\frac{1}{3}\right) \\ &= \left[3\left(\frac{2}{3}\right)^2 - 2\left(\frac{2}{3}\right)^3\right] - \left[3\left(\frac{1}{3}\right)^2 - 2\left(\frac{1}{3}\right)^3\right] \\ &= 0.48149. \end{aligned}$$

(d) We have already defined the CDF in (a). We can now simply plug in the x-values of interest:

```
cdf(2/3)-cdf(1/3)
```

(e) The variance can be calculated as follows:

$$\begin{aligned}
\mathrm{E}(X) &= \int_0^1 x6(x-x^2)\mathrm{d}x = 6\cdot\int_0^1 (x^2-x^3)\mathrm{d}x \\
&= 6\left[\frac{1}{3}x^3-\frac{1}{4}x^4\right]_0^1 = 6\cdot\left(\frac{1}{3}-\frac{1}{4}\right) = 0.5 \\
\mathrm{E}(X^2) &= \int_0^1 x^2 6(x-x^2)\mathrm{d}x = 6\cdot\int_0^1 (x^3-x^4)\mathrm{d}x \\
&= 6\left[\frac{1}{4}x^4-\frac{1}{5}x^5\right]_0^1 = 6\cdot\left(\frac{1}{4}-\frac{1}{5}\right) = 0.3 \\
\mathrm{Var}(X) &= \mathrm{E}(X^2)-[\mathrm{E}(X)]^2 = 0.3-0.5^2 = 0.05.
\end{aligned}$$

Solution to Exercise 7.4

(a) Two conditions need to be satisfied for $f(x)$ to be a proper PDF:

(i) $\int_0^2 f(x)\mathrm{d}x = 1$:

$$\begin{aligned}
\int_0^2 f(x)\mathrm{d}x &= \int_0^2 c\cdot x(2-x)\mathrm{d}x = c\int_0^2 x(2-x)\mathrm{d}x \\
&= c\int_0^2 (2x-x^2)\mathrm{d}x = c\left[x^2-\frac{1}{3}x^3\right]_0^2 \\
&= c\left[4-\frac{8}{3}-(0-0)\right] = c\cdot\frac{4}{3}\overset{!}{=}1 \\
\Longrightarrow c &= \frac{3}{4}.
\end{aligned}$$

(ii) $f(x)\geq 0$:

$$f(x) = \frac{3}{4}x(2-x)\geq 0 \quad \forall\ x\in[0,2].$$

(b) We calculate

$$\begin{aligned}
F(x) = P(X\leq x) &= \int_0^x f(t)\mathrm{d}t = \int_0^x \frac{3}{4}t(2-t)\mathrm{d}t \\
&= \frac{3}{4}\int_0^x (2t-t^2)\mathrm{d}t = \frac{3}{4}\left[t^2-\frac{1}{3}t^3\right]_0^x \\
&= \frac{3}{4}\left[x^2-\frac{1}{3}x^3-0\right] = \frac{3}{4}x^2\left(1-\frac{1}{3}x\right)
\end{aligned}$$

and therefore

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{3}{4}x^2(1 - \frac{1}{3}x) & \text{if } 0 \le x \le 2 \\ 1 & \text{if } 2 < x. \end{cases}$$

(c) The expectation is

$$\begin{aligned} E(X) &= \int_0^2 x f(x)\mathrm{d}x = \frac{3}{4}\int_0^2 (2x^2 - x^3)\mathrm{d}x \\ &= \frac{3}{4}\left[\frac{2}{3}x^3 - \frac{1}{4}x^4\right]_0^2 = \frac{3}{4}\left[\frac{2}{3}\cdot 8 - \frac{1}{4}\cdot 16 - 0\right] \\ &= \frac{3}{4}\left[\frac{16}{3} - \frac{12}{3}\right] = \frac{3}{4}\cdot\frac{4}{3} = 1. \end{aligned}$$

Using $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$, we calculate the variance as

$$\begin{aligned} E(X^2) &= \int_0^2 x^2 f(x)\mathrm{d}x = \frac{3}{4}\int_0^2 (2x^3 - x^4)\mathrm{d}x \\ &= \frac{3}{4}\left[\frac{2}{4}x^4 - \frac{1}{5}x^5\right]_0^2 = \frac{3}{4}\left[\frac{2}{4}\cdot 16 - \frac{1}{5}\cdot 32 - 0\right] \\ &= 6 - \frac{3\cdot 32}{4\cdot 5} = 6 - \frac{3\cdot 8}{5} = \frac{6}{5} \end{aligned}$$

$$\mathrm{Var}(X) = \frac{6}{5} - 1^2 = \frac{1}{5}.$$

(d)

$$P(|X - \mu| \le 0.5) \ge 1 - \frac{\sigma^2}{c^2} = 1 - \frac{(\frac{1}{5})}{(0.5)^2} = 1 - 0.8 = 0.2.$$

Solution to Exercise 7.5

(a) The marginal distributions are obtained by the row and column sums of the joint PDF, respectively. For example, $P(X = 1) = \sum_{j=1}^{J} p_{1j} = p_{1+} = 1/4$.

X	$P(X = x_i)$
0	3/4
1	1/4

Y	$P(Y = y_i)$
1	1/6
2	7/12
3	1/4

The marginal distribution of X tells us how many customers sought help via the telephone hotline (75 %) and via email (25 %). The marginal distribution of Y represents the distribution of the satisfaction level, highlighting that more than half of the customers (7/12) were "satisfied".

(b) To determine the 75 % quantile with respect to Y, we need to find the value $y_{0.75}$ for which $F(y_{0.75}) \geq 0.75$ and $F(y) < 0.75$ for $y < y_{0.75}$. Y takes the values 1, 2, 3. The quantile cannot be $y_{0.75} = 1$ because $F(1) = 1/6 < 0.75$. The 75 % quantile is $y_{0.75} = 2$ because $F(2) = 1/6 + 7/12 = 3/4 \geq 0.75$ and for all values which are smaller than 2 we get $F(x) < 0.75$.

(c) We can calculate the conditional distribution using $P(Y = y_j|X = 1) = p_{1j}/p_{1+} = p_{1j}/(1/6 + 1/12 + 0) = p_{1j}/(0.25)$. Therefore,

$$P(Y = 1|X = 1) = \frac{1/6}{1/4} = \frac{2}{3},$$
$$P(Y = 2|X = 1) = \frac{1/12}{1/4} = \frac{1}{3},$$
$$P(Y = 3|X = 1) = \frac{0}{1/4} = 0.$$

Among those who used the email customer service two-thirds were unsatisfied, one-third were satisfied, and no one was very satisfied.

(d) As we know from (7.27), two discrete random variables are said to be independent if $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$. However, in our example, $P(X = 0, Y = 1) = P(X = 0)P(X = 1) = \frac{3}{4} \cdot \frac{1}{6} \neq 0$. This means that X and Y are not independent.

(e) The covariance of X and Y is defined as $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$. We calculate

$$\mathrm{E}(X) = 0 \cdot \frac{3}{4} + 1 \cdot \frac{1}{4} = \frac{1}{4}$$
$$\mathrm{E}(Y) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{7}{12} + 3 \cdot \frac{1}{4} = \frac{25}{12}$$
$$\mathrm{E}(XY) = 0 \cdot 1 \cdot 0 + 1 \cdot 1 \cdot \frac{1}{6} + 0 \cdot 2 \cdot \frac{1}{2} + 1 \cdot 2 \cdot \frac{1}{12} + 0 \cdot 3 \cdot \frac{1}{4} + 1 \cdot 3 \cdot 0$$
$$= \frac{2}{6}$$
$$\mathrm{Cov}(X, Y) = \frac{2}{6} - \frac{1}{4} \cdot \frac{25}{12} = -\frac{3}{16}.$$

Since $\mathrm{Cov}(X, Y) < 0$, we conclude that there is a negative relationship between X and Y: the higher the values of X, the lower the values of Y—and vice versa. For example, those who use the email-based customer service ($X = 1$) are less satisfied than those who use the telephone customer service ($X = 0$). It is, however, evident that in this example the values of X have no order and therefore, care must be exercised in interpreting the covariance.

Solution to Exercise 7.6 Using Tschebyschev's inequality (7.24)

$$P(|X-\mu|<c) \geq 0.9 = 1 - \frac{\mathrm{Var}(X)}{c^2},$$

we can determine c as follows:

$$\begin{aligned} 1 - \frac{\mathrm{Var}(X)}{c^2} &= 0.9 \\ c^2 = \frac{\mathrm{Var}(X)}{0.1} &= \frac{4}{0.1} = 40 \\ c &= \pm\sqrt{40} = \pm 6.325. \end{aligned}$$

Thus, the interval is $[15 - 6.325; 15 + 6.325] = [8.675; 21.325]$.

Solution to Exercise 7.7

(a) The joint PDF is:

		Y		
		0	1	2
X	−1	0.3	0.2	0.2
	2	0.1	0.1	0.1

(b) The marginal distributions are obtained from the row and column sums of the joint PDF, respectively:

X	−1	2
$P(X=x)$	0.7	0.3

Y	0	1	2
$P(Y=y)$	0.4	0.3	0.3

(c) The random variables X and Y are independent if

$$P(X=x, Y=y) = P(X=x)P(Y=y) \quad \forall x, y.$$

However, in our example we have, for example,

$$P(X=-1, Y=0) = 0.3 \neq P(X=-1) \cdot P(Y=0) = 0.7 \cdot 0.4 = 0.28.$$

Hence, the two variables are not independent.

(d) The joint distribution of X and Y can be used to obtain the desired distribution of U. For example, If $X=-1$ and $Y=0$, then $U = X+Y = -1$. The respective probability is $P(U=-1) = 0.3$ because $P(U=-1) = P(X=-1, Y=0) = 0.3$ and there is no other combination of X- and Y-values which yields $X+Y=-1$. The distribution of U is therefore as follows:

k	-1	0	1	2	3	4
$P(U=k)$	0.3	0.2	0.2	0.1	0.1	0.1

(e) We calculate

$$\begin{aligned}
\mathrm{E}(U) &= \sum_{k=-1}^{4} k \cdot P(U=k) = 0.8 \\
\mathrm{E}(X) &= (-1)0.7 + 2 \cdot 0.3 = -0.1 \\
\mathrm{E}(Y) &= 0 \cdot 0.4 + 1 \cdot 0.3 + 2 \cdot 0.3 = 0.9
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(U^2) &= 0.3 \cdot (-1)^2 + \cdots + 0.1 \cdot 4^2 = 3.4 \\
\mathrm{E}(X^2) &= 0.7 \cdot (-1)^2 + 0.3 \cdot 2^2 = 1.9 \\
\mathrm{E}(Y^2) &= 0.4 \cdot 0^2 + 0.3 \cdot 1^2 + 0.3 \cdot 2^2 = 1.5
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(U) &= \mathrm{E}(U^2) - [\mathrm{E}(U)]^2 = 3.4 - (0.8)^2 = 2.76 \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 1.9 - (-0.1)^2 = 1.89 \\
\mathrm{Var}(Y) &= \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 = 1.5 - (0.9)^2 = 0.69.
\end{aligned}$$

It can be seen that $\mathrm{E}(X) + \mathrm{E}(Y) = -0.1 + 0.9 = 0.8 = \mathrm{E}(U)$. This makes sense because we know from (7.31) that $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$. However, $\mathrm{Var}(U) = 2.76 \neq \mathrm{Var}(X) + \mathrm{Var}(Y) = 1.89 + 0.69$. This follows from (7.7.1) which says that $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$ and therefore, $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ only if the covariance is 0. We know from (c) that X and Y are not independent and thus $\mathrm{Cov}(X, Y) \neq 0$.

Solution to Exercise 7.8

(a) The random variables X and Y are independent if the balls are drawn with replacement. This becomes clear by understanding that drawing with replacement implies that for both the draws, the same balls are in the urn and the conditions in each draw remain the same. The first draw has no implications for the second draw.

If we were drawing the balls without replacement, then the first draw could possibly have implications for the second draw: for instance, if the first ball drawn was red, then the second one could not be red because there is only one red ball in the urn. This means that drawing without replacement implies dependency of X and Y. This can also be seen by evaluating the independence assumption (7.27):

$$P(X=2, Y=2) = 0 \neq P(X=2) \cdot P(Y=2) = \frac{1}{8} \cdot \frac{1}{8}.$$

(b) The marginal probabilities $P(X = x_i)$ can be obtained from the given information. For example, 3 out of 8 balls are black and thus $P(X = 1) = 3/8$. The conditional distributions $P(Y|X = x_i)$ can be calculated easily by realizing that under the assumed dependency of X and Y, the second draw is always based on 7 balls (8 balls minus the one drawn under the condition $X = x_i$)—e.g. if the first ball drawn is black, then 7 balls, 2 of which are black, remain in the urn and $P(Y = 1|X = 1) = 2/7$. We thus calculate

$$\begin{aligned}
P(Y = 1, X = 1) &= P(Y = 1|X = 1)P(X = 1) = \frac{2}{7} \cdot \frac{3}{8} = \frac{6}{56} \\
P(Y = 1, X = 2) &= P(Y = 1|X = 2)P(X = 2) = \frac{3}{7} \cdot \frac{1}{8} = \frac{3}{56} \\
&\dots \\
P(Y = 3, X = 3) &= P(Y = 3|X = 3)P(X = 3) = \frac{3}{7} \cdot \frac{4}{8} = \frac{12}{56}
\end{aligned}$$

and obtain

		Y		
		1	2	3
X	1	$\frac{6}{56}$	$\frac{3}{56}$	$\frac{12}{56}$
	2	$\frac{3}{56}$	0	$\frac{4}{56}$
	3	$\frac{12}{56}$	$\frac{4}{56}$	$\frac{12}{56}$

(c) The expectations are

$$\begin{aligned}
\mathrm{E}(X) &= 1 \cdot \frac{3}{8} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{4}{8} = \frac{17}{8} \\
\mathrm{E}(Y) &= \mathrm{E}(X) = \frac{17}{8}.
\end{aligned}$$

To estimate $\rho(X, Y)$, we need $\mathrm{Cov}(X, Y)$ as well as $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$:

$$\begin{aligned}
\mathrm{E}(XY) &= 1\frac{6}{56} + 2\frac{3}{56} + 3\frac{12}{56} + 2\frac{3}{56} + 4 \cdot 0 + 6\frac{4}{56} + 3\frac{12}{56} + 6\frac{4}{56} + 9\frac{12}{56} \\
&= \frac{246}{56} \\
\mathrm{E}(X^2) &= \mathrm{E}(Y^2) = 1^2\frac{3}{8} + 2^2\frac{1}{8} + 3^2\frac{4}{8} = \frac{43}{8} \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \frac{43}{8} - \left(\frac{17}{8}\right)^2 = \frac{55}{64} \\
\mathrm{Var}(Y) &= \mathrm{Var}(X) = \frac{55}{64}.
\end{aligned}$$

Using (7.38) and $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$, we obtain

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}} = \frac{\frac{246}{56} - \frac{289}{64}}{\sqrt{\frac{55}{64} \cdot \frac{55}{64}}} = -0.143.$$

Solution to Exercise 7.9

(a) The constant c must satisfy

$$\int_{40}^{100} \int_{10}^{100} c \left(\frac{100 - x}{x} \right) \mathrm{d}x \, \mathrm{d}y = \int_{40}^{100} \int_{10}^{100} \frac{100c}{x} - c \, \mathrm{d}x \, \mathrm{d}y \stackrel{!}{=} 1$$

and therefore

$$\int_{40}^{100} [100c \ln(x) - cx]_{10}^{100} \, \mathrm{d}y = \int_{40}^{100} 100c \ln 100 - 100c - 100c \ln(10) + 10c \, \mathrm{d}y$$

which is

$$\begin{aligned} \int_{40}^{100} 100c \left(\ln \frac{100}{10} - 1 + \frac{1}{10} \right) \mathrm{d}y &= [100cy \, (\ln 10 - 9/10)]_{40}^{100} \\ &= 600c(10 \ln 10 - 9) \quad \rightarrow c \approx 0.00012. \end{aligned}$$

(b) The marginal distribution is

$$\begin{aligned} f_X(x) &= \int_{40}^{100} c \left(\frac{100 - x}{x} \right) \mathrm{d}y = \left[c \left(\frac{100 - x}{x} \right) y \right]_{40}^{100} \\ &= 100c \left(\frac{100 - x}{x} \right) - 40c \left(\frac{100 - x}{x} \right) \approx 0.00713 \left(\frac{100 - x}{x} \right) \end{aligned}$$

for $10 \leq x \leq 100$.

(c) To determine $\mathrm{P}(X > 75)$, we need the cumulative marginal distribution of X:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t) \mathrm{d}t = \int_{10}^{x} 0.00713 \left(\frac{100 - t}{t} \right) \mathrm{d}t \\ &= \int_{10}^{x} \frac{0.00713}{t} - 0.00713 \, \mathrm{d}t = [0.713 \ln(t) - 0.00713]_{10}^{x} \\ &= 0.713 \ln(x) - 0.00713x - 0.00713 \ln(10) + 0.00713 \cdot 10. \end{aligned}$$

Now we can calculate

$$\mathrm{P}(X > 75) = 1 - P(X \leq 75) = 1 - F_X(75) = 1 - 0.973 \approx 2.7 \, \%.$$

(d) The conditional distribution is

$$f_{Y|X}(x, y) = \frac{f(x, y)}{f(x)} = \frac{c\left(\frac{100-x}{x}\right)}{60c\left(\frac{100-x}{x}\right)} = \frac{1}{60}.$$

Solution to Exercise 7.10 If we evaluate the expectation with respect to Y, then both μ and σ can be considered to be constants. We can therefore write

$$\mathrm{E}(Y) = \mathrm{E}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}(E(X) - \mu).$$

Since $E(X) = \mu$, it follows that $E(Y) = 0$. The variance is

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right).$$

Applying $\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$ to this equation yields $a = \mu$, $b = \frac{1}{\sigma}$ and therefore

$$\mathrm{Var}(Y) = \frac{1}{\sigma^2}\mathrm{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1.$$

Chapter 8

Solution to Exercise 8.1 The random variable X: "number of packages with a toy" is binomially distributed. In each of $n = 20$ "trials", a toy can be found with probability $p = \frac{1}{6}$.

(a) We thus get

$$P(X = 4) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{20}{4}\binom{1}{6}^4\binom{5}{6}^{16} \approx 0.20.$$

(b) Similarly, we calculate

$$P(X = 0) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{20}{0}\binom{1}{6}^0\binom{5}{6}^{20} = 0.026.$$

(c) This question relates to a hypergeometric distribution: there are $N = 20$ packages with $M = 3$ packages with toys and $N - M = 17$ packages without a toy. The daughter gets $n = 5$ packages and we are interested in $P(X = 2)$. Hence, we get

$$P(X = 2) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{3}{2}\binom{17}{3}}{\binom{20}{5}} \approx 0.13.$$

Solution to Exercise 8.2 Given $X \sim N(42.1, 20.8^2)$, we get:

(a)

$$\begin{aligned} P(X \geq 50) &= 1 - P(X \leq 50) = 1 - \phi\left(\frac{x-\mu}{\sigma}\right) = 1 - \phi\left(\frac{50-42.1}{20.8}\right) \\ &= 1 - \phi(0.37) \approx 0.35. \end{aligned}$$

We obtain the same results in *R* as follows:

```
1-pnorm(50,42.1,20.8)
```

(b)

$$\begin{aligned} P(30 \leq X \leq 40) &= P(X \leq 40) - P(X \leq 30) \\ &= \phi\left(\frac{40-42.1}{20.8}\right) - \phi\left(\frac{30-42.1}{20.8}\right) \\ &= \phi(-0.096) - \phi(-0.577) = 1 - 0.538 - 1 + 0.718 \\ &\approx 18\,\%. \end{aligned}$$

We would have obtained the same results in *R* using:

```
pnorm(40,42.1,20.8)-pnorm(30,42.1,20.8)
```

Solution to Exercise 8.3 The random variable X follows a discrete uniform distribution because $p_i = \frac{1}{12}$ for each x_i. The expectation and variance are therefore

$$\begin{aligned} \mathrm{E}(X) &= \frac{k+1}{2} = \frac{12+1}{2} = 6.5, \\ \mathrm{Var}(X) &= \frac{1}{12}(12^2 - 1) \approx 11.92. \end{aligned}$$

Solution to Exercise 8.4 Each guess is a Bernoulli experiment where the right answer is given with a probability of 50 %. The number of correct guesses therefore follows a binomial distribution, i.e. $X \sim B(10; 0.5)$. The probability of giving the right answer at least 8 times is identical to the probability of not being wrong more

than 2 times. We can thus calculate $P(X \geq 8)$ as $P(X \leq 2)$:

$$P(X = 0) = \binom{10}{0} 0.5^0 (1 - 0.5)^{10} \approx 0.000977$$
$$P(X = 1) = \binom{10}{1} 0.5^1 (1 - 0.5)^9 \approx 0.009766$$
$$P(X = 2) = \binom{10}{2} 0.5^2 (1 - 0.5)^8 \approx 0.043945.$$

This relates to

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.000977 + 0.009766 + 0.043945 \approx 0.0547. \end{aligned}$$

We would have obtained the same results in *R* using:

```
pbinom(2,10,0.5)
1-pbinom(7,10,0.5)
```

Solution to Exercise 8.5

(a) It seems appropriate to model the number of fused bulbs with a Poisson distribution. We assume, however, that the probabilities of fused bulbs on two consecutive days are independent of each other; i.e. they only depend on λ but not on the time t.

(b) The arithmetic mean is

$$\bar{x} = \frac{1}{30}(0 + 1 \cdot 8 + 2 \cdot 8 + \cdots + 5 \cdot 1) = \frac{52}{30} = 1.7333$$

which means that, on an average, 1.73 bulbs are fused per day. The variance is

$$\begin{aligned} s^2 &= \frac{1}{30}(0 + 1^2 \cdot 8 + 2^2 \cdot 8 + \cdots + 5^2 \cdot 1) - 1.7333^2 \\ &= \frac{142}{30} - 3.0044 = 1.72889. \end{aligned}$$

We see that mean and variance are similar, which is an indication that the choice of a Poisson distribution is appropriate since we assume $\mathrm{E}(X) = \lambda$ and $\mathrm{Var}(X) = \lambda$.

(c) The following table lists the proportions (i.e. relative frequencies f_j) together with the probabilities $P(X = x)$ from a $Po(1.73)$-distribution. As a reference, we also list the probabilities from a $Po(2)$-distribution since it is not practically possible that 1.73 bulbs stop working and it may hence be an option to round the mean.

	f_i	$Po(1.73)$	$Po(2)$
$P(X=0)$	0.2	0.177	0.135
$P(X=1)$	0.267	0.307	0.27
$P(X=2)$	0.267	0.265	0.27
$P(X=3)$	0.167	0.153	0.18
$P(X=4)$	0.067	0.067	0.09
$P(X=5)$	0.033	0.023	0.036

One can see that observed proportions and expected probabilities are close together which indicates again that the choice of a Poisson distribution was appropriate. Chapter 9 gives more details on how to estimate parameters, such as λ, from data if it is unknown.

(d) Using $\lambda = 1.73$, we calculate

$$\begin{aligned} P(X > 5) &= 1 - P(X \leq 5) = 1 - \sum_{i=0}^{5} \frac{\lambda^i}{i!} \exp(-\lambda) \\ &= 1 - \exp(-1.73)\left(\frac{1.73^0}{0!} + \frac{1.73^1}{1!} + \cdots + \frac{1.73^5}{5!}\right) \\ &= 1 - 0.99 = 0.01. \end{aligned}$$

Thus, the bulbs are replaced on only 1 % of the days.

(e) If X follows a Poisson distribution then, given Theorem 8.2.1, Y follows an exponential distribution with $\lambda = 1.73$.

(f) The expectation of an exponentially distributed variable is

$$E(Y) = \frac{1}{\lambda} = \frac{1}{1.73} = 0.578.$$

This means that, on average, it takes more than half a day until one of the bulbs gets fused.

Solution to Exercise 8.6

(a) Let X be a random variable describing "the number x of winning tickets among n bought tickets"; then X follows the hypergeometric distribution $X \sim H(n, 500, 4000)$. We need to determine n for the conditions specified. We are interested in

$$P(X \geq 3) = 1 - P(X = 2) - P(X = 1) - P(X = 0).$$

Using the PMF of the hypergeometric distribution

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

this equates to

$$P(X \geq 3) = 1 - \frac{\binom{500}{2}\binom{4000-500}{n-2}}{\binom{4000}{n}} - \frac{\binom{500}{1}\binom{4000-500}{n-1}}{\binom{4000}{1}} - \frac{\binom{500}{0}\binom{4000-500}{n}}{\binom{4000}{n}}.$$

We have the following requirement:

$$1 - \frac{\binom{500}{2}\binom{4000-500}{n-2}}{\binom{4000}{n}} - \frac{\binom{500}{1}\binom{4000-500}{n-1}}{\binom{4000}{1}} - \frac{\binom{500}{0}\binom{4000-500}{n}}{\binom{4000}{n}} \overset{!}{\geq} 0.99.$$

To solve this equation, we can program this function for $P(X > 3; n)$ in R and evaluate it for different numbers of tickets sold, e.g. between 50 and 100 tickets:

```
raffle <- function(n){
p <- 1-((choose(500,2)*choose(3500,n-2))/(choose(4000,n)))
-((choose(500,1)*choose(3500,n-1))/(choose(4000,n)))
-((choose(500,0)*choose(3500,n))/(choose(4000,n)))
return(p)
}
raffle(50:100)
raffle(63:64)
```

The output shows that at least 64 tickets need to be bought to have a 99% guarantee that at least three tickets win. This equates to spending € 96.

(b) We can plot the function as follows:

```
nb <- seq(1:75)
plot(nb,tombola(nb),type='l')
```

Figure B.16 shows the relationship between the number of tickets bought and the probability of having at least three winning tickets.

(c) The solution of (a) shows that it is well worth taking part in the raffle: Marco pays €96 and with a probability of 99 % and he wins at least three prizes which are worth $€142 \cdot 3 = 426$. More generally, the money generated by the raffle is $€1.50 \times 4000 = 6000$, but the prizes are worth $€142 \cdot 500 = 71,000$. One may suspect that the company produces the appliances much more cheaply than they are sold for and is thus so generous.

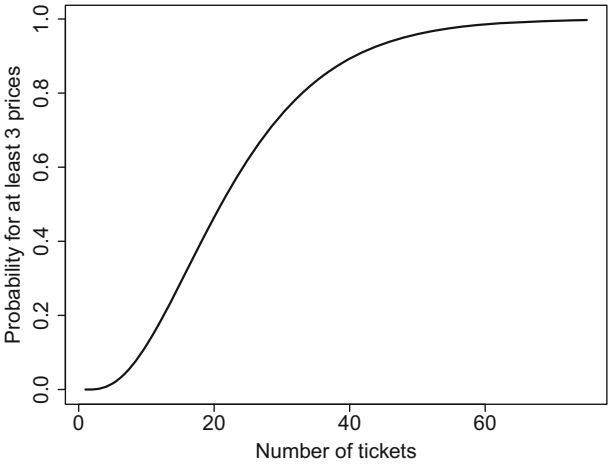


Fig. B.16 Probability to have at least three winning tickets given the number of tickets bought

Solution to Exercise 8.7 The probability of getting a girl is $p = 1 - 0.5122 = 0.4878$.

(a) We are dealing with a geometric distribution here. Since we are interested in $P(X \leq 3)$, we can calculate:

$$
\begin{aligned}
P(X = 1) &= 0.4878 \\
P(X = 2) &= 0.4878(1 - 0.4878) = 0.2498512 \\
P(X = 3) &= 0.4878(1 - 0.4878)^2 = 0.1279738 \\
P(X \leq 3) &= P(X = 1) + P(X = 2) + P(X = 3) = 0.865625.
\end{aligned}
$$

We would have obtained the same result in *R* using:

```
pgeom(2,0.4878)
```

Note that we have to specify "2" rather than "3" for x because *R* takes the number of unsuccessful trials rather the number of trials until success.

(b) Here we deal with a binomial distribution with $k = 2$ and $n = 4$. We can calculate $P(X = 2)$ as follows:

$$
\begin{aligned}
P(X = k) &= \binom{n}{k} p^k (1 - p)^{n-k} \\
&= \binom{4}{2} 0.4878^2 \cdot (1 - 0.4878)^2 = 0.3745536.
\end{aligned}
$$

R would have given us the same result using the `dbinom(2,4,0.4878)` command.

Solution to Exercise 8.8

(a) The random variable Y follows a Poisson distribution, see Theorem 8.2.1 for more details.

(b) The fisherman catches, on average, 3 fish an hour. We can thus assume that the rate λ is 3 and thus $\mathrm{E}(Y) = \lambda = 3$. Similarly, $\mathrm{E}(X) = \frac{1}{\lambda} = \frac{1}{3}$ which means that it takes, on average, 20 min to catch another fish.

(c) Using the PDF of the Poisson distribution, we get:

$$P(Y = 5) = \frac{3^5}{5!}\exp(-3) = 0.1 = 10\,\%$$
$$P(Y < 1) = P(Y = 0) = \frac{3^0}{0!}\exp(-3) \approx 0.0498 \approx 5\,\%.$$

We would have obtained the same results in *R* using the `dpois(5,3)` and `dpois(0,3)` commands.

Solution to Exercise 8.9 The random variable $\mathbf{X}$ = "choice of dessert" follows a multinomial distribution. More precisely, X_1 describes whether chocolate brownies were chosen, X_2 describes whether yoghurt was chosen, X_3 describes whether lemon tart was chosen and $\mathbf{X} = \{X_1, X_2, X_3\}$.

(a) Using the PMF of the multinomial distribution, we get

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1!n_2!\cdots n_k!} \cdot p_1^{n_1} \cdots p_k^{n_k}$$
$$P(X_1 = 2, X_2 = 1, X_3 = 2) = \frac{5!}{2!1!2!} \cdot 0.2^2 \cdot 0.3^1 \cdot 0.5^2$$
$$= 9\,\%.$$

We would have obtained the same results in *R* as follows:

```
dmultinom(c(2,1,2),prob=c(0.2,0.3,0.5))
```

(b) The probability of choosing lemon tart for the first two guests is 1. We thus need to determine the probability that 3 out of the remaining 3 guests order lemon tart:

$$P(X_1 = 0, X_2 = 0, X_3 = 3) = \frac{3!}{0!0!3!} \cdot 0.2^0 \cdot 0.3^0 \cdot 0.5^3$$
$$= 12.5\,\%.$$

Using `dmultinom(c(0,0,3),prob=c(0.2,0.3,0.5))` in *R*, we get the same result.

(c) The expectation vector is

$$\mathrm{E}(\mathbf{X}) = (np_1, \ldots, np_k) = (20 \cdot 0.2, 20 \cdot 3, 20 \cdot 0.5) = (4, 6, 10).$$

This means we expect 4 guests to order brownies, 6 to order yoghurt, and 10 to order lemon tart. The covariance matrix can be determined as follows:

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j. \end{cases}$$

Using $n = 20$, $p_1 = 0.2$, $p_2 = 0.3$ and $p_3 = 0.5$, we obtain the covariance matrix as:

$$\begin{pmatrix} 3.2 & -1.2 & -2 \\ -1.2 & 4.2 & -3 \\ -2 & -3 & 5 \end{pmatrix}$$

Solution to Exercise 8.10

$$\begin{aligned} P(S \geq 1, W \geq 1) &\overset{\text{indep.}}{=} P(S \geq 1) \cdot P(W \geq 1) \\ &= (1 - P(S = 0)) \cdot (1 - P(W = 0)) \\ &= (1 - e^{-3}\frac{3^0}{0!}) \cdot (1 - e^{-4}\frac{4^0}{0!}) \\ &\approx 0.93. \end{aligned}$$

Solution to Exercise 8.11

(a) Random numbers of a normal distribution can be generated using the `rnorm` command. By default $\mu = 0$ and $\sigma = 1$ (see `?rnorm`), so we do not need to specify these parameters. We simply need to set $n = 1000$. The mean of the 1000 realizations can thus be obtained using `mean(rnorm(1000))`. We can, for example, write a `for` loop to repeat this process 1000 times. An empty (`=NA`) vector of size 1000 can be used to store and evaluate the results:

```
set.seed(24121980)
R <- 1000
means <- c(rep(NA,R))
for(i in 1:R){means[i] <- mean(rnorm(1000))}
mean(means)
[1] -0.0007616465
var(means)
[1] 0.0009671311
plot(density(means))
```

We see that the mean of the arithmetic means is close to zero, but not exactly zero. The variance is approximately $\sigma^2/n = 1/1000 = 0.001$, as one would expect

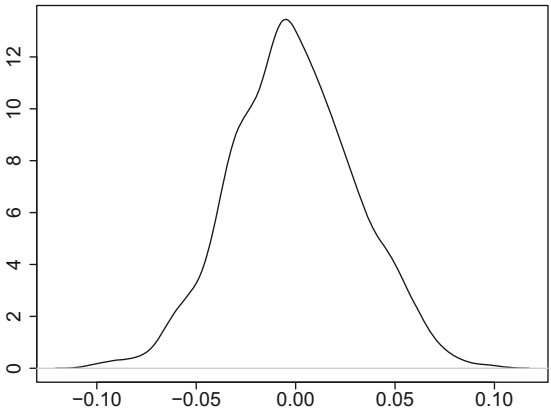


Fig. B.17 Kernel density plot of the distribution simulated in Exercise 8.11a

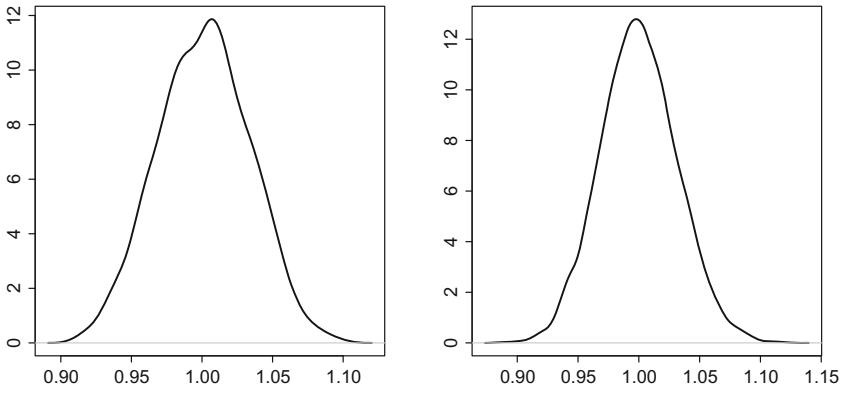
from the Central Limit Theorem. The distribution is symmetric, similar to a normal distribution, see Fig. B.17. It follows that $\bar{X}_n$ is approximately $N(\mu, \frac{\sigma^2}{n})$ distributed, as one could expect from the Theorem of Large Numbers and the Central Limit Theorem. It is important to understand that $\bar{X}$ is not fixed but a random variable which follows a distribution, i.e. the normal distribution.

(b) We can use the same code as above, except we use the exponential instead of the normal distribution:

```
means2 <- c(rep(NA,R))
for(i in 1:R){means2[i] <- mean(rexp(1000))}
mean(means2)
[1] 1.001321
var(means2)
[1] 0.001056113
plot(density(means))
```

The realizations are i.i.d. observations. Once can see that, as in a), $\bar{X}_n$ is approximately $N(\mu, \frac{\sigma^2}{n}) = N(1, 1/1000)$ distributed. It is evident that the X_i do not necessarily need to follow a normal distribution for $\bar{X}$ to follow a normal distribution, see also Fig. B.18a.

(c) Increasing the number of repetitions makes the distribution look closer to a normal distribution, see Fig. B.18b. This visualizes that as n tends to infinity $\bar{X}_n$ gets closer to a $N(\mu, \frac{\sigma^2}{n})$-distribution.



(a) 1,000 repetitions (b) 10,000 repetitions

Fig. B.18 Kernel density plots for Exercises 8.11b and 8.11c

Chapter 9

Solution to Exercise 9.1

(a) The exercise tells us that $X_i \overset{iid}{\sim} Po(\lambda),\ i = 1, 2, \ldots, n$. Let us look at the realizations $x_1, x_2, \ldots, x_n$: under the assumption of independence, which we know is fulfilled because the X_i's are i.i.d., and we can write the likelihood function as the product of the n PMF's of the Poisson distribution:

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum x_i}}{\prod x_i!} e^{-n\lambda}.$$

It is better to work on a log-scale because it is easy to differentiate. The results are identical no matter whether we use the likelihood function or the log-likelihood function because the log transformation is monotone in nature. The log-likelihood function is:

$$\ln L = \sum x_i \ln \lambda - \ln(x_1! \cdots x_n!) - n\lambda.$$

Differentiating with respect to λ yields

$$\frac{\partial \ln L}{\partial \lambda} = \frac{1}{\lambda} \sum x_i - n \overset{!}{=} 0$$

which gives us the ML estimate:

$$\hat{\lambda} = \frac{1}{n} \sum x_i = \bar{x}.$$

We need to confirm that the second derivative is < 0 at $\hat{\lambda} = \bar{x}$; otherwise, the solution would be a minimum rather than a maximum. We get

$$\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{1}{\hat{\lambda}^2} \sum x_i = -\frac{n}{\hat{\lambda}} < 0.$$

It follows that the arithmetic mean $\bar{x} = \hat{\lambda}$ is the maximum likelihood estimator for the parameter λ of a Poisson distribution.

(b) Using the results from (a) we can write the log-likelihood function for $x_1 = 4, x_2 = 3, x_3 = 8, x_4 = 6, x_5 = 6$ as:

$$\ln L = 27 \ln \lambda - \ln(4!\,3!\,8!\,6!\,6!) - 5\lambda.$$

because $\sum x_i = 27$. We can write down this function in *R* as follows:

```
MLP <- function(lambda){
27*log(lambda) - log(factorial(4)*...*factorial(6)) -
5*lambda
}
```

The function can be plotted using the `curve` command:

```
curve(MLP, from=0, to=10)
```

Figure B.19 shows the log-likelihood function. It can be seen that the function reaches its maximum at $\bar{x} = 5.4$.

(c) Using (a) we can write the likelihood function as

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum x_i}}{\prod x_i!} e^{-n\lambda} = \underbrace{\lambda^{\sum x_i} e^{-n\lambda}}_{g(t,\lambda)} \; \underbrace{\frac{1}{\prod x_i!}}_{h(x_1,\ldots,x_n)} .$$

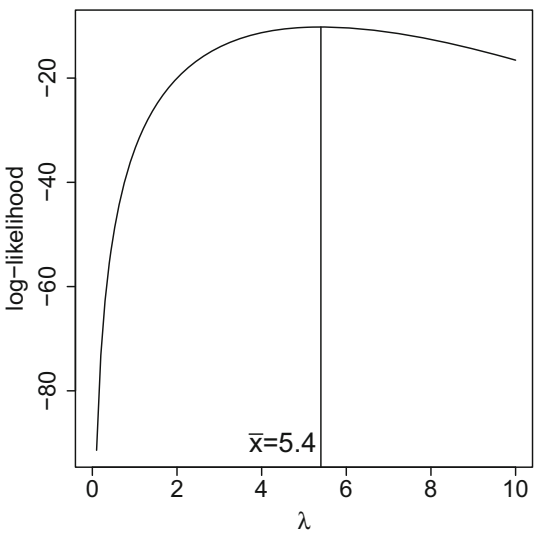


Fig. B.19 Illustration of the log-likelihood function

This means $T = \sum_{i=1}^{n} x_i$ is sufficient for λ. The arithmetic mean, which is the maximum likelihood estimate, is a one-to-one function of T and therefore sufficient too.

Solution to Exercise 9.2

(a) The probability density function of a normal distribution equates to

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

with $\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma^2 > 0$. The likelihood function is therefore

$$L(x_1, x_2, \ldots, x_n|\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

To find the maximum of $L(x_1, x_2, \ldots, x_n|\mu, \sigma^2)$, it is again easier to work with the *log*-likelihood function, which is

$$l = \ln L(x_1, x_2, \ldots, x_n|\mu, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{n}\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

Assuming σ^2 to be 1, differentiating the log-likelihood function with respect to μ, and equating it to zero gives us

$$\frac{\partial l}{\partial \mu} = 2\sum_{i=1}^{n}\left(\frac{x_i-\mu}{1^2}\right) = 0 \quad \Leftrightarrow \quad n\mu = \sum_{i=1}^{n} x_i.$$

The ML estimate is therefore $\hat{\mu} = \bar{x}$.

(b) Looking at the differentiated log-likelihood function in (a) shows us that the ML estimate of μ is always the arithmetic mean, no matter whether σ^2 is 1 or any other number.

(c) Differentiating the log-likelihood function from (a) with respect to σ^2 yields

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0.$$

Using $\hat{\mu} = \bar{x}$ we calculate $\frac{\partial l}{\partial \sigma^2} = 0$ as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Since the parameter $\theta = (\mu, \sigma^2)$ is two-dimensional, it follows that one needs to solve two ML equations, where $\hat{\mu}_{\mathrm{ML}}$ is estimated first, and $\hat{\sigma}^2_{\mathrm{ML}}$ second (as we did above). It follows further that one needs to look at the positive definiteness

of a matrix (the so-called information matrix) when checking that the second-order derivatives of the estimates are positive and therefore the estimates yield a maximum rather then a minimum. However, we omit this lengthy and time-consuming task here.

Solution to Exercise 9.3 The probability density function of $U(0, \theta)$ is

$$f(x) = \frac{1}{\theta} \quad \text{if} \quad 0 < x < \theta \quad \text{and} \quad 0 \quad \text{otherwise.}$$

Note that this equates to the PDF from Definition 8.2.1 for $a = 0$ and $b = \theta$. The likelihood function is therefore

$$L(x_1, x_2, \ldots, x_n|\theta) = \left(\frac{1}{\theta}\right)^n \quad \text{if} \quad 0 < x_i < \theta \quad \text{and} \quad 0 \quad \text{otherwise.}$$

One can see that $L(x_1, x_2, \ldots, x_n|\theta)$ increases as θ decreases. The maximum of the likelihood function is therefore achieved for the smallest valid θ. In particular, θ is minimized when $\theta \geq \max(x_1, x_2, \ldots, x_n) = x_{(n)}$. This follows from the definition of the PDF which requires that $0 < x_i < \theta$ and therefore $\theta > x_i$. Thus, the maximum likelihood estimate of θ is $x_{(n)}$, the greatest observed value in the sample.

Solution to Exercise 9.4

(a) $T_n(X)$ is unbiased, and therefore also asymptotically unbiased, because

$$\mathrm{E}(T_n(X)) = \mathrm{E}(nX_{\min}) \overset{(7.29)}{=} n\frac{1}{n\lambda} = \frac{1}{\lambda} = \mu.$$

Similarly, $V_n(X)$ is unbiased, and therefore also asymptotically unbiased, because

$$\mathrm{E}(V_n(X)) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{(7.29)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}n\frac{1}{\lambda} = \mu.$$

(b) To calculate the MSE we need to determine the bias and the variance of the estimators as we know from Eq. (9.5). It follows from (a) that both estimators are unbiased and hence the bias is 0. For the variances we get:

$$\mathrm{Var}(T_n(X)) = \mathrm{Var}(nX_{\min}) \overset{(7.33)}{=} n^2\mathrm{Var}(X_{\min}) = n^2\frac{1}{n^2\lambda^2} = \mu^2.$$

$$\mathrm{Var}(V_n(X)) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{(7.33)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\frac{1}{\lambda^2} = \frac{1}{n}\mu^2.$$

Since the mean squared error consists of the sum of the variance and squared bias, the MSE for $T_n(X)$ and $V_n(X)$ are μ^2 and $n^{-1}\mu^2$, respectively. One can see that the larger n, the more superior $V_n(X)$ over $T_n(X)$ in terms of the mean squared error. In other words, $V_n(X)$ is more efficient than $T_n(X)$ because its variance is lower for any $n > 1$.

(c) Using the results from (b), we get

$$\lim_{n\to\infty} \text{MSE}(V_n(X)) = \lim_{n\to\infty} \frac{1}{n}\mu^2 = 0.$$

This means the MSE approaches 0 as n tends to infinity. Therefore, $V_n(X)$ is MSE consistent for μ. Since $V_n(X)$ is MSE consistent, it is also weakly consistent.

Solution to Exercise 9.5

(a) The point estimate of μ is $\bar{x}$ which is

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{24}(450 + \cdots + 790) = 667.92.$$

The variance of σ^2 can be estimated unbiasedly using s^2:

$$\begin{aligned}\hat{\sigma}^2 = s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{23}((450 - 667.92)^2 + \cdots + (790 - 667.92)^2) \approx 18,035.\end{aligned}$$

(b) The variance is unknown and needs to be estimated. We thus need the t-distribution to construct the confidence interval. We can determine $t_{23;0.975} \approx 2.07$ using `qt(0.975,23)` or Table C.2 (though the latter is not detailed enough), $\alpha = 0.05$, $\bar{x} = 667.97$ and $\hat{\sigma}^2 = 18,035$. This yields

$$I_l(X) = \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 667.92 - t_{23;0.975} \cdot \frac{\sqrt{18,035}}{\sqrt{24}} \approx 611.17,$$

$$I_u(X) = \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} 667.92 - t_{23;0.975} \cdot \frac{\sqrt{18,035}}{\sqrt{24}} \approx 724.66.$$

The confidence interval for μ is thus [611.17; 724.66].

(c) We can reproduce these results in *R* as follows:

```
eland <- c(450,730,700,600,620,,790)
t.test(eland)$conf.int
```

Solution to Exercise 9.6

- Let us start with the confidence interval for the "Brose Baskets Bamberg". Using $t_{15;0.975} = 2.1314$ (`qt(0.975,15)` or Table C.2) and $\alpha = 0.05$, we can determine the confidence interval as follows:

$$I_l(Ba) = \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 199.06 - t_{15;0.975} \cdot \frac{7.047}{\sqrt{16}} = 195.305,$$
$$I_u(Ba) = \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 199.06 + t_{15;0.975} \cdot \frac{7.047}{\sqrt{16}} = 202.815.$$

  Thus, we get [195.305; 202.815].

- For the "Bayer Giants Leverkusen", we use $t_{13;0.975} = 2.1604$ to get

$$I_l(L) = \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 196 - t_{13;0.975} \cdot \frac{9.782}{\sqrt{14}} = 190.352,$$
$$I_u(L) = \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 196 + t_{13;0.975} \cdot \frac{9.782}{\sqrt{14}} = 201.648.$$

  This leads to a confidence interval of [190.352; 201.648].

- For "Werder Bremen", we need to use the quantile $t_{22,0.975} = 2.0739$ which yields a confidence interval of

$$I_l(Br) = \bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 187.52 - t_{22;0.975} \cdot \frac{5.239}{\sqrt{23}} = 185.255,$$
$$I_u(Br) = \bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}} = 187.25 + t_{22;0.975} \cdot \frac{5.239}{\sqrt{23}} = 189.786.$$

  The interval is therefore [185.255; 189.786].

- The mean heights of the basketball teams are obviously larger than the mean height of the football team. The two confidence intervals of the basketball teams overlap, whereas the intervals of the football team with the two basketball teams do not overlap. It is evident that this indicates that the height of football players is substantially less than the height of basketball players. In Chap. 10, we will learn that confidence intervals can be used to test hypotheses about mean differences.

Solution to Exercise 9.7

(a) Using $n = 98$, we calculate an unbiased point estimate for p using $\bar{x} = \hat{p}$. Note that $x_i = 1$ if the wife has to wash the dishes.

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{98} \cdot 59 = \frac{59}{98} \approx 0.602.$$

(b) Since $n\hat{p}(1-\hat{p}) = 98 \cdot 0.602 \cdot 0.398 = 23.48 > 9$ and p is sufficiently large, we can use the normal approximation to calculate a confidence interval for p. Using $z_{1-\alpha/2} = z_{0.975} = 1.96$, we obtain

$$I_l(X) = 0.602 - 1.96\sqrt{\frac{0.602 \cdot 0.398}{98}} = 0.505,$$
$$I_u(X) = 0.602 + 1.96\sqrt{\frac{0.602 \cdot 0.398}{98}} = 0.699.$$

This yields a confidence interval of $[0.505, 0.699]$. Note that the confidence interval does not contain $p = 0.5$ which is the probability that would be expected if the coin was fair. This is a clear sign that the coin might be unfair.

(c) If the coin is fair, we can use $\hat{p} = 0.5$ as our prior judgement. We would then need

$$n \geq \left[\frac{z_{1-\alpha/2}}{\Delta}\right]^2 \hat{p}(1-\hat{p})$$
$$\geq \left[\frac{1.96}{0.005}\right]^2 0.5^2 = 38,416$$

dinners to get the desired precision—which is not possible as this would constitute a time period of more than 100 years. This shows that the expectation of such a high precision is unrealistic and needs to be modified. However, as the results of (b) show, the coin may not be fair. If the coin is indeed unfair, we may have, for example, $p = 0.6$ and $1 - p = 0.4$ which gives a smaller sample size. We can thus interpret the sample size as a conservative estimate of the highest number of dinners needed.

Solution to Exercise 9.8 If a student fails then $x_i = 1$. We know that $\sum x_i = 11$ and $n = 104$.

(a) Using $\bar{x} = p$ as point estimate, we get $\hat{p} = \frac{11}{104} \approx 0.106 = 10.6\ \%$. Using $z_{1-\alpha/2} = z_{0.975} = 1.96$, we can calculate the confidence interval as

$$0.106 \pm 1.96 \cdot \sqrt{\frac{0.106 \cdot 0.894}{104}} = [0.047; 0.165]\,.$$

Using *R* we get:

```
binom.test(11,104)$conf.int
[1] 0.05399514 0.18137316
```

This result is different because the above command does not use the normal approximation. In addition, p is rather small which means that care must be exercised when using the results of the confidence interval with normal approximation in this context.

(b) The point estimate of 10.6 % is substantially higher than 3.2 %. The lower bound confidence interval is still larger than the proportion of failures at county level. This indicates that the school is doing worse than most other schools in the county.

Solution to Exercise 9.9

(a) Whether the ith household has switched on the TV and watches "Germany's next top model" (GNTM) relates to a random variable X_i with

$$X_i = 1 : \text{if TV switched on and household watching GNTM}$$
$$X_i = 0 : \text{if TV switched off or household watches another show.}$$

It follows that $X = \sum_{i=1}^{2500} X_i$ is the random variable describing the number of TVs, out of 2500 TVs, which are switched on and show GNTM. Since the X_i's can be assumed to be i.i.d., we can say that X follows a binomial distribution, i.e. $X \sim B(2500; p)$ with p unknown. The length of the confidence interval for p,

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

is

$$L = 2z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Unfortunately, $\hat{p}(1-\hat{p})$ is unknown but $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$ because the maximum value for $\hat{p}(1-\hat{p})$ is 0.25 if $\hat{p} = 0.5$. Hence

$$L \leq 2z_{1-\alpha/2}\sqrt{\frac{\frac{1}{4}}{n}} = \frac{1.96}{\sqrt{2500}} = 0.0392.$$

This means the precision is half the length, i.e. $\pm 0.0196 = \pm 1.96$ %, in the worst case.

(b) There is the danger of selection bias. A total of 500 households refused to take part in the study. It may be that their preferences regarding TV shows are different from the other 2500 households. For example, it may well be possible that those watching almost no TV refuse to be included; or that those watching TV shows which are considered embarrassing by society are refusing as well. In general, missing data may cause point estimates to be biased, depending on the underlying mechanism causing the absence.

Solution to Exercise 9.10

(a) Using $z_{1-\alpha/2} = 1.96$ and $\hat{\sigma} = 0.233$, we obtain an optimal sample size of at least

$$n_{opt} \geq \left[2\frac{z_{1-\alpha/2}\sigma_0}{\Delta}\right]^2 = \left[2 \cdot \frac{1.96 \cdot 0.233}{0.2}\right]^2 = 20.85.$$

To calculate a confidence interval with a width of not more than 0.2 s, the results of at least 21 athletes are needed.

(b) The sample size is 30. Thus, the confidence interval width should be smaller than 0.2 s. This is indeed true as the calculations show:

$$[10.93 \pm \underbrace{t_{0.975;29}}_{2.045} \cdot \frac{0.233}{\sqrt{30}}] = [10.84; 11.02].$$

The width is only 0.18 s.

(c) If we calculate a 80 % confidence interval, then the lower confidence limit corresponds to the running time which is achieved by only 10 % of the athletes (of the population). Using $t_{0.9;29} \approx 1.31$ we get

$$[10.93 \pm 1.31 \cdot \frac{0.233}{\sqrt{30}}] = [10.87; 10.99].$$

The athlete's best time is thus below the lower confidence limit. He is among the top 10 % of all athletes, using the results of the confidence interval.

Solution to Exercise 9.11

(a) The odds ratio is

$$OR = \frac{163 \cdot 477}{475 \cdot 151} \approx 1.08.$$

This means the chances that a pizza arrives in time are about 1.08 times higher for Laura compared with Melissa. To calculate the 95 % confidence interval, we need $\hat{\theta}_0 = \ln(1.08) \approx 0.077$, $z_{1-\alpha/2} \approx 1.96$, and

$$\hat{\sigma}_{\hat{\theta}_0} = \left(\frac{1}{163} + \frac{1}{475} + \frac{1}{151} + \frac{1}{477}\right)^{\frac{1}{2}} = 0.13.$$

The interval for the log odds ratio is

$$[\ln(1.08) \pm 1.96 \cdot 0.13] \approx [-0.18; 0.33]\,.$$

Exponentiating the interval gives us the 95 % confidence interval for the odds ratio which is [0.84; 1.39]. This indicates that the odds of Laura's pizzas arriving earlier than Melissa's are not much different from one. While the point estimate tells us that Laura's pizzas are delivered 1.08 times faster, the confidence interval tells us that there is uncertainty around this estimate in the sense that it could also be smaller than 1 and Melissa may not necessarily work more slowly than Laura.

(b) We can reproduce the results in *R* by attaching the pizza data, creating a categorical delivery time variable (using `cut`) and then applying the `oddsratio` command from the library `epitools` onto the contingency table:

```
attach(pizza)
timecat <- cut(time, breaks=c(-1,30,150))
library(epitools)
oddsratio(table(timecat,operator), method='wald')
```

Chapter 10

Solution to Exercise 10.1 A type I error is defined as the probability of rejecting H_0 if H_0 is true. This error occurs if A thinks that B does confess, but B does not. In this scenario, A confesses, goes free, and B serves a three-year sentence. A type II error is defined as the probability of accepting H_0, despite the fact that H_0 is wrong. In this case, B does confess, but A does not. In this scenario, B goes free and A serves a three-year sentence. A type II error is therefore worse for A. With a statistical test, we always control the type I error, but not the type II error.

Solution to Exercise 10.2

(a) The hypotheses are

$$H_0 : \mu = 100 \quad \text{versus} \quad H_1 : \mu \neq 100.$$

(b) It is a one-sample problem for μ: thus, for known variance, the Gauss test can be used; the t-test otherwise, see also Appendix D. Since, in this exercise, σ is assumed to be known we can use the Gauss test; i.e. we can compare the test statistic with the normal distribution (as opposed to the t-distribution when using the t-test). The sample size is small: we must therefore assume a normal distribution for the data.

(c) To calculate the realized test statistic, we need the arithmetic mean $\bar{x} = 98.08$. Then we obtain

$$t(x) = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{n} = \frac{98.08 - 100}{2} \cdot \sqrt{15} = \frac{-1.92}{2} \cdot \sqrt{15} = -3.72.$$

We reject H_0 if $|t(x)| > z_{1-\frac{\alpha}{2}} = 1.96$. Since $|t(x)| = 3.72 > 1.96$, the null hypothesis can be rejected. The test is significant. There is enough evidence to suggest that the observed weights do not come from a normal distribution with mean 100 g.

(d) To show that $\mu < 100$, we need to conduct a one-sided test. Most importantly, the research hypothesis of interest needs to be stated as the *alternative* hypothesis:

$$H_0 : \mu \geq 100 \qquad \text{versus} \qquad H_1 : \mu < 100.$$

(e) The test statistic is the same as in (b): $t(x) = -3.72$. However, the critical region changes. H_0 gets rejected if $t(x) < -z_{1-\alpha} = -1.64$. Again, the null hypothesis is rejected. The producer was right in hypothesizing that the average weight of his chocolate bars was lower than 100 g.

Solution to Exercise 10.3

(a) We calculate

	No auction	Auction
$\bar{x}$	16.949	10.995
s^2	2.948	2.461
s	1.717	1.569
v	0.101	0.143

Note that we use the unbiased estimates for the variance and the standard deviation as explained in Chap. 9; i.e. we use $1/(n-1)$ rather than $1/n$. It is evident that the mean price of the auctions (μ_a) is lower than the mean non-auction price (μ_{na}), and also lower than the price from the online book store. There is, however, a higher variability in relation to the mean for the auction prices. One may speculate that the best offers are found in the auctions, but that there are no major differences between the online store and the internet book store, i.e.

- $\mu_{na} \neq$ €16.95,
- $\mu_a <$ €16.95,
- $\mu_{na} > \mu_a$.

(b) We can use the t-test to test the hypotheses

$$H_0 : \mu_{na} = 16.95 \qquad \text{versus} \qquad H_1 : \mu_{na} \neq 16.95.$$

The test statistic is

$$t(x) = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{n} = \frac{16.949 - 16.95}{1.717} \cdot \sqrt{14} = -0.002.$$

Using the decision rules from Table 10.2, we conclude that H_0 gets rejected if $|t(x)| > t_{13,0.975} = 2.16$. We can calculate $t_{13,0.975}$ either by using Table C.2 or by using R (`qt(0.975,13)`). Since $|t(x)| = 0.002 \not> 2.16$, we keep the null hypothesis. There is not enough evidence to suggest that the prices of the online store differ from €16.95 (which is the price from the internet book store).

(c) Using (9.6) we can calculate the upper and lower limits of the confidence interval as,

$$\bar{x} + t_{n-1;1-\alpha/2} \cdot \frac{s_X}{\sqrt{n}} = 16.949 + \underbrace{t_{13;0.975}}_{=2.16} \cdot \frac{1.717}{\sqrt{14}} = 17.94$$

$$\bar{x} - t_{n-1;1-\alpha/2} \cdot \frac{s_X}{\sqrt{n}} = 16.949 - \underbrace{t_{13;0.975}}_{=2.16} \cdot \frac{1.717}{\sqrt{14}} = 15.96$$

respectively. The confidence interval *does* contain $\mu_0 = 16.95$; hence, the null hypothesis $\mu = \mu_0 = 16.95$ cannot be rejected. This is the same conclusion as obtained from the one-sample t-test from above.

(d) We test the following hypotheses:

$$H_0 : \mu_a \geq 16.95 \qquad \text{versus} \qquad H_1 : \mu_a < 16.95.$$

The realized test statistic is

$$t(x) = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{n} = \frac{10.995 - 16.95}{1.569} \cdot \sqrt{14} = -14.201.$$

Table 10.2 tells us that H_0 gets rejected if $t(x) < -t_{13,0.95}$. Since $-14.201 < -2.16$ we reject H_0. The test confirms our hypothesis that the mean auction prices are lower than €16.95, i.e. the price from the book store.

(e) We need to conduct two tests: the two-sample t-test for the assumption of equal variances and the Welch test for the assumption of different variances.

(i) *Two-sample t-test*:

$$H_0 : \mu_{na} \leq \mu_a \qquad \text{versus} \qquad H_1 : \mu_{na} > \mu_a.$$

To calculate the test statistic (10.4), we need to determine the pooled sample variance:

$$\begin{aligned} s^2 &= \frac{(n_{na} - 1)s_{na}^2 + (n_a - 1)s_a^2}{n_{na} + n_a - 2} \\ &= \frac{(14 - 1)2.948 + (14 - 1)2.461}{14 + 14 - 2} \approx 2.705 \end{aligned}$$

The test statistic is

$$\begin{aligned} t(x) &= \frac{\bar{x}_{na} - \bar{x}_a}{s} \cdot \sqrt{\frac{n_{na} \cdot n_a}{n_{na} + n_a}} = \frac{16.949 - 10.995}{\sqrt{2.705}} \cdot \sqrt{\frac{14 \cdot 14}{14 + 14}} \\ &= \frac{5.954}{1.645} \cdot \sqrt{\frac{196}{28}} = 3.621 \cdot \sqrt{7} = 9.578. \end{aligned}$$

We reject H_0 if $t(x) > t_{n_{na}+n_a-2,0.95} = t_{26,0.95} = 1.71$. Table C.2 does not list the quantile; thus, one uses R (`qt(0.95,26)`) to determine the quantile. Since $9.578 > 1.71$, we reject the null hypothesis. The test is significant. There is enough evidence to conclude that the mean auction prices are lower than the mean non-auction prices.

(ii) *Welch test*: For the Welch test, we calculate the test statistic, using (10.6) as:

$$t(x) = \frac{|\bar{x}_{na} - \bar{x}_a|}{\sqrt{\frac{s_{na}^2}{n_{na}} + \frac{s_a^2}{n_a}}} = \frac{|16.949 - 10.995|}{\sqrt{\frac{2.948}{14} + \frac{2.461}{14}}} = 9.578.$$

To calculate the critical value, we need the degrees of freedom:

$$\begin{aligned} v &= \left(\frac{s_{na}^2}{n_{na}} + \frac{s_a^2}{n_a}\right)^2 \Big/ \left(\frac{\left(s_{na}^2/n_{na}\right)^2}{n_{na}-1} + \frac{\left(s_a^2/n_a\right)^2}{n_a - 1}\right) \\ &= \left(\frac{2.948}{14} + \frac{2.461}{14}\right)^2 \Big/ \left(\frac{(2.948/14)^2}{13} + \frac{(2.461/14)^2}{13}\right) \\ &\approx 25.79 \end{aligned}$$

We reject H_0 if $t(x) > t_{v,0.95}$. Using $t_{v,0.95} = 1.706$ (in *R* obtained as `qt(0.95,25.79)`) and $t(x) = 9.578$, we reject the null hypothesis. The conclusions are identical to using the two-sample t-test (in this example). Interestingly, the two test statistics are similar which indicates that the assumption of equal variances in case (i) was not unreasonable.

(f) The F-test relies on the assumption of the normal distribution and tests the hypotheses:

$$H_0 : \sigma_{na}^2 = \sigma_a^2 \qquad \text{versus} \qquad H_1 : \sigma_{na}^2 \neq \sigma_a^2.$$

We can calculate the test statistic as described in Appendix C:

$$t(x) = \frac{s_{na}^2}{s_a^2} = \frac{2.949}{2.461} = 1.198.$$

The F-distribution is not symmetric around 0; thus, we need to calculate two critical values: $f_{n_1-1;n_2-1;1-\alpha/2}$ and $f_{n_1-1;n_2-1;\alpha/2}$. Using *R* we get $f_{13,13,0.975} = 3.115$ (`qf(0.975,13,13)`) and $f_{13,13,0.025} = \frac{1}{3.115} = 0.321$ (`qf(0.025,13,13)`). H_0 is rejected if $t(x) > f_{n_1-1;n_2-1;1-\alpha/2}$ or $t(x) < f_{n_1-1;n_2-1;\alpha/2}$. Since $0.321 < 1.198 < 3.115$ we do not reject H_0. This means there is strong evidence that the two variances are equal. This is also reassuring with respect to using the two-sample t-test above (which assumes equal variances). However, note that testing the equality of variances with the F-test and then using the two-sample t-test or Welch test based on the outcome of the F-test is not ideal and not necessarily correct. In practice, it is best to simply use the Welch test (rather than the t-test), which is also the default option in the *R* function `t.test`.

(g) We need the following table to calculate the rank sums:

Value	9.3	9.52	9.54	10.01	10.5	10.5	10.55	10.59	11.02	11.03
Sample	a	a	a	a	a	a	a	a	a	a
Rank	1	2	3	4	5	6	7	8	9	10
Value	11.89	11.99	12	13.79	15.49	15.9	15.9	15.9	15.9	15.9
Sample	a	a	a	na	a	na	na	na	na	na
Rank	11	12	13	14	15	16	17	18	19	20
Value	15.99	16.98	16.98	17.72	18.19	18.19	19.97	19.97		
Sample	na	na	na	na	na	na	na	na		
Rank	21	22	23	24	25	26	27	28		

We can calculate the rank sums as $R_{na+} = 13 + 15 + \cdots + 28 = 300$ and $R_{a+} = 1 + 2 + \cdots + 13 + 15 = 106$, respectively. Thus

$$U_1 = 14^2 + \frac{14 \cdot 15}{2} - 106 = 195; \quad U_2 = 14^2 + \frac{14 \cdot 15}{2} - 300 = 1.$$

With $U = \min(195, 1) = 1$ we can calculate the test statistic, which is approximately normally distributed, as:

$$t(x, y) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{1 - \frac{14^2}{2}}{\sqrt{\frac{14 \cdot 14 \cdot 29}{12}}} \approx -4.46$$

Since $|t(x, y)| = 4.46 > z_{1-\alpha/2} = 1.96$, the null hypothesis can be rejected. We conclude that the locations of the two distributions are shifted. Obviously the prices of the auction are smaller in general, and so is the median.

(h) We can type in the data and evaluate the summary statistics using the `mean`, `var`, and `sd` commands:

```
na <- c(18.19,16.98,19.97,...,17.72)
a <- c(10.5,12.0, 9.54,..., 11.02)
mean(na)
mean(a)
var(na)
...
```

The `t.test` command can be used to answer questions (b)–(e). For (b) and (c) we use

```
t.test(na,mu=16.95)
```

```
        One-sample t-test

data:  na
t = -0.0031129, df = 13, p-value = 0.9976
alternative hypothesis: true mean is not equal to 16.95
95 percent confidence interval:
 15.95714 17.94001
sample estimates:
mean of x
 16.94857
```

The test decision can be made by means of either the p-value ($= 0.9976 > 0.05$) or the confidence interval ([15.95; 17.94], which covers 16.95). To answer (d) and (e) we need to make use of the option `alternative` which specifies the alternative hypothesis:

```
t.test(a,mu=16.95,alternative='less')
t.test(na,a, alternative='greater')
```

Note that the two-sample test provides a confidence interval for the difference of the means. Questions (f) and (g) can be easily solved by using the `var.test` and `wilcox.test` commands:

```
var.test(na,a)
wilcox.test(na,a)
```

```
        F-test to compare two variances

data:  na and a
F = 1.198, num df = 13, denom df = 13, p-value = 0.7496
alternative hypothesis: true ratio of variances not equal to 1
95 percent confidence interval:
 0.3845785 3.7317371
sample estimates:
ratio of variances
          1.197976
```

```
        Wilcoxon rank sum test with continuity correction

data:  na and a
W = 195, p-value = 8.644e-06
alternative hypothesis: true location shift is not equal to 0
```

The test decision is best made by using the p-value.

Solution to Exercise 10.4 Since the data before and after the diet is dependent (weight measured on the same subjects), we need to use the paired t-test. To calculate the test statistic, we need to first determine the weight differences:

Person i	1	2	3	4	5	6	7	8	9	10
Before diet	80	95	70	82	71	70	120	105	111	90
After diet	78	94	69	83	65	69	118	103	112	88
Differences d	2	1	1	−1	6	1	2	2	−1	2

Using $\bar{d} = 1.5$ and

$$s_d^2 = \frac{1}{10-1} \cdot (0.5^2 + 0.5^2 + 0.5^2 + 2.5^2 + 4.5^2 + \cdots + 0.5^2) = 3.83,$$

we calculate the test statistic as

$$t(d) = \frac{\bar{d}}{s_d}\sqrt{n} = \frac{1.5}{\sqrt{3.83}}\sqrt{10} = 2.42.$$

The null hypothesis is rejected because $|t(d)| = 2.42 > t_{9;0.975} = 2.26$. This means the mean weight before and after the diet is different. We would have obtained the same results by calculating the confidence interval for the differences:

$$\left[\bar{d} \pm t_{n-1;1-\alpha/2}\frac{s_d}{\sqrt{n}}\right] \Leftrightarrow \left[1.5 \pm 2.26 \cdot \frac{\sqrt{3.83}}{\sqrt{10}}\right] = [0.1; 2.9].$$

Since the confidence interval does not overlap with zero, we reject the null hypothesis; there is enough evidence that the mean difference is different (i.e. greater) from zero. While the test is significant and suggests a weight difference, it is worth noting that the mean benefit from the diet is only 1.5 kg. Whether this is a relevant reduction in weight has to be decided by the ten people experimenting with the diet.

Solution to Exercise 10.5

(a) The production is no longer profitable if the probability of finding a deficient shirt is greater than 10 %. This equates to the hypotheses:

$$H_0 : p \leq 0.1 \qquad \text{versus} \qquad H_0 : p > 0.1.$$

The sample proportion of deficient shirts is $\hat{p} = \frac{35}{230} = \frac{7}{46}$. Since $np(1-p) = 230 \cdot \frac{1}{10} \cdot \frac{9}{10} > 9$ we can use the test statistic

$$t(x) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{\frac{7}{46} - \frac{1}{10}}{\sqrt{\frac{1}{10} \cdot \frac{9}{10}}} \cdot \sqrt{230}$$
$$= \frac{6}{115} \cdot \frac{10}{3} \cdot \sqrt{230} = \frac{4}{23} \cdot \sqrt{230} = 2.638.$$

The null hypothesis gets rejected because $t(x) = 2.638 > z_{0.95} = 1.64$. It seems the production is no longer profitable.

(b) The test statistic is $t(x) = \sum x_i = 30$. The critical region can be determined by calculating

$$P_{H_0}(Y \leq c_l) \leq 0.025 \qquad \text{and} \qquad P_{H_0}(Y \geq c_r) \leq 0.975.$$

Using R we get

```
qbinom(p=0.975,prob=0.1,size=230)
[1] 32
qbinom(p=0.025,prob=0.1,size=230)
[1] 14
```

The test statistic ($t(x) = 30$) does not fall outside the critical region ([14; 32]); therefore, the null hypothesis is *not* rejected. The same result is obtained by using the binomial test in R: `binom.test(c(30,200),p=0.1)`. This yields a p-value of 0.11239 and a confidence interval covering 10 % ([0.09; 0.18]). Interestingly, the approximate binomial test rejects the null hypothesis, whereas the exact test keeps it. Since the latter test is more precise, it is recommended to follow its outcome.

(c) The research hypothesis is that the new machine produces fewer deficient shirts:

$$H_0 : p_{\text{new}} \geq p_{\text{old}} \qquad \text{versus} \qquad H_1 : p_{\text{new}} < p_{\text{old}}.$$

To calculate the test statistic, we need the following:

$$\hat{d} = \frac{x_{\text{new}}}{n_{\text{new}}} - \frac{x_{\text{old}}}{n_{\text{old}}} = \frac{7}{115} - \frac{7}{46} = -\frac{21}{230}$$
$$\hat{p} = \frac{x_{\text{new}} + x_{\text{old}}}{n_{\text{new}} + n_{\text{old}}} = \frac{7 + 35}{230 + 115} = \frac{42}{345} = \frac{14}{115}.$$

This yields:

$$t(x_{\text{new}}, x_{\text{old}}) = \frac{\hat{d}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_{\text{new}}} + \frac{1}{n_{\text{old}}})}} = \frac{-\frac{21}{230}}{\sqrt{\frac{14}{115} \cdot \frac{101}{115}(\frac{1}{115} + \frac{1}{230})}}$$
$$= \frac{-\frac{21}{230}}{\sqrt{0.1069 \cdot 0.013}} = -\frac{0.0913}{0.0373} = -2.448.$$

The null hypothesis is rejected because $t(x_{\text{new}}, x_{\text{old}}) = -2.448 < z_{0.05} = -z_{0.95} = -1.64$. This means we accept the alternative hypothesis that the new machine is better than the old one.

(d) The data can be summarized in the following table:

	Machine 1	Machine 2
Deficient	30	7
Fine	200	112

To test the hypotheses established in (c), we apply the `fisher.test` command onto the contingency table:

```
fisher.test(matrix(c(30,200,7,112),ncol=2))
```

This yields a p-value of 0.0438 suggesting, as the test in (c), that the null hypothesis should be rejected. Note that the confidence interval for the odds ratio, also reported by R, does not necessarily yield the same conclusion as the test of Fisher.

Solution to Exercise 10.6

(a) To compare the two means, we should use the Welch test (since the variances cannot be assumed to be equal). The test statistic is

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}}} = \frac{|103 - 101.8|}{\sqrt{\frac{12599.56}{10} + \frac{62.84}{10}}} \approx 0.0337.$$

The alternative hypothesis is $H_1 : \mu_2 > \mu_1$; i.e. we deal with a one-sided hypothesis. The null hypothesis is rejected if $t(x, y) > t_{v;1-\alpha}$. Calculating

$$\begin{aligned} v &= \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Big/ \left(\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}\right) \\ &= \left(\frac{62.844}{10} + \frac{12599}{10}\right)^2 \Big/ \left(\frac{(62.844/10)^2}{9} + \frac{(12599/10)^2}{9}\right) \approx 9.09 \end{aligned}$$

yields $t_{9.09;0.95} = 1.831$ (using `qt(0.95,9.09)` in R; or looking at Table C.2 for 9 degrees of freedom). Therefore, $t(x, y) < t_{9.09;0.95}$ and the null hypothesis is not rejected. This means player 2 could not prove that he scores higher on average.

(b) We have to first rank the merged data:

Value	6	29	40	47	47	64	87	88	91	98
Sample	2	2	2	2	2	2	2	1	1	2
Rank	1	2	3	4	5	6	7	8	9	10
Value	99	99	101	104	105	108	111	112	261	351
Sample	1	1	1	1	1	1	1	1	2	2
Rank	11	12	13	14	15	16	17	18	19	20

This gives us $R_{1+} = 8 + 9 + \cdots + 18 = 133$, $R_{2+} = 1 + 2 + \cdots + 20 = 77$, $U_1 = 10^2 + (10 \cdot 11)/2 - 133 = 22$, $U_2 = 10^2 + (10 \cdot 11)/2 - 77 = 78$, and therefore $U = 22$. The test statistic can thus be calculated as

$$t(x, y) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{22 - \frac{10^2}{2}}{\sqrt{\frac{10 \cdot 10 \cdot 21}{12}}} \approx= -2.12.$$

Since $|t(x, y)| = 2.12 > z_{1-\alpha} = 1.64$, the null hypothesis is rejected. The test supports the alternative hypothesis of higher points for player 1. The U-test has the advantage of not being focused on the expectation μ. The two samples are clearly different: the second player scores with much more variability and his distribution of scores is clearly not symmetric and normally distributed. Since the sample is small, and the assumption of a normal distribution is likely not met, it makes sense to *not* use the t-test. Moreover, because the distribution is skewed the mean may not be a sensible measure of comparison. The two tests yield different conclusions in this example which shows that one needs to be careful in framing the right hypotheses. A drawback of the U-test is that it uses only ranks and not the raw data: it thus uses less information than the t-test which would be preferred when comparing means of a reasonably sized sample.

Solution to Exercise 10.7 Otto speculates that the probability of finding a bear of colour i is 0.2, i.e. $p_{\text{white}} = 0.2$, $p_{\text{red}} = 0.2$, $p_{\text{orange}} = 0.2$, $p_{\text{yellow}} = 0.2$, and $p_{\text{green}} = 0.2$. This hypothesis can be tested by using the χ^2 goodness-of-fit test. The test statistic is

$$t(x) = \chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} = \frac{1}{250}(222 - 250)^2 + (279 - 250)^2 + (251 - 250)^2 + (232 - 250)^2 + (266 - 250)^2 = 8.824.$$

The null hypothesis cannot be rejected because $t(x) = 8.824 \not> c_{4-1-0;0.95} = 9.49$. While the small number of white bears might be disappointing, the test suggests that there is not enough evidence to reject the hypothesis of equal probabilities.

Solution to Exercise 10.8

(a) To answer this question, we need to conduct the χ^2-independence test. The test statistic $t(x, y)$ is identical to Pearson's χ^2 statistic, introduced in Chap. 4. In Exercise 4.4, we have calculated this statistic already, $\chi^2 \approx 182$, see p. 350 for the details. The null hypothesis of independence is rejected if $t(x, y) = \chi^2 > c_{(I-1)(J-1);1-\alpha}$. With $I = 2$ (number of rows), and $J = 4$ (number of columns) we get $c_{3;0.95} = 7.81$ using Table C.3 (or `qchisq(0.95,3)` in *R*). Since $182 > 7.81$, we reject the null hypothesis of independence.

(b) The output refers to a χ^2 test of homogeneity: the null hypothesis is that the proportion of passengers rescued is identical for the different travel classes. This hypothesis is rejected because p is smaller than $\alpha = 0.05$. It is evident that the proportions of rescued passengers in the first two classes (60 %, 43.9 %) are much higher than in the other classes (25 %, 23.8 %). One can see that the test statistic (182.06) is identical to (a). This is not surprising: the χ^2-independence test and the χ^2 test of homogeneity are technically identical, but the null hypotheses differ. In (a), we showed that "rescue status" and "travel class" are independent; in (b), we have seen that the conditional distributions of rescue status given travel class differ by travel class, i.e. that the proportions of those rescued differ by the categories 1. class/2. class/3. class/staff.

(c) The summarized table is as follows:

	1. Class/2. Class	3. Class/Staff	Total
Rescued	327	391	718
Not rescued	295	1215	1510
Total	622	1606	2228

Using (4.7) we get

$$t(x, y) = \chi^2 = \frac{2228(327 \cdot 1215 - 295 \cdot 391)^2}{718 \cdot 1510 \cdot 622 \cdot 1606} = 163.55.$$

The χ^2-independence test and the χ^2 test of homogeneity are technically identical: H_0 is rejected if $t(x, y) > c_{(I-1)(J-1);1-\alpha}$. Since $163.55 > c_{1;0.95} = 3.84$ the null hypothesis is rejected. As highlighted in (b) this has two interpretations: (i) "rescue status" and "travel class" are independent (independence test) and (ii) the conditional distributions of rescue status given travel class differ by travel class (homogeneity test). The second interpretation implies that the probability of being rescued differs by travel class. The null hypothesis of the same probabilities of being rescued is also rejected using the test of Fisher. Summarizing the data in a matrix and applying the `fisher.test` command yields a p-value smaller than $\alpha = 0.05$.

```
fisher.test(matrix(c(327,295,391,1215),ncol=2,nrow=2))
```

Solution to Exercise 10.9

(a) The hypotheses are:

$$H_0 : \mu_X = \mu_{Y1} \qquad \text{versus} \qquad H_1 : \mu_X \neq \mu_{Y1}.$$

The pooled variance,

$$s^2 = \frac{19 \cdot 2.94 + 19 \cdot 2.46}{39} = \frac{102.6}{39} = 2.631,$$

is needed to calculate the test statistic:

$$t(x, y) = \frac{4.97 - 4.55}{1.622} \cdot \sqrt{\frac{400}{40}} = \frac{0.42}{1.622} \cdot \sqrt{10} = 0.8188.$$

H_0 is rejected if $|t(x, y)| > t_{38,0.975} \approx 2.02$. Since $|t(x, y)| = 0.8188$ we do not reject the null hypothesis.

(b) The hypotheses are:

$$H_0 : \mu_X = \mu_{Y2} \qquad \text{versus} \qquad H_1 : \mu_X \neq \mu_{Y2}.$$

The pooled variance,

$$s^2 = \frac{19 \cdot 2.94 + 19 \cdot 3.44}{39} = 3.108,$$

is needed to calculate the test statistic:

$$t(x, y) = \frac{4.97 - 3.27}{1.763} \cdot \sqrt{10} = 3.049.$$

H_0 is rejected because $|t(x, y)| = 3.049 > t_{38,0.975} \approx 2.02$.

(c) In both (a) and (b), there exists a true difference in the mean. However, only in (b) is the t-test able to detect the difference. This highlights that smaller differences can only be detected if the sample size is sufficiently large. However, if the sample size is very large, it may well be that the test detects a difference where there is no difference.

Solution to Exercise 10.10

(a) After reading in and attaching the data, we can simply use the `t.test` command to compare the expenditure of the two samples:

```
theatre <- read.csv('theatre.csv')
attach(theatre)
t.test(Culture[Sex==1],Culture[Sex==0])
```

```
        Welch Two-Sample t-test

data:  Culture[Sex == 1] and Culture[Sex == 0]
t = -1.3018, df = 667.43, p-value = 0.1934
alternative hypothesis: true difference not equal to 0
95 percent confidence interval:
 -12.841554   2.602222
sample estimates:
mean of x mean of y
 217.5923  222.7120
```

We see that the null hypothesis is *not* rejected because $p = 0.1934 > \alpha$ (also, the confidence interval overlaps with "0").

(b) A two-sample t-test and a U-test yield the same conclusion. The p-values, obtained with

```
wilcox.test(Culture[Sex==1],Culture[Sex==0])
t.test(Culture[Sex==1],Culture[Sex==0],var.equal=TRUE)
```

are 0.1946 and 0.145, respectively. Interestingly, the test statistic of the two-sample t-test is almost identical to the one from the Welch test in (a) (-1.2983)—this indicates that the assumption of equal variances may not be unreasonable.

(c) We can use the usual `t.test` command together with the option `alternative = 'greater'` to get the solution.

```
t.test(Theatre[Sex==1],Theatre[Sex==0],
alternative='greater')
```

Here, p is much smaller than 0.001; hence, the null hypothesis can be rejected. Women spend more on theatre visits than men.

(d) We deal with dependent (paired) data because different variables (expenditure this year versus expenditure last year) are measured on the same observations. Thus, we need to use the paired t-test—which we can use in *R* by specifying the `paired` option:

```
t.test(Theatre,Theatre_ly,paired=TRUE)
```

```
            Paired t-test

data:  Theatre and Theatre_ly
t = 1.0925, df = 698, p-value = 0.275
alternative hypothesis: true difference in means is != 0
95 percent confidence interval:
 -2.481496  8.707533
sample estimates:
mean of the differences
              3.113019
```

Both the p-value (which is greater than 0.05) and the confidence interval (overlapping with "0") state that the null hypothesis should be kept. The mean difference in expenditure (3.1 SFR) is not large enough to suggest that this difference is not caused by chance. Visitors spend, on average, about the same in the two years.

Solution to Exercise 10.11

(a) We can use the one-sample t-test to answer both questions:

```
t.test(temperature,mu=65,alternative='greater')
t.test(time,mu=30,alternative='less')
```

```
        One-sample t-test

data:  temperature
t = -11.006, df = 1265, p-value = 1
alternative hypothesis: true mean is greater than 65
...

data:  time
t = 23.291, df = 1265, p-value = 1
alternative hypothesis: true mean is less than 30
```

We cannot confirm that the mean delivery time is less than 30 min and that the mean temperature is greater than 65 °C. This is not surprising: we have already seen in Exercise 3.10 that the manager should be unsatisfied with the performance of his company.

(b) We can use the exact binomial test to investigate $H_0 : p \geq 0.15$ and $H_1 : p < 0.15$. For the `binom.test` command, we need to know the numbers of successes and failures, i.e. the number of deliveries where a free wine should have been

given to the customer. Applying the `table` commands yields 229 and 1037 deliveries, respectively.

```
table(free_wine)
binom.test(c(229,1037),p=0.15,alternative='less')
```

```
        Exact binomial test

data:  c(229, 1037)
number of successes = 229, number of trials = 1266,
p-value = 0.9988 alternative hypothesis: true probability
is less than 0.15
95 percent confidence interval:
 0.0000000 0.1996186
sample estimates, probability of success: 0.1808847
```

The null hypothesis cannot be rejected because $p = 0.9988 > \alpha = 0.05$ and because the confidence interval covers $p = 0.15$. We get the same results if we use the variable "got_wine" instead of "free_wine". While the test says that we cannot exclude the possibility that the probability of receiving a free wine is less than 15 %, the point estimate of 18 % suggests that the manager still has to improve the timeliness of the deliveries or stop the offer of free wine.

(c) We first need to create a new categorical variable (using `cut`) which divides the temperatures into two parts: below and above 65 °C. Then we can simply apply the test commands (`fisher.test`, `chisq.test`, `prop.test`) to the table of branch and temperature:

```
hot <- cut(temperature,breaks=c(-Inf,65,Inf))
fisher.test(table(hot,operator))
chisq.test(table(hot,operator))
prop.test(table(hot,operator))
```

We know that the two χ^2 tests lead to identical results. For both of them the p-value is 0.2283 which suggests that we should keep the null hypothesis. There is not enough evidence which would support that the proportion of hot pizzas differs by operator, i.e. that the two variables are independent! The test of Fisher yields the same result ($p = 0.2227$).

(d) The null hypothesis is that the proportion of deliveries is the same for each branch: $H_0 : p_{\text{East}} = p_{\text{West}} = p_{\text{Centre}}$. To test this hypothesis, we need a χ^2 goodness-of-fit test:

```
chisq.test(table(branch))
```

```
        Chi-squared test for given probabilities

data:  table(branch)
X-squared = 0.74408, df = 2, p-value = 0.6893
```

The null hypothesis of equal proportions is therefore not rejected ($p > \alpha = 0.05$).

(e) To compare three proportions we need to use the χ^2 homogeneity test:

```
prop.test(table(branch, operator))
```

```
X-squared = 0.15719, df = 2, p-value = 0.9244
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.5059382 0.5097561 0.4965517
```

We can see that the proportions are almost identical and that the null hypothesis is not rejected ($p = 0.9244$).

(f) To test this relationship, we use the χ^2-independence test:

```
chisq.test(table(driver, branch))
```

```
X-squared = 56.856, df = 8, p-value = 1.921e-09
```

The null hypothesis of independence is rejected.

Solution to Exercise 10.12 To test the hypothesis $p_{\text{Shalabh}} = p_{\text{Heumann}} = p_{\text{Schomaker}} = 1/3$ we use the χ^2 goodness-of-fit test. The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} = \frac{1}{111}[(110-111)^2 + (118-111)^2 + (105-111)^2]$$
$$= (1 + 49 + 36)/111 \approx 0.77$$

H_0 gets rejected if $\chi^2 = 0.77 > c_{3-1-0,1-0.01} = 9.21$. Thus, the null hypothesis is not rejected. We accept that all three authors took about one-third of the pictures.

Chapter 11

Solution to Exercise 11.1

(a) Calculating $\bar{x} = \frac{1}{6}(26 + 23 + 27 + 28 + 24 + 25) = 25.5$ and $\bar{y} = \frac{1}{6}(170 + 150 + 160 + 175 + 155 + 150) = 160$, we obtain the following table needed for the estimation of $\hat{\alpha}$ and $\hat{\beta}$:

Body mass index			Systolic blood pressure			
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	v_i
26	0.5	0.25	170	10	100	5
23	−2.5	6.25	150	−10	100	25
27	1.5	2.25	160	0	0	0
28	2.5	6.25	175	15	225	37.5
24	−1.5	2.25	155	−5	25	7.5
25	−0.5	0.25	150	−10	100	5
Total 153		17.5	960		550	80

With $\sum_i v_i = \sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 80$, it follows that $S_{xy} = 80$. Moreover, we get $S_{xx} = \sum_i (x_i - \bar{x})^2 = 17.5$ and $S_{yy} = \sum_i (y_i - \bar{y})^2 = 550$. The parameter estimates are therefore

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{80}{17.5} \approx 4.57,$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 160 - 4.57 \cdot 25.5 = 43.465.$$

A one-unit increase in the BMI therefore relates to a 4.57 unit increase in the blood pressure. The model suggests a positive association between BMI and systolic blood pressure. It is impossible to have a BMI of 0; therefore, $\hat{\alpha}$ cannot be interpreted meaningfully here.

(b) Using (11.14), we obtain R^2 as

$$R^2 = r^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}\right)^2 = \left(\frac{80}{\sqrt{17.5 \cdot 550}}\right)^2 \approx 0.66.$$

Thus 66 % of the data's variability can be explained by the model. The goodness of fit is good, but not perfect.

Solution to Exercise 11.2

(a) To estimate $\hat{\beta}$, we use the second equality from (11.6):

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Calculating $\bar{x} = 1.333$, $\bar{y} = 5.556$, $\sum_{i=1}^{n} x_i\, y_i = 62.96$, $\sum_{i=1}^{n} x_i^2 = 24.24$, and $\sum_{i=1}^{n} y_i^2 = 281.5$ leads to

$$\begin{aligned}
\hat{\beta} &= \frac{62.91 - 9 \cdot 1.333 \cdot 5.556}{24.24 - 9 \cdot 1.333^2} \approx \frac{-3.695}{8.248} \approx -0.45,\\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 5.556 + 0.45 \cdot 1.333 \approx 6.16,\\
\hat{y}_i &= 6.16 - 0.45 x_i.
\end{aligned}$$

For children who spend no time on the internet at all, this model predicts 6.16 h of deep sleep. Each hour spent on the internet decreases the time in deep sleep by 0.45 h which is 27 min.

(b) Using the results from (a) and $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \approx 3.678$ yields:

$$R^2 = r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{(-3.695)^2}{8.248 \cdot 3.678} \approx 0.45.$$

About 45 % of the variation can be explained by the model. The fit of the model to the data is neither very good nor very bad.

(c) After collecting the data in two vectors (`c()`), printing a summary of the linear model (`summary(lm())`) reproduces the results. A scatter plot can be produced by plotting the two vectors against each other (`plot()`). The regression line can be added with `abline()`:

```
it <- c(0.3,2.2,...,2.3)
sleep <- c(5.8,4.4,...,6.1)
summary(lm(sleep~it))
plot(it,sleep)
abline(a=6.16,b=-0.45)
```

The plot is displayed in Fig. B.20.

(d) Treating X as a binary variable yields the following values: 0, 1, 0, 0, 0, 1, 1, 0, 1. We therefore have $\bar{x} = 0.444$, $\sum x_i^2 = 4$, and $\sum x_i\, y_i = 4.4 + 5.0 + 4.8 + 6.1 = 20.3$. Since the Y-values stay the same we calculate

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{20.3 - 9 \cdot 5.556 \cdot 0.444}{4 - 9 \cdot 0.444^2} \approx -0.85.$$

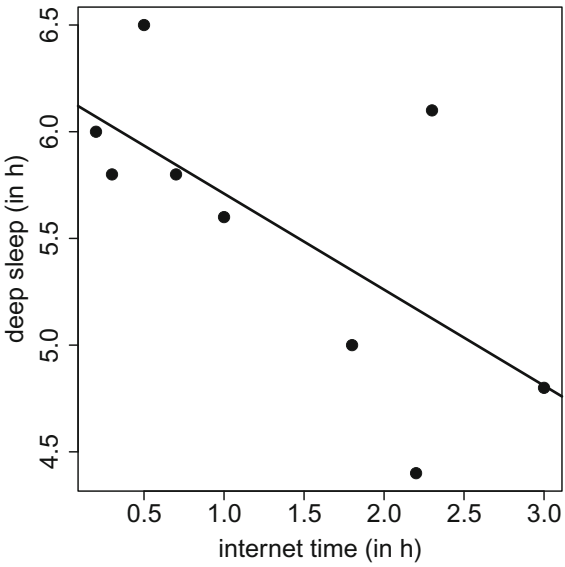


Fig. B.20 Scatter plot and regression line for the association of internet use and deep sleep

Thus, those children who are on the internet for a long time (i.e. >1 h) sleep on average 0.85 h (=51 min) less. If we change the coding of 1's and 0's, $\hat{\beta}$ will just have a different sign: $\hat{\beta} = 0.85$. In this case, we can conclude that children who spend less time on the internet sleep on average 0.85 h longer than children who spend more time on the internet. This is the same conclusion and highlights that the choice of coding does not affect the interpretation of the results.

Solution to Exercise 11.3

(a) The correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{yy}S_{xx}}} = \frac{170,821 - 17 \cdot 166.65 \cdot 60.12}{\sqrt{(62,184 - 17 \cdot 60.12^2)(472,569 - 17 \cdot 166.65^2)}}$$
$$= \frac{498.03}{\sqrt{738.955 \cdot 441.22}} = 0.87.$$

This indicates strong positive correlation: the higher the height, the higher the weight. Since $R^2 = r^2 = 0.87^2 \approx 0.76$, we already know that the fit of a linear regression model will be good (no matter whether height or weight is treated as outcome). From (11.11), we also know that $\hat{\beta}$ will be positive.

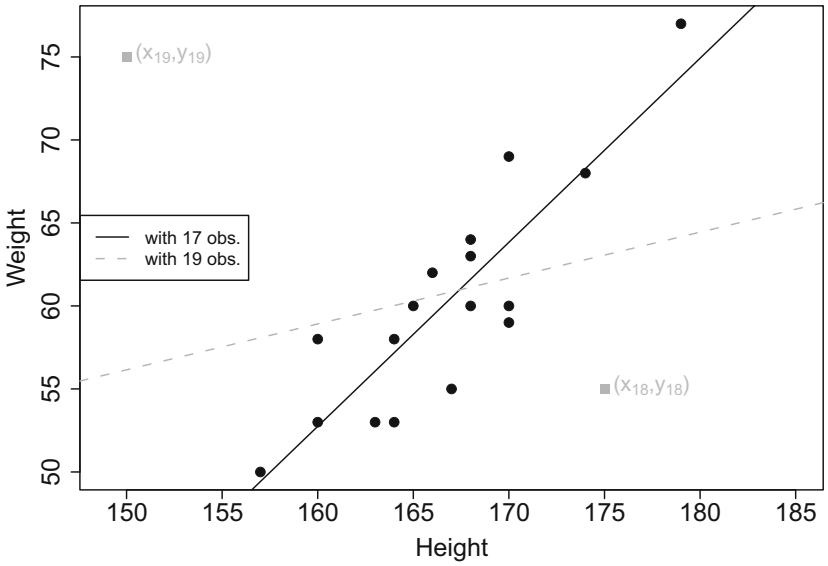


Fig. B.21 Scatter plot and regression line for both 17 and 19 observations

(b) We know from (a) that $S_{xy} = 498.03$ and that $S_{xx} = 441.22$. The least squares estimates are therefore

$$\hat{\beta} = \frac{498.03}{441.22} = 1.129,$$
$$\hat{\alpha} = 60.12 - 166.65 \cdot 1.129 = -128.03.$$

Each centimetre difference in height therefore means a 1.129 kg difference in weight. It is not possible to interpret $\hat{\alpha}$ meaningfully in this example.

(c) The prediction is

$$-128.03 + 1.129 \cdot 175 = 69.545\,\text{kg}.$$

(d)–(g) The black dots in Fig. B.21 show the scatter plot of the data. There is clearly a positive association in that greater height implies greater weight. This is also emphasized by the regression line estimated in (b). The two additional points appear in dark grey in the plot. It is obvious that they do not match the pattern observed in the original 17 data points. One may therefore speculate that with the inclusion of the two new points $\hat{\beta}$ will be smaller. To estimate the new regression line we need

$$\bar{x} = \frac{1}{19}(17 \cdot 166.65 + 150 + 175) = 166.21,$$
$$\bar{y} = \frac{1}{19}(17 \cdot 60.12 + 75 + 55) = 60.63.$$

This yields

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{191696 - 19 \cdot 166.21 \cdot 60.63}{525694 - 19 \cdot 166.21^2} = \frac{227.0663}{804.4821}$$
$$\approx 0.28.$$

This shows that the two added points shrink the estimate from 1.129 to 0.28. The association becomes less clear. This is an insightful example showing that least squares estimates are generally *sensitive to outliers* which can potentially affect the results.

Solution to Exercise 11.4

(a) The point estimate of β suggests a 0.077 % increase of hotel occupation for each one degree increase in temperature. However, the null hypothesis of $\beta = 0$ cannot be rejected because $p = 0.883 > 0.05$. We therefore cannot show an association between temperature and hotel occupation.

(b) The average hotel occupation is higher in Davos (7.9 %) and Polenca (0.9 %) compared with Basel (reference category). However, these differences are not significant. Both $H_0 : \beta_{\text{Davos}} = 0$ and $H_0 : \beta_{\text{Polenca}} = 0$ cannot be rejected. The model cannot show a significant difference in hotel occupation between Davos/Polenca and Basel.

(c) The analysis of variance table tells us that the null hypothesis of equal average temperatures in the three cities ($\beta_1 = \beta_2 = 0$) cannot be rejected. Note that in this example the overall F-test would have given us the same results.

(d) In the multivariate model, the main conclusions of (a) and (b) do not change: testing $H_0 : \beta_j = 0$ never leads to the rejection of the null hypothesis. We cannot show an association between temperature and hotel occupation (given the city); and we cannot show an association between city and hotel occupation (given the temperature).

(e) Stratifying the data yields considerably different results compared to (a)–(c): In Davos, where tourists go for skiing, each increase of 1 °C relates to a drop in hotel occupation of 2.7 %. The estimate $\hat{\beta} \approx -2.7$ is also significantly different from zero ($p = 0.000231$). In Polenca, a summer holiday destination, an increase of 1 °C implies an increase of hotel occupation of almost 4 %. This estimate is also significantly different from zero ($p = 0.00114 < 0.05$). In Basel, a business destination, there is a somewhat higher hotel occupation for higher temperatures ($\hat{\beta} = 1.3$); however, the estimate is not significantly different from zero. While there is no overall association between temperature and hotel occupation (see (a) and (c)), there is an association between them if one looks at the different cities separately. This suggests that an interaction between temperature and city should be included in the model.

(f) The design matrix contains a column of 1's (to model the intercept), the temperature and two dummies for the categorical variable "city" because it has three categories. The matrix also contains the interaction terms which are both the product of temperature and Davos and temperature and Polenca. The matrix has 36 rows because there are 36 observations: 12 for each city.

$$
\begin{array}{c}
 \\
1 \\
2 \\
\vdots \\
12 \\
13 \\
\vdots \\
24 \\
25 \\
\vdots \\
36
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
\text{Int.} & \text{Temp.} & \text{Davos} & \text{Polenca} & \text{Temp.} \times \text{Davos} & \text{Temp.} \times \text{Polenca}
\end{array} \\
\left(\begin{array}{cccccc}
1 & -6 & 1 & 0 & -6 & 0 \\
1 & -5 & 1 & 0 & -5 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 10 & 0 & 1 & 0 & 10 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 12 & 0 & 1 & 0 & 12 \\
1 & 1 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 4 & 0 & 0 & 0 & 0
\end{array}\right)
\end{array}
$$

(g) Both interaction terms are significantly different from zero ($p = 0.000375$ and $p = 0.033388$). The estimate of temperature therefore differs by city, and the estimate of city differs by temperature. For the reference city of Basel, the association between temperature and hotel occupation is estimated as 1.31; for Davos it is $1.31 - 4.00 = -2.69$ and for Polenca $1.31 + 2.66 = 3.97$. Note that these results are identical to (d) where we fitted three different regressions—they are just summarized in a different way.

(h) From (f) it follows that the point estimates for $\beta_{\text{temperature}}$ are 1.31 for Basel, -2.69 for Davos, and 3.97 for Polenca. Confidence intervals for these estimates can be obtained via (11.29):

$$(\hat{\beta}_i + \hat{\beta}_j) \pm t_{n-p-1;1-\alpha/2} \cdot \hat{\sigma}_{(\hat{\beta}_i + \hat{\beta}_j)}.$$

We calculate $t_{n-p-1;1-\alpha/2} = t_{36-5-1,0.975} = t_{30,0.975} = 2.04$. With $\text{Var}(\beta_{\text{temp.}}) = 0.478$ (obtained via 0.6916^2 from the model output or from the second row and second column of the covariance matrix), $\text{Var}(\beta_{\text{temp:Davos}}) = 0.997$, $\text{Var}(\beta_{\text{Polenca}}) = 1.43$, $\text{Cov}(\beta_{\text{temp.}}, \beta_{\text{temp:Davos}}) = -0.48$, and also $\text{Cov}(\beta_{\text{temp.}}, \beta_{\text{temp:Polenca}}) = -0.48$ we obtain:

$$
\begin{aligned}
\hat{\sigma}_{(\hat{\beta}_{\text{temp.}} + \hat{\beta}_{\text{Davos}})} &= \sqrt{0.478 + 0.997 - 2 \cdot 0.48} \approx 0.72, \\
\hat{\sigma}_{(\hat{\beta}_{\text{temp.}} + \hat{\beta}_{\text{Polenca}})} &= \sqrt{0.478 + 1.43 - 2 \cdot 0.48} \approx 0.97, \\
\hat{\sigma}_{(\hat{\beta}_{\text{temp.}} + \hat{\beta}_{\text{Basel}})} &= \sqrt{0.478 + 0 + 0} \approx 0.69.
\end{aligned}
$$

The 95 % confidence intervals are therefore:

$$\begin{aligned} \text{Davos:} \quad & [-2.69 \pm 2.04 \cdot 0.72] \approx [-4.2; -1.2], \\ \text{Polenca:} \quad & [3.97 \pm 2.04 \cdot 0.97] \approx [2.0; 5.9], \\ \text{Basel:} \quad & [1.31 \pm 2.04 \cdot 0.69] \approx [-0.1; 2.7]. \end{aligned}$$

Solution to Exercise 11.5

(a) The missing value [1] can be calculated as

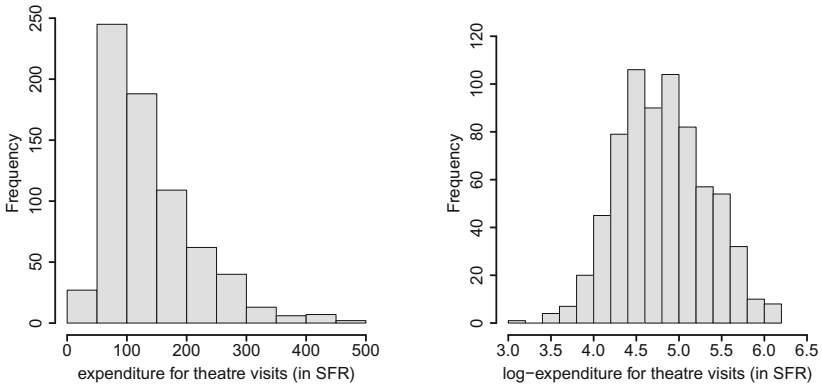
$$T = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} = \frac{0.39757 - 0}{0.19689} = 2.019.$$

Since $t_{699-5-1,0.975} = 1.96$ and $2.019 > 1.96$, it is clear that the p-value from [2] is smaller than 0.05. The exact p-value can be calculated in *R* via `(1-pt(2.019, 693))*2` which yields 0.0439. The `pt` command gives the probability value for the quantile of 2.019 (with 693 degrees of freedom): 0.978. Therefore, with probability $(1 - 0.978)$ % a value is right of 2.019 in the respective t-distribution which gives, multiplied by two to account for a two-sided test, the p-value.

(b)–(c) The plot on the left shows that the residuals are certainly not normally distributed as required by the model assumptions. The dots do not approximately match the bisecting line. There are too many high positive residuals which means that we are likely dealing with a right-skewed distribution of residuals. The plot on the right looks alright: no systematic pattern can be seen; it is a random plot. The histogram of both theatre expenditure and log(theatre expenditure) suggests that a log-transformation may improve the model, see Fig. B.22. Log-transformations are often helpful when the outcome's distribution is skewed to the right.

(d) Since the outcome is log-transformed, we can apply the interpretation of a log-linear model:

- Each year's increase in age yields an $\exp(0.0038) = 1.0038$ times higher (=0.38 %) expenditure on theatre visits. Therefore, a 10-year age difference relates to an $\exp(10 \cdot 0.0038) = 1.038$ times higher expenditure (=3.8 %).
- Women (gender = 1) spend on average (given the other variables) $\exp(0.179) \approx 1.20$ times more money on theatre visits.
- Each 1000 SFR more yearly income relates to an $\exp(0.0088) = 1.0088$ times higher expenditure on theatre visits. A difference in 10,000 SFR per year therefore amounts to an 8.8 % difference in expenditure.
- Each extra Swiss Franc spent on cultural activities is associated with an $\exp(0.00353) = 1.0035$ times higher expenditure on theatre visits.



(a) Expenditure (b) log(expenditure)

Fig. B.22 Histogram of the outcome

- Except for theatre expenditure from the preceding year, all β_j are significantly different from zero.

(e) While in (b) the residuals were clearly not normally distributed, this assumption seems to be fulfilled now: the QQ-plot shows dots which lie approximately on the bisecting line. The fitted values versus residuals plot remains a chaos plot. In conclusion, the log-transformation of the outcome helped to improve the quality of the model.

Solution to Exercise 11.6

(a) The multivariate model is obtained by using the `lm()` command and separating the covariates with the + sign. Applying `summary()` to the model returns the comprehensive summary.

```
mp <- lm(time ~ temperature + branch + day + operator + driver
+ bill + pizzas + discount_customer)
summary(mp)
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       40.42270    2.00446  20.166  < 2e-16 ***
temperature       -0.20823    0.02594  -8.027 2.28e-15 ***
branchEast        -1.60263    0.42331  -3.786 0.000160 ***
branchWest        -0.11912    0.37330  -0.319 0.749708
dayMonday         -1.15858    0.63300  -1.830 0.067443 .
daySaturday        0.88163    0.50161   1.758 0.079061 .
daySunday          1.01655    0.56103   1.812 0.070238 .
dayThursday        0.78895    0.53006   1.488 0.136895
dayTuesday         0.79284    0.62538   1.268 0.205117
dayWednesday       0.25814    0.60651   0.426 0.670468
operatorMelissa   -0.15791    0.34311  -0.460 0.645435
driverDomenico    -2.59296    0.73434  -3.531 0.000429 ***
driverLuigi       -0.80863    0.58724  -1.377 0.168760
driverMario       -0.39501    0.43678  -0.904 0.365973
driverSalvatore   -0.50410    0.43480  -1.159 0.246519
bill               0.14102    0.01600   8.811  < 2e-16 ***
pizzas             0.55618    0.11718   4.746 2.31e-06 ***
discount_customer -0.28321    0.36848  -0.769 0.442291
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

Residual standard error: 5.373 on 1248 degrees of freedom
Multiple R-squared:  0.3178,    Adjusted R-squared:  0.3085
F-statistic: 34.2 on 17 and 1248 DF,  p-value: < 2.2e-16
```

The output shows that lower temperature, higher bills, and more ordered pizzas increase the delivery times. The branch in the East is the fastest, and so is the driver Domenico. While there are differences with respect to day, discount customers, and the operator, they are not significant at the 5 % level.

(b) The confidence intervals are calculated as: $\hat{\beta}_i \pm t_{n-p-1;1-\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}}$. We know from the model output from (a) that there are 1248 degrees of freedom (1266 observations − 18 estimated coefficients). The respective quantile from the t-distribution is obtained with the `qt()` function. The coefficients are accessed via the `coefficients` command (alternatively: `mp$coefficients`); the variances of the coefficients are either accessed via the diagonal elements of the covariance matrix (`diag(vcov(mp))`) or the model summary (`summary(mp)[[4]][,2]`)—both of which are laborious. The summary of coefficients, lower confidence limit (lcl), and upper confidence limit (ucl) may be summarized in a matrix, e.g. via merging the individual columns with the `cbind` command.

```
lcl <- coefficients(mp) - qt(0.975,1248)*sqrt(diag(vcov(mp)))
ucl <- coefficients(mp) + qt(0.975,1248)*sqrt(diag(vcov(mp)))
cbind(coefficients(mp),lcl,ucl)
```

```
                                lcl        ucl
(Intercept)       40.4227014 36.4902223 44.3551805
temperature       -0.2082256 -0.2591146 -0.1573366
branchEast        -1.6026299 -2.4331162 -0.7721436
branchWest        -0.1191190 -0.8514880  0.6132501
dayMonday         -1.1585828 -2.4004457  0.0832801
...
```

(c) The variance is estimated as the residual sum of squares divided by the degrees of freedom, see also (11.27). Applying the `residuals` command to the model and using other basic operations yields an estimated variance of 28.86936.

```
sum(residuals(mp)^2)/(mp$df.residual)
```

Taking the square root of the result yields $\sqrt{28.86936} = 5.37$ which is also reported in the model output from (a) under "Residual standard error".

(d) The sum of squares error is defined as $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The total sum of squares is $\sum_{i=1}^{n}(y_i - \bar{y})^2$. This can be easily calculated in R. The goodness of fit is then obtained as $R^2 = 1 - SQ_{\text{Error}}/SQ_{\text{Total}} = 0.3178$. Dividing SQ_{Error} by $n - p - 1 = 1266 - 17 - 1 = 1248$ and SQ_{Total} by $n - 1 = 1265$ yields $R^2_{\text{adj}} = 0.3085$. This corresponds to the model output from (a).

```
SQE <- sum(residuals(mp)^2)
SQT <- sum((time-mean(time))^2)
1-(SQE/SQT)
1-((SQE/1248)/(SQT/1265))
```

(e) Applying `stepAIC` to the fitted model (with option "back" for backward selection) executes the model selection by means of AIC.

```
library(MASS)
stepAIC(mp, direction='back')
```

The output shows that the full model has an AIC of 4275.15. The smallest AIC is achieved by removing the operator variable from the model.

```
Step:  AIC=4275.15
time ~ temperature + branch + day + operator + driver + bill +
    pizzas + discount_customer

                    Df Sum of Sq   RSS    AIC
- operator           1      6.11 36035 4273.4
- discount_customer  1     17.05 36046 4273.8
<none>                           36029 4275.2
- day                6    448.79 36478 4278.8
- driver             4    363.91 36393 4279.9
- branch             2    511.10 36540 4289.0
- pizzas             1    650.39 36679 4295.8
- temperature        1   1860.36 37889 4336.9
- bill               1   2241.30 38270 4349.6
```

The reduced model has an AIC of 4273.37. Removing the discount customer variable from the model yields an improved AIC ($4272.0 < 4273.37$).

```
Step:  AIC=4273.37
time ~ temperature + branch + day + driver + bill + pizzas +
    discount_customer

                    Df Sum of Sq   RSS    AIC
- discount_customer  1     17.57 36053 4272.0
<none>                           36035 4273.4
- day                6    452.00 36487 4277.1
- driver             4    364.61 36400 4278.1
- branch             2    508.57 36544 4287.1
- pizzas             1    649.54 36685 4294.0
- temperature        1   1869.98 37905 4335.4
- bill               1   2236.19 38271 4347.6
```

The model selection procedure stops here as removing any variable would only increase the AIC, not decrease it.

```
Step:  AIC=4271.98
time ~ temperature + branch + day + driver + bill + pizzas

              Df Sum of Sq   RSS    AIC
<none>                    36053 4272.0
- day          6    455.62 36508 4275.9
- driver       4    368.18 36421 4276.8
- branch          2 513.17 36566 4285.9
- pizzas       1    657.07 36710 4292.8
- temperature  1   1878.24 37931 4334.3
- bill         1   2228.88 38282 4345.9
```

The final model, based on backward selection with AIC, includes day, driver, branch, number of pizzas ordered, temperature, and bill.

(f) Fitting the linear model with the variables obtained from (e) and obtaining the summary of it yields an R^2_{adj} of 0.3092.

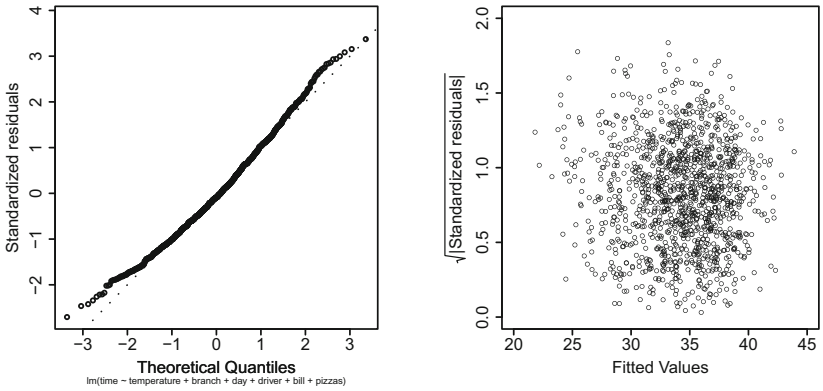
```
mps <- lm(time ~ temperature + branch + day + driver + bill +
pizzas)
summary(mps)
```

This is only marginally higher than the goodness of fit from the full model ($0.3092 > 0.3085$). While the selected model is better than the model with all variables, both, with respect to AIC and R^2_{adj}, the results are very close and remind us of the possible instability of applying automated model selection procedures.

(g) Both the normality assumption and heteroscedasticity can be checked by applying `plot()` to the model. From the many graphs provided we concentrate on the second and third of them:

```
plot(mps, which=2)
plot(mps, which=3)
```

Figure B.23a shows that the residuals are approximately normally distributed because the dots lie approximately on the bisecting line. There are some smaller deviations at the tails but they are not severe. The plot of the fitted values versus



(a) QQ-plot (b) Fitted values vs. residuals

Fig. B.23 Checking the model assumptions

the square root of the absolute values of the residuals shows no pattern; it is a random plot (Fig. B.23b). There seems to be no heteroscedasticity.

(h) Not all variables identified in (e) represent necessarily a "cause" for delayed or improved delivery time. It makes sense to speculate that *because* many pizzas are being delivered (and need to be made!) the delivery time increases. There might also be reasons why a certain driver is improving the delivery time: maybe he does not care about red lights. This could be investigated further given the results of the model above. However, high temperature does not *cause* the delivery time to be shorter; likely it is the other way around: the temperature is hotter because the delivery time is shorter. However, all of these considerations remain speculation. A regression model only exhibits associations. If there is a significant association, we know that given an accepted error (e.g. 5 %), values of **x** are higher when values of **y** are higher. This is useful but it does not say whether **x** caused **y** or vice versa.

(i) To check whether it is worth to add a polynomial, we simply add the squared temperature to the model. To make *R* understand that we apply a transformation we need to use `I()`.

```
mps2 <- lm(time ~ temperature + I(temperature^2) +
I(temperature^3) + branch + day + driver + bill + pizzas)
summary(mps2)
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -18.954965   8.795301  -2.155  0.03134 *
temperature        1.736692   0.282453   6.149 1.05e-09 ***
I(temperature^2)  -0.015544   0.002247  -6.917 7.36e-12 ***
branchEast        -1.429772   0.416107  -3.436  0.00061 ***
...
```

It can be seen that the null hypothesis $H_0 : \beta_{\text{temp}^2} = 0$ is rejected. This indicates that it is worthwhile to assume a quadratic relationship between temperature and delivery time.

(j) The prediction can be obtained by the `predict` command as follows:

```
predict(mps,pizza[1266,])
```

The prediction is 36.5 min and therefore 0.8 min higher than the real delivery time.

Appendix: Technical Appendix

C

More details on Chap. 3

Proof of equation (3.27).

$$\tilde{s}^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{x_i \in K_j}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{x_i \in K_j}(x_i - \bar{x}_j + \bar{x}_j - \bar{x})^2$$

$$= \underbrace{\frac{1}{n}\sum_{j=1}^{k}\sum_{x_i \in K_j}(x_i - \bar{x}_j)^2}_{[i]} + \underbrace{\frac{1}{n}\sum_{j=1}^{k}\sum_{x_i \in K_j}(\bar{x}_j - \bar{x})^2}_{[ii]}$$

$$+ \underbrace{\frac{2}{n}\sum_{j=1}^{k}\sum_{x_i \in K_j}(x_i - \bar{x}_j)(\bar{x}_j - \bar{x})}_{[iii]}$$

We obtain the following expressions for $[i]$–$[iii]$:

$$[i] = \frac{1}{n}\sum_{j=1}^{k} n_j \frac{1}{n_j}\sum_{x_i \in K_j}(x_i - \bar{x}_j)^2 = \frac{1}{n}\sum_{j=1}^{k} n_j \tilde{s}_j^2 \,,$$

$$[ii] = \frac{1}{n}\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 \,,$$

$$[iii] = \frac{2}{n}\sum_{j=1}^{k}(\bar{x}_j - \bar{x}) \sum_{x_i \in K_j}(x_i - \bar{x}_j) = \frac{2}{n}\sum_{j=1}^{k}(\bar{x}_j - \bar{x})\, 0 = 0.$$

Since $[i]$ is the within-class variance and $[ii]$ is the between-class variance, Eq. (3.27) holds.

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5

More details on Chap. 7

Proof of Theorem 7.2.3. Consider the interval $(x_0 - \delta, x_0]$ with $\delta \geq 0$. From (7.12) it follows that $P(x_0 - \delta < X \leq x_0) = F(x_0) - F(x_0 - \delta)$ and therefore

$$\begin{aligned} P(X = x_0) &= \lim_{\delta \to 0} P(x_0 - \delta < X \leq x_0) \\ &= \lim_{\delta \to 0} [F(x_0) - F(x_0 - \delta)] \\ &= F(x_0) - F(x_0) = 0. \end{aligned}$$

Proof of Theorem 7.3.1.

$$\begin{aligned} \mathrm{Var}(X) &\overset{(7.17)}{=} E(X - \mu)^2 \\ &= E(X^2 - 2\mu X + \mu^2) \\ &\overset{(7.28-7.31)}{=} E(X^2) - 2\mu E(X) + E(\mu^2) \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - [E(X)]^2. \end{aligned}$$

Proof of Theorem 7.4.1. We define a discrete random variable Y as

$$Y = \begin{cases} 0 & \text{if } |X - \mu| < c \\ c^2 & \text{if } |X - \mu| \geq c. \end{cases} \tag{C.1}$$

The respective probabilities are $P(|X - \mu| < c) = p_1$ and $P(|X - \mu| \geq c) = p_2$. The definition of Y in (C.1) implies that

$$Y \leq |X - \mu|^2$$

since for $|X - \mu|^2 < c^2$ Y takes the value $y_1 = 0$ and therefore $Y \leq |X - \mu|^2$. If $|X - \mu|^2 \geq c^2 Y$ takes the value $y_2 = c^2$, and therefore $Y \leq |X - \mu|^2$. Using this knowledge, we obtain

$$\mathrm{E}(Y) \leq \mathrm{E}(X - \mu)^2 = \mathrm{Var}(X).$$

However, for Y we also have

$$\mathrm{E}(Y) = 0 \cdot p_1 + c^2 \cdot p_2 = c^2 P(|X - \mu| \geq c)$$

which means that we can summarize the above findings in the following inequality:

$$c^2 P(|X - \mu| \geq c) \leq \mathrm{Var}(X).$$

This equates to Tschebyschev's inequality. Using $P(\bar{A}) = 1 - P(A)$, i.e.

$$P(|X - \mu| \geq c) = 1 - P(|X - \mu| < c),$$

we obtain the second formula of Tschebyschev's inequality:

$$P(|X - \mu| < c) \geq 1 - \frac{\mathrm{Var}(X)}{c^2}.$$

Proof of rule (7.30). For discrete variables, we have

$$\mathrm{E}(a+bX) \overset{(7.16)}{=} \sum_i (a+bx_i)\,p_i = \sum ap_i + bx_i\,p_i = a\sum_i p_i + b\sum_i x_i\,p_i.$$

Since $\sum_i p_i = 1$ and $\sum_i x_i\,p_i = \mathrm{E}(X)$, we obtain $\mathrm{E}(a+bX) = a + b\mathrm{E}(X)$, which is rule (7.30). In the continuous case, we have

$$\begin{aligned}\mathrm{E}(a+bX) &= \int_{-\infty}^{\infty} (a+bx)\,f(x)\mathrm{d}x = \int_{-\infty}^{\infty} (af(x)dx + bxf(x)\mathrm{d}x)\\ &= a\int_{-\infty}^{\infty} f(x)\mathrm{d}x + b\int_{-\infty}^{\infty} xf(x)\mathrm{d}x = a + b\mathrm{E}(X).\end{aligned}$$

Proof of rule (7.33). Using $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$, we can write the variance of bX as

$$\mathrm{Var}(bX) = \mathrm{E}([bX]^2) + \mathrm{E}(bX)^2.$$

Using (7.29), we get $\mathrm{E}([bX]^2) = b^2\mathrm{E}(X^2)$ and $\mathrm{E}(bX)^2 = (b\mathrm{E}(X))^2$. Therefore

$$\mathrm{Var}(bX) = b^2(\mathrm{E}(X^2) - \mathrm{E}(X)^2) = b^2\mathrm{Var}(X).$$

Proof of $\rho = 1$ *for a perfect linear relationship.* If $Y = aX + b$ with $a \neq 0$, we get

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]\\ &= a\mathrm{E}[(X-\mu_X)(X-\mu_X)]\end{aligned}$$

because $\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y)$, see p. 147. Then

$$\begin{aligned}\mathrm{Cov}(X,Y) &= a\mathrm{Var}(X),\\ \mathrm{Var}(Y) &\overset{(7.33)}{=} a^2\mathrm{Var}(X),\end{aligned}$$

and therefore

$$\rho(X,Y) = \frac{a\mathrm{Var}(X)}{\sqrt{a^2\mathrm{Var}(X)\mathrm{Var}(X)}} = \frac{a}{|a|} = 1$$

if $a > 0$. Similarly, if $Y = aX + b$ with $a < 0$ we get $\rho(X,Y) = -1$.

More details on Chap. 8

Theorem of Large Numbers. To explain the Theorem of Large Numbers, we first need to first define **stochastic convergence**.

Definition C.1 A sequence of random variables, $(X_n)_{n\in\mathbb{N}}$, converges stochastically to 0, if for any $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n| > \epsilon) = 0 \tag{C.2}$$

holds.

This is equivalent to $\lim_{n\to\infty} P(|X_n| \leq \epsilon) = 1$.

Theorem C.1 (Theorem of large numbers) *Consider n i.i.d. random variables $X_1, X_2, \ldots, X_n$ with $\mathrm{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. It holds that*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < c) = 1, \quad \forall\ c \geq 0. \tag{C.3}$$

This implies that $\bar{X}_n$ converges stochastically to μ. As another motivation, recall Definition 7.6.1 where we define random variables $X_1, X_2, \ldots, X_n$ to be i.i.d. (independently identically distributed) if all X_i follow the same distribution and are independent of each other. Under this assumption, we showed in (7.36) that $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$. It follows that the larger n, the smaller the variance. If we apply Tschebyschev's inequality (Theorem 7.4.1) to $\bar{X}$, we get the following equation for $(\bar{X}_n - \mu)_{n\in\mathbb{N}}$:

$$P(|\bar{X}_n - \mu| < c) \geq 1 - \frac{\mathrm{Var}(\bar{X}_n)}{c^2} = 1 - \frac{\sigma^2}{nc^2}. \tag{C.4}$$

This means that for each $c \geq 0$, the right-hand side of the above equation tends to 1 as $n \to \infty$ which gives a similar interpretation as the Theorem of Large Numbers.

Central Limit Theorem. Let X_i $(i = 1, 2, \ldots, n)$ be n i.i.d. random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. If we consider the sum $\sum_{i=1}^n X_i$, we obtain $\mathrm{E}(\sum_{i=1}^n X_i) = n\mu$ and $\mathrm{Var}(\sum_{i=1}^n X_i) = n\sigma^2$. If we want to standardize $\sum_{i=1}^n X_i$ we can use Theorem 7.3.2 to obtain

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}}, \tag{C.5}$$

i.e. it holds that $\mathrm{E}(Y_n) = 0$ and $\mathrm{Var}(Y_n) = 1$.

Theorem C.2 (Central Limit Theorem) *Let X_i $(i = 1, 2, \ldots, n)$ be n i.i.d. random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Y_n denotes the standardized sum of X_i, $i = 1, 2, \ldots, n$. The CDF of Y_n is*

$$\lim_{n\to\infty} P(Y_n \leq y) = \Phi(y), \quad \forall\ y,$$

where $\Phi(y)$ denotes the CDF of the standard normal distribution.

This theorem tells us that Y_n is approximately standard normal distributed if n is large, i.e.

$$Y_n \sim N(0, 1) \quad \text{as} \quad n \to \infty.$$

This is equivalent to saying that $\sum_{i=1}^n X_i$ is approximately $N(n\mu, n\sigma^2)$ distributed if n is large, i.e.

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2). \tag{C.6}$$

As a consequence $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is $N(\mu, \frac{\sigma^2}{n})$ distributed for large n, i.e.

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

In summary, the Theorem of Large Numbers tells us that $\bar{X}_n$ converges stochastically to μ, whereas the Central Limit Theorem tells us that $\bar{X}_n$ converges to a $N(\mu, \frac{\sigma^2}{n})$-distribution as n tends to infinity.

PDF of the χ^2-Distribution. The PDF of the χ^2-distribution, with n degrees of freedom, is defined as

$$f(x) = \begin{cases} \frac{x^{n/2-1}\exp(-x/2)}{\Gamma(n/2)2^{n/2}} & \text{if} \quad x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{C.7}$$

Note that $\Gamma(n)$ is the Gamma function, defined as $\Gamma(n) = (n-1)!$ for positive integers and $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)\mathrm{d}t$ otherwise.

PDF of the t-Distribution. The PDF of the t-distribution, with n degrees of freedom, is defined as

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n/2)\sqrt{n\pi}}\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} \qquad -\infty < x < \infty. \tag{C.8}$$

Note that $\Gamma(n)$ is the Gamma function, defined as $\Gamma(n) = (n-1)!$ for positive integers and $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)\mathrm{d}t$ otherwise.

PDF of the F-Distribution. The PDF of the F-distribution, with n and m degrees of freedom, respectively, is defined as

$$f(x) = \frac{\Gamma(\frac{n+m}{2})(\frac{n}{m})^{n/2}x^{n/2-1}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})(1+\frac{nx}{m})^{(n+m)/2}}, \quad x > 0. \tag{C.9}$$

The PDF of the F-distribution with m and n degrees of freedom can be derived by interchanging the roles of m and n.

More details on Chap. 9

Another Example of Sufficiency. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where μ and σ^2 are both unknown. We attempt to find a sufficient statistic for μ and σ^2.

$$f(x_1, x_2 \ldots, x_n; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\right]$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right].$$

Here the joint density depends on $x_1, x_2, \ldots, x_n$ through two statistics $t_1(x_1, x_2, \ldots, x_n) = \sum_{i=1}^n x_i$ and $t_2(x_1, x_2, \ldots, x_n) = \sum_{i=1}^n x_i^2$ with $h(x_1, x_2, \ldots, x_n) = 1$. Thus $T_1 = \sum_{i=1}^n X_i$ and $T_2 = \sum_{i=1}^n X_i^2$ are jointly sufficient for μ and σ^2. Therefore, $\bar{X}$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are jointly sufficient for μ and σ^2 as they are a one-to-one function of T_1 and T_2. They are the maximum likelihood estimates for μ and σ^2.

More details on Chap. 10

Exact Test of Fisher. Similar to the approximate two-sample binomial test in Sect. 10.4.2, we assume two samples following a binomial distribution with parameters (n_1, p_1) and (n_2, p_2), respectively.

$$\mathbf{X} = (X_1, X_2, \ldots, X_{n_1}), \quad X_i \sim B(1; p_1)$$
$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{n_2}), \quad Y_i \sim B(1; p_2).$$

For the sums of these random variables, we get:

$$X = \sum_{i=1}^{n_1} X_i \sim B(n_1; p_1), \quad Y = \sum_{i=1}^{n_2} Y_i \sim B(n_2; p_2).$$

Let $Z = X + Y$. The Exact Test of Fisher uses the fact that the row marginal frequencies n_1 and n_2 in the following table

	Success	Failure	Total
Population A	X	$n_1 - X$	n_1
Population B	$Z - X = Y$	$n_2 - (Z - X)$	n_2
Total	Z	$(n_1 + n_2 - Z)$	$n_1 + n_2$

are fixed by the sample sizes n_1 and n_2. Conditional on the total number of successes $Z = z$ (i.e. the column margins are assumed to be fixed), the only remaining random variable is X (since the other three entries of the table are then determined by the realization x of X and the margins). Under $H_0 : p_1 = p_2 = p$, it can be shown that

$$X \sim H(n, n_1, z),$$

i.e.

$$P(X = x|Z = z) = \frac{\binom{n_1}{x}\binom{n-n_1}{z-x}}{\binom{n}{z}}.$$

The proof is straightforward using the idea of conditioning on $Z = z$ and the assumption of independence of the two samples:

$$\begin{aligned} P(X = x|Z = z) = \frac{P(X = x, Z = z)}{P(Z = z)} &= \frac{P(X = x, Y = z - x)}{P(Z = z)} \\ &= \frac{P(X = x)P(Y = z - x)}{P(Z = z)} \\ &= \frac{\binom{n_1}{x} p^x (1-p)^{n_1 - x} \binom{n_2}{z-x} p^{z-x} (1-p)^{n_2 - (z-x)}}{\binom{n}{z} p^z (1-p)^{n-z}} \\ &= \frac{\binom{n_1}{x}\binom{n_2}{z-x}}{\binom{n}{z}} = \frac{\binom{n_1}{x}\binom{n-n_1}{z-x}}{\binom{n}{z}}. \end{aligned}$$

Note that in the equation above we use the fact that under H_0, $Z = X + Y$ is $B(n, p)$ distributed; see the additivity theorem for the binomial distribution, i.e. Theorem 8.1.1.

Example C.1 Consider two competing lotteries A and B. Say we buy 10 tickets from each lottery and test the hypothesis of equal winning probabilities. The data can be summarized in a 2×2 table:

	Winning	Not winning	Total
Lottery A	1	24	25
Lottery B	7	18	25
Total	8	42	50

In *R*, we can use the command `fisher.test` to perform the test. Using the example data and $H_1 : p_1 \neq p_2$, we get

```
ft <- matrix(nrow=2,ncol=2,data=cbind(c(1,7), c(24,18)))
fisher.test(x=ft)
```

with output

```
    Fisher's Exact Test for Count Data

data:  fisher.table
p-value = 0.0488
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
 0.002289885 0.992114690
sample estimates:
odds ratio
 0.1114886
```

For the example data and $\alpha = 0.05$, the null hypothesis is rejected, since the p-value is lower than α. For the calculation of the p-value, the one-sided and two-sided cases have to be distinguished. The idea is that while fixing the margins at the observed values (i.e. 25, 25, 8, 42), we have to calculate the sum of the probabilities of all tables which have lower probability than the observed table. In R, one can use the functions `dhyper` and `phyper` for calculating (cumulative) probabilities. For example, $P(X = 1|Z = 8)$ can be calculated as

```
dhyper(1,25,25,8)
```

```
[1] 0.02238402
```

A more extreme table than the observed one is

0	25	25
8	17	25
8	42	50

with probability $P(X = 0) =$ `dhyper(0,25,25,8)` $= 0.002$, which is lower than $P(X = 1)$. The sum is $0.0224 + 0.002 = 0.0244$ which is the (left) one-sided p-value. In this case (not generally true!), the two-sided p-value is simply two times the one-sided value, i.e. $2 \cdot 0.0244 = 0.0488$.

Remark C.1 The two-sided version of the Exact Test of Fisher can also be used as a test of independence of two binary variables. It is equivalent to the test of equality of two proportions, see Example 10.8.2.

One-Sample χ^2-Test for Testing Hypothesis About the Variance. We assume a normal population, i.e. $X \sim N(\mu, \sigma^2)$ and an i.i.d. sample $(X_1, X_2, \ldots, X_n)$ distributed as X. We only introduce the test for two-sided hypotheses

$$H_0 : \sigma^2 = \sigma_0^2$$

versus

$$H_1 : \sigma^2 \neq \sigma_0^2.$$

The test statistic

$$T(\mathbf{X}) = \frac{(n-1)S_X^2}{\sigma_0^2}$$

follows a χ^2_{n-1}-distribution under H_0. The critical region is constructed by taking the $\alpha/2$- and $(1-\alpha/2)$ quantile as critical values; i.e. H_0 is rejected, if

$$t(x) < c_{n-1;\alpha/2}$$

or if

$$t(x) > c_{n-1;1-\alpha/2},$$

where $c_{n-1;\alpha/2}$ and $c_{n-1;1-\alpha/2}$ are the desired quantiles of a χ^2-distribution. In *R*, the test can be called by the `sigma.test` function in the `TeachingDemos` library or the `varTest` function in library `EnvStats`. Both functions also return a confidence interval for the desired confidence level. Note that the test is biased. An unbiased level α test would not take $\alpha/2$ at the tails but two different tail probabilities α_1 and α_2 with $\alpha_1 + \alpha_2 = \alpha$.

***F*-Test for Comparing Two Variances.** Comparing variances can be of interest when comparing the variability, i.e. the "precision" of two industrial processes; or when comparing two medical treatments with respect to their reliability. Consider two populations characterized by two independent random variables X and Y which follow normal distributions:

$$X \sim N(\mu_X, \sigma_X^2), \qquad Y \sim N(\mu_Y, \sigma_Y^2).$$

For now, we distinguish the following two hypotheses:

$$\begin{aligned} H_0 &: \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2, \qquad \text{two-sided} \\ H_0 &: \sigma_X^2 \leq \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 > \sigma_Y^2, \qquad \text{one-sided.} \end{aligned}$$

The third hypothesis with $H_1 : \sigma_X^2 < \sigma_Y^2$ is similar to the second hypothesis where X and Y are replaced with each other.

Test Statistic

Let $(X_1, X_2, \ldots, X_{n_1})$ and $(Y_1, Y_2, \ldots, Y_{n_2})$ be two independent random samples of size n_1 and n_2. The test statistic is defined as the ratio of the two sample variances

$$T(\mathbf{X}, \mathbf{Y}) = \frac{S_X^2}{S_Y^2}, \tag{C.10}$$

which is, under the null hypothesis, F-distributed with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, see also Sect. 8.3.3.

Critical Region

Two-Sided Case. The motivation behind the construction of the critical region for the two-sided case, H_0: $\sigma_X^2 = \sigma_Y^2$ vs. H_1: $\sigma_X^2 \neq \sigma_Y^2$, is that if the null hypothesis is true (i.e. the two variances are equal) then the test statistic (C.10) would be 1; also, $T(\mathbf{X}, \mathbf{Y}) > 0$. Therefore, very small (but positive) and very large values of $T(\mathbf{X}, \mathbf{Y})$ should lead to a rejection of H_0. The critical region can then be written as $K = [0, k_1) \cup (k_2, \infty)$, where k_1 and k_2 are critical values such that

$$P(T(\mathbf{X}, \mathbf{Y}) < k_1 | H_0) = \alpha/2$$
$$P(T(\mathbf{X}, \mathbf{Y}) > k_2 | H_0) = \alpha/2.$$

Here k_1 and k_2 can be calculated from the quantile function of the F-distribution as

$$k_1 = f_{n_1-1, n_2-1, \alpha/2}, \qquad k_2 = f_{n_1-1, n_2-1, 1-\alpha/2}.$$

Example C.2 Let $n_1 = 50$, $n_2 = 60$ and $\alpha = 0.05$. Using the `qf` command in *R*, we can determine the critical values as:

```
qf(q=0.025, df1=50-1, df2=60-1)
qf(q=0.975, df1=50-1, df2=60-1)
```

The results are $k_1 = 0.5778867$ and $k_2 = 1.706867$.

Remark C.2 There is the following relationship between quantiles of the F-distribution:

$$f_{n_1-1;n_2-1;\alpha/2} = \frac{1}{f_{n_2-1;n_1-1;1-\alpha/2}}.$$

One-Sided Case. In the one-sided case, the alternative hypothesis is always formulated in such a way that the variance in the numerator is greater than the variance in the denominator. The hypotheses are H_0: $\sigma_X^2 \leq \sigma_Y^2$ versus H_1: $\sigma_X^2 > \sigma_Y^2$ or H_0: $\sigma_X^2/\sigma_Y^2 \leq 1$ versus H_1: $\sigma_X^2/\sigma_Y^2 > 1$. This means it does not matter whether $H_1 : \sigma_X^2 > \sigma_Y^2$ or $H_1 : \sigma_X^2 < \sigma_Y^2$; by constructing the test statistic in the correct way, we implicitly specify the hypothesis. The critical region K consists of the largest values of $T(\mathbf{X}, \mathbf{Y})$, i.e. $K = (k, \infty)$, where k has to fulfil the condition

$$P(T(\mathbf{X}, \mathbf{Y}) > k | H_0) = \alpha.$$

The resulting critical value is denoted by $k = f_{n_1-1;n_2-1;1-\alpha}$.

Observed Test Statistic

Using the sample variances, the realized test statistic t is calculated as:

$$t(x, y) = \frac{s_x^2}{s_y^2}, \quad s_x^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2.$$

Test Decisions

Case	H_0	H_1	Reject H_0, if
(a)	$\sigma_X = \sigma_Y$	$\sigma_X \neq \sigma_Y$	$t(x, y) > f_{n_1-1;n_2-1;1-\alpha/2}$ or
			$t(x, y) < f_{n_1-1;n_2-1;\alpha/2}$
(b)	$\sigma_X \leq \sigma_Y$	$\sigma_X > \sigma_Y$	$t(x, y) > f_{n_1-1;n_2-1;1-\alpha}$

Remark C.3 We have tacitly assumed that the expected values μ_X and μ_y are unknown and have to be estimated. However, this happens rarely, if ever, in practice. When estimating the expected values by the arithmetic means, it would be appropriate to increase the degrees of freedom from $n_1 - 1$ to n_1 and $n_2 - 1$ to n_2. Interestingly, standard software will not handle this case correctly.

Example C.3 A company is putting baked beans into cans. Two independent machines at two sites are used. The filling weights are assumed to be normally distributed with mean 1000 g. It is speculated that one machine is more precise than the other. Two samples of the two machines give the following results:

Sample	n	$\bar{x}$	s^2
X	20	1000.49	72.38
Y	25	1000.26	45.42

With $\alpha = 0.1$ and the hypotheses

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2,$$

we get the following quantiles

```
qf(0.05, 20-1, 25-1)
[1] 0.4730049
qf(0.95, 20-1, 25-1)
[1] 2.039858
```

The observed test statistic is

$$t = \frac{72.38}{45.42} = 1.59.$$

Therefore, H_0 is not rejected, since $k_1 \leq t \leq k_2$. We cannot reject the hypothesis of equal variability of the two machines. In *R*, the F-test can be used using the command `var.test`.

Remark C.4 For the t-test, we remarked that the assumption of normality is not crucial because the test statistic is approximately normally distributed, even for moderate sample sizes. However, the F-test relies heavily on the assumption of normality. This is why alternative tests are often used, for example the Levene's test.

More details on Chap. 11

Obtaining the Least Squares Estimates in the Linear Model. The function $S(a, b)$ describes our optimization problem of minimizing the residual sum of squares:

$$S(a, b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Minimizing $S(a, b)$ is achieved using the principle of maxima and minima which involves taking the partial derivatives of $S(a, b)$ with respect to both a and b and setting them equal to 0. The partial derivatives are

$$\frac{\partial}{\partial a} S(a, b) = \sum_{i=1}^{n} \frac{\partial}{\partial a} (y_i - a - bx_i)^2 = -2 \sum_{i=1}^{n} (y_i - a - bx_i), \tag{C.11}$$

$$\frac{\partial}{\partial b} S(a, b) = \sum_{i=1}^{n} \frac{\partial}{\partial b} (y_i - a - bx_i)^2 = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i. \tag{C.12}$$

Now we set (C.11) and (C.12) as equal to zero, respectively:

$$\begin{aligned} &\text{(I)} \quad \textstyle\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i) = 0, \\ &\text{(II)} \quad \textstyle\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i) x_i = 0. \end{aligned}$$

This equates to

$$\begin{aligned} &(\mathrm{I}') \qquad n\hat{a} + \hat{b}\textstyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ &(\mathrm{II}') \; \hat{a}\textstyle\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{aligned}$$

Multiplying (I′) by $\frac{1}{n}$ yields

$$\hat{a} + \hat{b}\bar{x} = \bar{y}$$

which gives us the solution for a:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Putting this solution into (II′) gives us

$$(\bar{y} - \hat{b}\bar{x}) \sum_{i=1}^{n} x_i + \hat{b} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Using $\sum_{i=1}^{n} x_i = n\bar{x}$ leads to

$$\hat{b} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

If we use

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

and

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = S_{xy},$$

we eventually obtain the least squares estimate of b:

$$\begin{aligned} \hat{b}\, S_{xx} &= S_{xy} \\ \hat{b} &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \end{aligned}$$

Remark C.5 To show that the above solutions really relate to a minimum, and not to a maximum, we would need to look at all the second-order partial derivatives of $S(a, b)$ and prove that the bordered Hessian matrix containing these derivatives is always positive definite. We omit this proof however.

Variance Decomposition.

We start with the following equation:

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}).$$

If we square both sides we obtain

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 - 2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}).$$

The last term on the right-hand side is

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) &\overset{(11.8)}{=} \sum_{i=1}^{n}(y_i - \bar{y})\hat{b}(x_i - \bar{x}) \\ &= \hat{b}\, S_{xy} \overset{(11.6)}{=} \hat{b}^2 S_{xx} \overset{(11.8)}{=} \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.\end{aligned}$$

We therefore obtain

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

which equates to

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

The Relation between R^2 and r.

$$\begin{aligned}SQ_{\text{Residual}} &= \sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2 \overset{(11.8)}{=} \sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})]^2 \\ &= S_{yy} + \hat{b}^2 S_{xx} - 2\hat{b}S_{xy} \\ &= S_{yy} - \hat{b}^2 S_{xx} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \\ SQ_{\text{Regression}} &= S_{yy} - SQ_{\text{Residual}} = \frac{(S_{xy})^2}{S_{xx}}.\end{aligned}$$

We therefore obtain

$$R^2 = \frac{SQ_{\text{Regression}}}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx}S_{yy}} = r^2.$$

The Least Squares Estimators are Unbiased.

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})$$

Given that $\mathbf{X}$ in the model is assumed to be fixed (i.e. non-stochastic and not following any distribution), we obtain

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{y}).$$

Since $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ it follows that $\mathrm{E}\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ and therefore

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

How to Obtain the Variance of the Least Squares Estimator. With the same arguments as above (i.e $\mathbf{X}$ is fixed and non-stochastic) and applying the rule $\mathrm{Var}(bX) = b^2\mathrm{Var}(X)$ from the scalar case to matrices we obtain:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Maximum Likelihood Estimation in the Linear Model. The linear model follows a normal distribution:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

Therefore, the likelihood function of $\mathbf{y}$ also follows a normal distribution:

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}.$$

The log-likelihood function is given by

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

To obtain the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2, one needs to obtain the maxima of the above function using the principle of maxima and minima that involves setting the partial derivatives equal to zero and finding the solution:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{2\sigma^2} 2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0},$$
$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

We therefore have

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}, \text{ or } \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

which give us the ML estimates of $\boldsymbol{\beta}$ and σ^2 in the linear regression model. Here, we also need to check the Hessian matrix of second-order partial derivatives to show that we really found a minimum and not a maximum. We omit this proof however.

Distribution Tables

See Tables C.1, C.2 and C.3.

Table C.1 CDF values for the standard normal distribution, $\Phi(z)$. These values can also be obtained in *R* by using the `pnorm(p)` command

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965620	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691

(continued)

Table C.1 (continued)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999

Table C.2 $(1-\alpha)$ quantiles for the t-distribution. These values can also be obtained in *R* using the `qt(p,df)` command.

df	$1-\alpha$			
	0.95	0.975	0.99	0.995
1	6.3138	12.706	31.821	63.657
2	2.9200	4.3027	6.9646	9.9248
3	2.3534	3.1824	4.5407	5.8409
4	2.1318	2.7764	3.7469	4.6041
5	2.0150	2.5706	3.3649	4.0321
6	1.9432	2.4469	3.1427	3.7074
7	1.8946	2.3646	2.9980	3.4995
8	1.8595	2.3060	2.8965	3.3554
9	1.8331	2.2622	2.8214	3.2498
10	1.8125	2.2281	2.7638	3.1693
11	1.7959	2.2010	2.7181	3.1058
12	1.7823	2.1788	2.6810	3.0545
13	1.7709	2.1604	2.6503	3.0123
14	1.7613	2.1448	2.6245	2.9768
15	1.7531	2.1314	2.6025	2.9467
16	1.7459	2.1199	2.5835	2.9208
17	1.7396	2.1098	2.5669	2.8982
18	1.7341	2.1009	2.5524	2.8784
19	1.7291	2.0930	2.5395	2.8609
20	1.7247	2.0860	2.5280	2.8453
30	1.6973	2.0423	2.4573	2.7500
40	1.6839	2.0211	2.4233	2.7045
50	1.6759	2.0086	2.4033	2.6778
60	1.6706	2.0003	2.3901	2.6603
70	1.6669	1.9944	2.3808	2.6479
80	1.6641	1.9901	2.3739	2.6387
90	1.6620	1.9867	2.3685	2.6316
100	1.6602	1.9840	2.3642	2.6259
200	1.6525	1.9719	2.3451	2.6006
300	1.6499	1.9679	2.3388	2.5923
400	1.6487	1.9659	2.3357	2.5882
500	1.6479	1.9647	2.3338	2.5857

Table C.3 $(1-\alpha)$ quantiles of the χ^2-distribution. These values can also be obtained in *R* using the `qchisq(p,df)` command

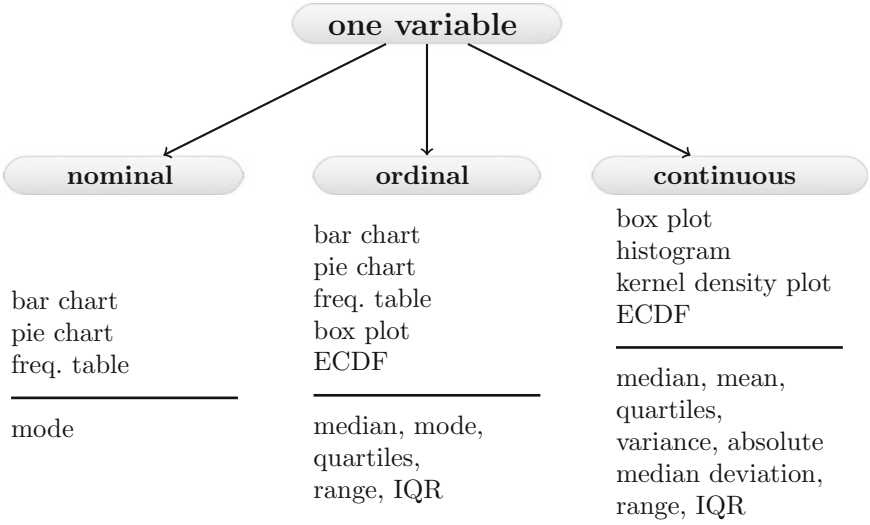
df	$1-\alpha$					
	0.01	0.025	0.05	0.95	0.975	0.99
1	0.0001	0.001	0.004	3.84	5.02	6.62
2	0.020	0.051	0.103	5.99	7.38	9.21
3	0.115	0.216	0.352	7.81	9.35	11.3
4	0.297	0.484	0.711	9.49	11.1	13.3
5	0.554	0.831	1.15	11.1	12.8	15.1
6	0.872	1.24	1.64	12.6	14.4	16.8
7	1.24	1.69	2.17	14.1	16.0	18.5
8	1.65	2.18	2.73	15.5	17.5	20.1
9	2.09	2.70	3.33	16.9	19.0	21.7
10	2.56	3.25	3.94	18.3	20.5	23.2
11	3.05	3.82	4.57	19.7	21.9	24.7
12	3.57	4.40	5.23	21.0	23.3	26.2
13	4.11	5.01	5.89	22.4	24.7	27.7
14	4.66	5.63	6.57	23.7	26.1	29.1
15	5.23	6.26	7.26	25.0	27.5	30.6
16	5.81	6.91	7.96	26.3	28.8	32.0
17	6.41	7.56	8.67	27.6	30.2	33.4
18	7.01	8.23	9.39	28.9	31.5	34.8
19	7.63	8.91	10.1	30.1	32.9	36.2
20	8.26	9.59	10.9	31.4	34.2	37.6
25	11.5	13.1	14.6	37.7	40.6	44.3
30	15.0	16.8	18.5	43.8	47.0	50.9
40	22.2	24.4	26.5	55.8	59.3	63.7
50	29.7	32.4	34.8	67.5	71.4	76.2
60	37.5	40.5	43.2	79.1	83.3	88.4
70	45.4	48.8	51.7	90.5	95.0	100.4
80	53.5	57.2	60.4	101.9	106.6	112.3
90	61.8	65.6	69.1	113.1	118.1	124.1
100	70.1	74.2	77.9	124.3	129.6	135.8

Quantiles of the *F*-Distribution. These quantiles can be obtained in *R* using the `qf(p,df1,df2)` command.

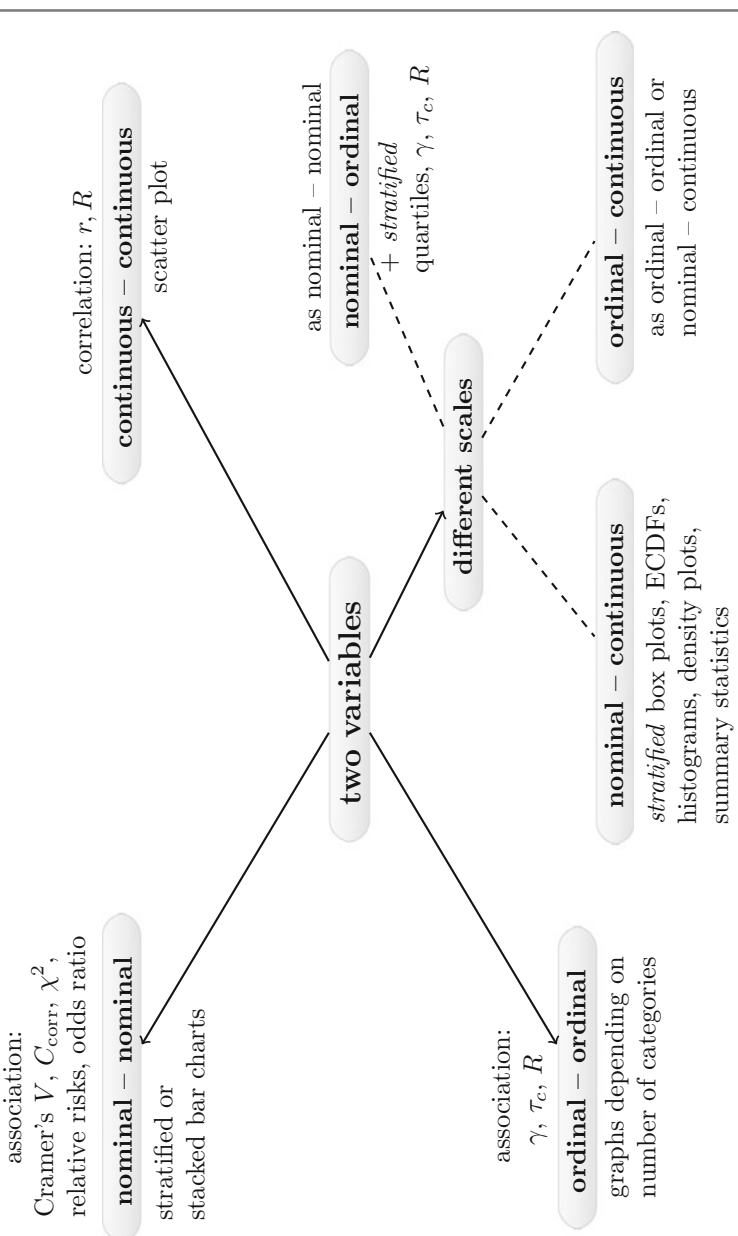
Appendix: Visual Summaries

D

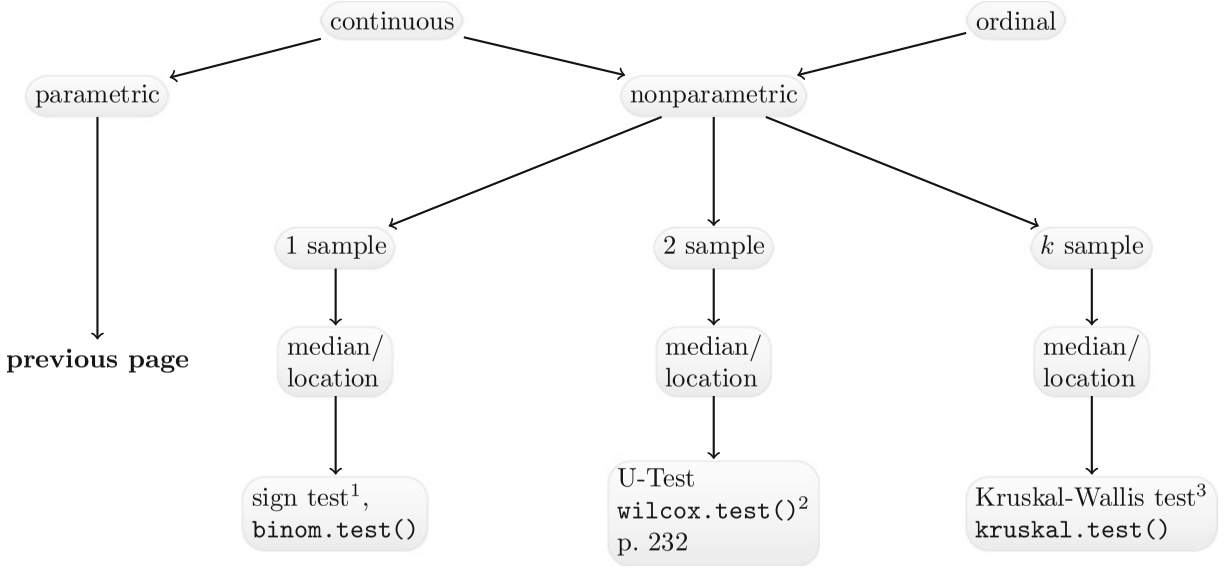
Descriptive Data Analysis



C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5

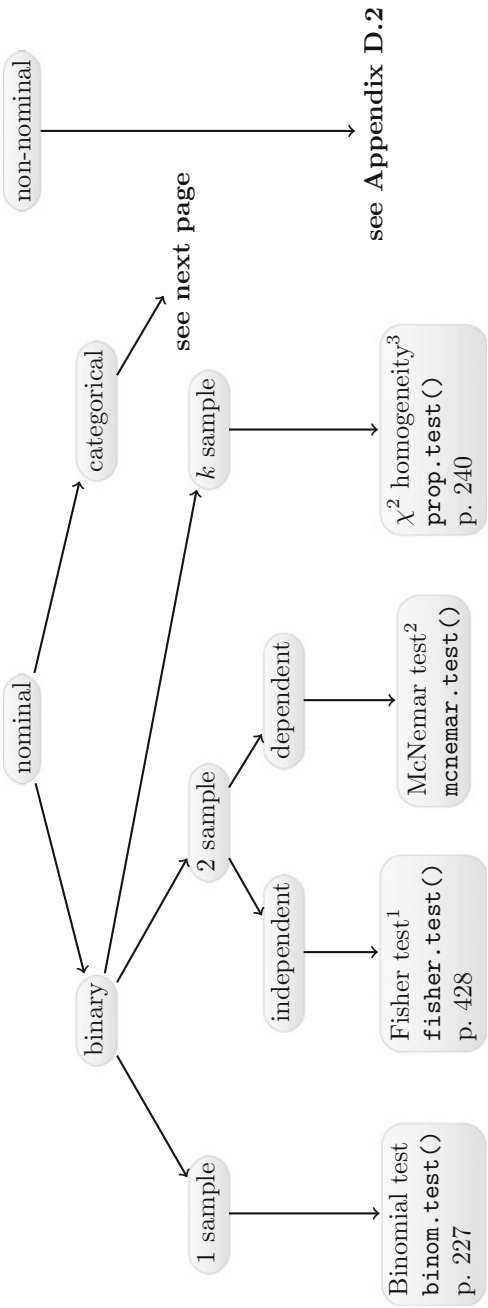

two variables
nominal – nominal
stratified or
stacked bar charts
association:
Cramer's V, C_{corr}, χ^2,
relative risks, odds ratio
continuous – continuous
scatter plot
correlation: r, R
different scales
nominal – ordinal
+ *stratified*
quartiles, γ, τ_c, R
as nominal – nominal
ordinal – ordinal
graphs depending on
number of categories
association:
γ, τ_c, R
nominal – continuous
stratified box plots, ECDFs,
histograms, density plots,
summary statistics
ordinal – continuous
as ordinal – ordinal or
nominal – continuous

Summary of Tests for Continuous and Ordinal Variables



[1] not explained in this book; alternative: Mood's median test
[2] use option `paired=TRUE` for dependent data
[3] not explained in this book; use Friedman test for dependent data

Summary of Tests for Nominal Variables



[1] alternative: χ^2-independence test (`chisq.test()`)
(test decision equivalent to χ^2-homogeneity test)
[2] not explained in this book
[3] test decision equivalent to χ^2-independence test

non-nominal
see Appendix D.2
binary
see previous page
nominal
k sample
χ²-independence[1]
chisq.test()
p. 238
categorical
2 sample
χ²-independence[1]
chisq.test()
p. 238
1 sample
χ² GoF test
chisq.test()
p. 235
[1] test decision equivalent to χ²-homogeneity test
(prop.test)

References

Adler, J. (2012). *R in a Nutshell*. Boston: O'Reilly.

Albert, J., & Rizzo, M. (2012). *R by Example*. New York: Springer.

Bock, J. (1997). *Bestimmung des Stichprobenumfangs*. Munich: Oldenbourg Verlag. (in German).

Casella, G., & Berger, R. (2002). *Statistical inference*. Boston, MA: Cengage Learning.

Chow, S., Wang, H., & Shao, J. (2007). *Sample size calculations in clinical research*. London: Chapman and Hall.

Crawley, M. (2013). *The R book*. London: Wiley.

Dalgaard, P. (2008). *Introductory statistics with R*. Berlin: Springer.

Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. New York: Springer.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2009). *Survey methodology. Wiley series in survey methodology*. Hoboken, NJ: Wiley.

Hernan, M., & Robins, J. (2017). *Causal inference*. Boca Raton: Chapman and Hall/CRC.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, *50*, 361–365.

Kauermann, G., & Küchenhoff, H. (2011). *Stichproben - Methoden und praktische Umsetzung in R*. Heidelberg: Springer. (in German).

Ligges, U. (2008). *Programmieren in R*. Heidelberg: Springer. (in German).

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Young, G., & Smith, R. (2005). *Essentials of statistical inference*. Cambridge: Cambridge University Press.

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5

Index

A
Absolute
  deviation, 51
  mean deviation, 51
  median deviation, 51
Additivity theorem, 116, 147
Akaike's information criterion (AIC), 282
Analysis of variance, 274
ANOVA, 274
Arithmetic mean, 38
  properties, 40
  weighted, 38
Association, 67, 249

B
Backward selection, 282
Bar chart, 24
Behrens-Fisher problem, 222
Bernoulli distribution, 156
Bias, 184
Binomial
  coefficient, 102
  distribution, 157
Bivariate random variables, 140
Box plot, 56

C
Calculation rules
  CDF, 133
  expectation, 144
  normal random variables, 168
  probabilities, 117
  variance, 144
Causation, 288
CDF, 129
  calculation rules, 133
  joint, 141
  quantile, 137
  quartile, 137
Central limit theorem, 426
Central tendency, 38
Certain event, 110
χ^2
  distribution, 171, 427
  goodness-of-fit test, 235
  independence test, 238
  test of homogeneity, 240, 241
  variance test, 430
Coding
  dummy, 266
  effect, 267
Coefficient
  binomial, 102
  of variation, 55
  regression, 251
Combinations, 102
  with order, 103, 104
  with replacement, 103, 104
  without order, 102, 103
  without replacement, 102, 103
Combinatorics, 97
Complementary event, 110
Composite event, 110
Conditional
  distribution, 141, 143
  frequency distribution, 70
  probability, 117
  relative frequency distribution, 70
Confidence
  bound, 196
  interval, 196, 197
  interval for μ; σ^2 known, 197
  interval for μ; σ^2 unknown, 198

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5

interval for p, 199
interval for the odds ratio, 201
level, 196
Consistency, 189
Contingency
coefficient, 77
table, 68, 140
Continuous variable, 6
Convergence
stochastic, 425
Correlation coefficient, 148
of Bravais–Pearson, 82
of Spearman, 84
product moment, 82
Covariance, 146, 147
Covariate, 251
Cramer's V, 77
Cross tabulation, *see* contingency table
Cumulative
distribution function, 129
frequency, *see* frequency, cumulative
marginal distribution, 143

D
Data
matrix, 3, 9
observation, 3
set, 3, 9
transformation, 11
unit, 3
Decomposition
complete, 113
Degenerate distribution, 156
Density, 129
Design matrix, 263, 270
Dispersion, 48
absolute deviation, 51
absolute mean deviation, 51
absolute median deviation, 51
mean squared error, 51
measure, 49
range, 49
standard deviation, 51
Distribution, 19
Bernoulli, 156
Binomial, 157
χ^2, 171, 427
conditional, 141, 143
conditional frequency, 70
conditional relative frequency, 70
continuous, 165
cumulative marginal, 143
degenerate, 156
exponential, 170
F, 427
Gauss, 166
geometric, 163
hypergeometric, 163
independent and identical, 145, 426
joint relative frequency, 70
marginal, 140, 142
marginal frequency, 70
marginal relative frequency, 70
multinomial, 161
normal, 166
Poisson, 160
standard, 154
Student, 172
t, 172, 427
uniform discrete, 154
Duality, 216
Dummy variable, 266

E
Efficiency, 185
Elementary event, 110
Empirical cumulative distribution function (ECDF), 19
Epitools, *see* R packages
Error
type I, 213
type II, 213
Estimation
interval, 195
least squares, 252, 253, 264
maximum likelihood, 192
method of moments, 195
nonparametric, 182
parametric, 182
Event, 110
additive theorem, 116
certain, 110
composite, 110
disjoint, 112
elementary, 110
impossible, 110
simple, 110
sure, 110
theorem of additivity, 115, 116
Expectation, 134
calculation rules, 144
Expected frequencies, 74

Experiment
  Laplace, 114
  random, 109
Exponential distribution, 170

F
Factorial function, *see* function, factorial
F-distribution, 173, 427
Fisher
  exact test, 230
Foreign, *see* R packages
Frequency
  absolute, 18, 69, 113
  cumulative, 19
  expected, 74
  relative, 18, 21, 113
  table, 19, 69
F-Test, 272, 431
Function
  cumulative distribution, *see* CDF
  empirical cumulative distribution, *see* ECDF
  factorial, 100
  joint cumulative distribution, 141
  joint probability distribution, 140, 141
  probability mass, *see* PMF
  step, 133

G
Gamma
  of Goodman and Kruskal, 87
Gauss test
  one-sample, 216
  two-sample, 221
Generalized method of moments, 195
Geometric distribution, 163
Ggplot2, *see* R packages
Gini coefficient, 60
  standardized, 61
Goodman and Kruskal's γ, 87
Goodness of fit
  adjusted measure, 281
  measure, 258
  test, 235
Graph
  bar chart, 24
  box plot, 56
  histogram, 27
  kernel density plot, 29
  Lorenz curve, 58
  pie chart, 26
  QQ-plot, 44
  random plot, 286
  scatter plot, 80
Growth
  factor, 46
  rate, 46

H
Heteroscedasticity, 286
Histogram, 27
Homoscedasticity, 286
Hypergeometric distribution, 163
Hypothesis, 210
  alternative, 211
  linear, 280
  null, 211
  one-sided, 211
  two-sided, 211

I
i.i.d., 145, 153, 426
Impossible event, 110
Independence, 73, 121
  pairwise, 122
  random variables, 143
  stochastic, 121, 144
Ineq, *see* R packages
Inequality
  Tschebyschev, 139
Inference, 181, 182
  least squares, 252, 264
  maximum likelihood, 192
  method of moments, 195
Interaction, 276
Intercept, 251
Interquartile range, 49
Interval estimation, 195

J
Joint
  cumulative distribution function, 141
  frequency distribution, 70
  probability distribution function, 140, 141
  relative frequency distribution, 70

K
Kernel density plot, 29
Kolmogorov–Smirnov test, 237

L
Laplace
  experiment, 114
  probability, 114
Lattice, *see* R packages
Least squares, 252
Life time, 170
Likelihood, 192
Line of equality, 59
Linear
  hypotheses, 280
Linear model, 251
  residuals, 270
Linear regression, 249
  interaction, 276
Location parameter, 38
Log-linear model, 269
Lorenz curve, 58

M
Mann-Whitney U-test, 232
Marginal
  distribution, 140, 142
  frequency distribution, 70
  relative frequency distribution, 70
MASS, *see* R packages
Matrix
  covariance, 147
  design, 263
Maximum likelihood estimation (MLE), 192
Mean
  arithmetic, 38
  properties, 40
  weighted arithmetic, 38
Mean squared error (MSE), 51, 184
Measure
  dispersion, 49
  symmetric, 76
Measure of association
  χ^2 coefficient, 76
  contingency coefficient C, 77
  correlation coefficient, 82
  Cramer's V, 77
  odds ratio, 78
  rank correlation coefficient, 84
  relative risk, 78
Memorylessness, 170
Method of moments, 195
Model
  fitted regression model, 253
  fitted value, 253
  linear, 251
  log-linear, 269
  nonlinear, 251
Multinomial distribution, 161
Multiple linear regression, 262
Multiplication theorem of probability, 119
Multivariate, 249
Mvtnorm, *see* R packages

N
Namibia, 206
Newton–Raphson, 193
Nominal variable, 6
Normal distribution, 166

O
Observation, 3
Odds ratio, 78
One-sample problem, 209, 210
Ordered
  set, 99
  values, 20
Ordinal variable, 6
Outcome, 251

P
Parameter
  location, 38
  regression, 251
  space, 184
PDF, 129
  joint, 140, 141
Percentile, 43
Permutation, 101
  with replacement, 101
  without replacement, 101
Pie chart, 26
Plot
  kernel density, 29
  QQ, 44
  random, 286
  scatter, 79
  trumpet, 286, 288
Poisson distribution, 160
Polynomial regression, 267
Population, 4
Power, 213
Probability
  calculation rules, 117
  conditional, 117, 119

density function, 129
Laplace, 114
mass function, 132
posterior, 120
prior, 120
Probability theory
axioms, 115
p-value, 215

Q
QQ-plot, 44
Quantile, 42, 137
Quartile, 43, 137
Quintile, 43

R
R^2
adjusted, 281
Random variables, 127
bivariate, 140
continuous, 129
discrete, 131
i.i.d, 153
independence, 144
standardization, 138
Range, 49
Real stories, 63, 124, 176, 244, 245
Realization, 128
Reference category, 266
Regression
line, 251
linear, 249
multiple linear, 262
polynomial, 267
Regressor, 251
Relationship, 249
Relative risk, 78
Residuals, 253, 270
standardized, 285
Response, 251
R packages
compositions, 219
epitools, 202
foreign, 12
ggplot2, 35, 73, 199
ineq, 59, 62
lattice, 73
MASS, 28, 283
mvtnorm, 316
ryouready, 87, 88
TeachingDemos, 445
vcd, 77

S
Sample
estimate, 182
pooled variance, 222
space, 110
variance, 51
Sampling
without replacement, 163
Scale
absolute, 7
continuous, 6
interval, 7
nominal, 6
ordinal, 6
ratio, 7
Scatter plot, 80
Set
ordered, 100
unordered, 100
Significance level, 213
Simple event, 110
Slope, 251
Standard deviation, 51, 136
Standard error, 189, 205
Standardization, 54, 138
Standard normal distribution, 166
Statistic, 146, 183
Step function, 20, 133
Stochastic convergence, 425
Stuart's τ_c, 87
Sufficiency, 190
Sure event, 110

T
Table
contingency, 68, 69, 140
frequency, 19, 69
τ_c
of Stuart, 87
T-distribution, 172, 427
Test
ANOVA, 274
Binomial, 227
χ^2 for variance, 232, 430
χ^2 goodness of fit, 235
χ^2 independence, 238
χ^2 of homogeneity, 240
duality, 216
equivalence, 213

F, 272, 431
Fisher, 230, 428
Friedman, 446
Kolmogorov–Smirnov, 237
Kruskal–Wallis, 446
Mann–Whitney U, 232
McNemar, 447
Mood, 446
one-sample Gauss test, 216
one-sample t-test, 219
one-sided, 211
overall F, 272
paired t-test, 225
sign, 446
significance, 213
two-sample binomial, 230
two-sample Gauss test, 221
two-sample t-test, 222
two-sided, 211
U, 232
Welch, 222
Wilcoxon rank sum, 232
Wilcoxon–Mann–Whitney, 232
Test distributions, 317
Theorem
additivity, 116, 147
additivity of χ^2 variables, 172
additivity of disjoint events, 115
Bayes, 120
central limit, 426
i.i.d., 185
large numbers, 425, 426
law of total probability, 119
multiplication for probabilities, 119
Neyman–Fisher Factorization, 191
PDF, 129
standardization, 138
Student, 172
Tschebyschev, 139
variance decomposition, 136
Tie, 85, 86
Transformation, 11
T-test
one-sample, 219
paired, 225
two-sample, 222
Two-sample problem, 210

U
Uniform distribution
continuous, 165
discrete, 154
Unit, 3
Unordered set, 100
U test, 232

V
Variable, 4
binary, 7
bivariate random, 140
categorical, 7
continuous, 6
dependent, 251
discrete, 6
dummy, 266
grouped, 7
independent, 251
nominal, 6
ordinal, 6
random, 127
response, 251
standardized, 54
Variance, 51, 135
additivity theorem, 147
between classes, 53
calculation rules, 144
decomposition, 257
dispersion, 48
pooled, 222
within classes, 53
Vcd, *see* R packages

W
Welch test, 222
Whiskers, 56
Wilcoxon–Mann–Whitney test, 232