1

The R-Package, Sampling Procedures, and Random Variables

1.1 Introduction

In this chapter we give an overview of the software package R and introduce basic knowledge about random variables and sampling procedures.

1.2 The Statistical Software Package R

In practical investigations, professional statistical software is used to design experiments or to analyse data already collected. We apply here the software package R. Anybody can extend the functionality of R without any restrictions using free software tools; moreover, it is also possible to implement special statistical methods as well as certain procedures of C and FORTRAN. Such tools are offered on the internet in standardised archives. The most popular archive is probably CRAN (Comprehensive R Archive Network), a server net that is supervised by the R Development Core Team. This net also offers the package OPDOE (optimal design of experiments), which was thoroughly described in Rasch *et al*. (2011). Further it offers the following packages used in this book: `car, lme4, DunnettTests, VCA, lmerTest, mvtnorm, seqtest, faraway, MASS, glm2, geoR, gstat`.

Apart from only a few exceptions, R contains implementations for all statistical methods concerning analysis, evaluation, and planning. We refer for details to Crawley (2013).

The software package R is available free of charge from http://cran.r-project.org for the operating systems Linux, MacOS X, and Windows. The installation under Microsoft Windows takes place via 'Windows'. Choosing 'base' the installation platform is reached. Using 'Download R 2.X.X for Windows' (X stands for the required version number) the setup file can be downloaded. After this file is started the setup assistant runs through the installation steps. In this book, all standard settings are adopted. The interested reader will find more information about R at http://www.r-project.org or in Crawley (2013).

After starting R the input window will be opened, presenting the red coloured input request: '>'. Here commands can be written up and carried out by pressing the enter button. The output is given directly below the command line. However, the user can also realise line changes as well as line indents for increasing clarity. Not all this influences the functional procedure. A command to read for instance data $y = (1, 3, 8, 11)$ is as follows:

Applied Statistics: Theory and Problem Solutions with R, First Edition.
Dieter Rasch, Rob Verdooren, and Jürgen Pilz.

```
> y <- c(1,3,8,11)
```

The assignment operator in R is the two-character sequence '<-' or '='.

The Workspace is a special working environment in R. There, certain objects can be stored that were obtained during the current work with R. Such objects contain the results of computations and data sets. A Workspace is loaded using the menu

```
File – Load Workspace...
```

In this book the R-commands start with >. Readers who like to use R-commands must only type or copy the text after > into the R-window.

An advantage of R is that, as with other statistical packages like SAS and IBM-SPSS, we no longer need an appendix with tables in statistical books. Often tables of the density or distribution function of the standard normal distribution appear in such appendices. However, the values can be easily calculated using R.

The notation of this and the following chapters is just that of Rasch and Schott (2018).

Problem 1.1 Calculate the value $\varphi(z)$ of the density function of the standard normal distribution for a given value z.

Solution

Use the command > `dnorm(z, mean = 0, sd = 1)`. If the `mean` or `sd` is not specified they assume the default values of 0 and 1, respectively. Hence > `dnorm(z)` can be used in Problem 1.1.

Example

We calculate the value $\varphi(1)$ of the density function of the standard normal distribution using

```
> dnorm(1)
[1] 0.2419707
```

Problem 1.2 Calculate the value $\Phi(z)$ of the distribution function of the standard normal distribution for a given value z.

Solution

Use the command > `pnorm(z, mean = 0, sd = 1)`.

Example

We calculate the value $\Phi(1)$ of the distribution function of the standard normal distribution by > `pnorm(1, mean = 0, sd = 1)` or using the default values using > `pnorm(1)`.

```
> pnorm(1)
[1] 0.8413447
```

Also, for other continuous distributions, we obtain using `d` with the R-name of a distribution, the value of the density function and, using `p` with the R-name of a distribution, the value of the distribution function. We demonstrate this in the next problem for the lognormal distribution.

Problem 1.3 Calculate the value of the density function of the lognormal distribution whose logarithm has mean equal to `meanlog = 0` and standard deviation equal to `sdlog = 1` for a given value *z*.

Solution
Use the command > `dlnorm(z, meanlog = 0, sdlog = 1)` or use the default values `meanlog = 0` and `sdlog = 1` using > `dlnorm(z).`

Example
We calculate the value of the density function of the lognormal distribution with `meanlog = 0` and `sdlog = 1` using

```
> dlnorm(1)
[1] 0.3989423
```

Problem 1.4 Calculate the value of the distribution function of the lognormal distribution whose logarithm has mean equal to `meanlog = 0` and standard deviation equal to `sdlog = 1` for a given value *z*.

Solution
Use the command > `plnorm(z, meanlog = 0, sdlog = 1)` or use the default values `meanlog = 0` and `sdlog = 1` using > `plnorm(z).`

Example
We calculate the value of the distribution function for z = 1 of the lognormal distribution with `meanlog = 0` and `sdlog = 1` using

```
> plnorm(1)
[1] 0.5
```

From most of the other distributions we need the quantiles (or percentiles) $q_P = P(\boldsymbol{y} \le P)$.

This can be done by writing *q* followed by the R-name of the distribution.

Problem 1.5 Calculate the *P*%-quantile of the *t*-distribution with df degrees of freedom and optional non-centrality parameter ncp.

Solution
Use the command > `qt(P,df, ncp)` and for a central *t*-distribution use the default by omitting `ncp`.

Example
Calculate the 95%-quantile of the central *t*-distribution with 10 degrees of freedom.

```
> qt(0.95,10)
[1] 1.812461
```

We demonstrate the procedure for the chi-square and the *F*-distribution.

Problem 1.6 Calculate the *P*%-quantile of the χ^2-distribution with df degrees of freedom and optional non-centrality parameter ncp.

Solution
Use the command `> qchisq(P,df, ncp)` and for the central χ^2-distribution with df degrees of freedom use `> qchisq(P,df)`.

Example
Calculate the 95%-quantile of the central χ^2-distribution with 10 degrees of freedom.

```
> qchisq(0.95,10)
[1] 18.30704
```

Problem 1.7 Calculate the *P%*-quantile of the *F*-distribution with df1 and df2 degrees of freedom and optional non-centrality parameter ncp.

Solution
Use the command `> qf(P,df1,df2, ncp)`, and for the central *F*-distribution with df1 and df2 degrees of freedom use `> qf(P,df1,df2)`.

Example
Calculate the 95%-quantile of the *central F*-distribution with 10 and 20 degrees of freedom!

```
> qf(0.95,10,20)
[1] 2.347878
```

For the calculation of further values of probability functions of discrete random variables or of distribution functions and quantiles the commands can be found by using the help function in the tool bar of R, and then you may call up the 'manual' or use Crawley (2013).

1.3 Sampling Procedures and Random Variables

Even if we, in this book, we mainly discuss how to plan experiments and to analyse observed data, we still need basic knowledge about random variables because, without this, we could not explain unbiased estimators or the expected length of a confidence interval or how to define the risks of a statistical tests.

Definition 1.1 A sampling procedure without replacement (wor) or with replacement (wr) is a rule of selecting a proper subset, named sample, from a well-defined finite basic set of objects (population, universe). It is said to be at random if each element of the basic set has the same probability p to be drawn into the sample. We also can say that in a random sampling procedure each possible sample has the same probability to be drawn.

A (concrete) sample is the result of a sampling procedure. Samples resulting from a random sampling procedure are said to be (concrete) random samples or shortly samples.

If we consider all possible samples from a given finite universe, then, from this definition, it follows that each possible sample has the same probability to be drawn.

There are several random sampling procedures that can be used in practice. Basic sets of objects are mostly called (statistical) populations or, synonymously, (statistical) universes.

Concerning random sampling procedures, we distinguish (among other cases):

- Simple (or pure) random sampling with replacement (wr) where each of the N elements of the population is selected with probability $\frac{1}{N}$.
- Simple random sampling without replacement (wor) where each unordered sample of n different objects has the same probability to be chosen.
- In cluster sampling, the population is divided into disjoint subclasses (clusters). Random sampling without replacement is done among these clusters. In the selected clusters, all objects are taken into the sample. This kind of selection is often used in area sampling. It is only random corresponding to Definition 1.1 if the clusters contain the same number of objects.
- In multi-stage sampling, sampling is done in several steps. We restrict ourselves to two stages of sampling where the population is decomposed into disjoint subsets (primary units). Part of the primary units is sampled randomly without replacement (wor) and within them pure random sampling without replacement (wor) is done with the secondary units. A multi-stage sampling is favourable if the population has a hierarchical structure (e.g. country, province, towns in the province). It is at random corresponding to Definition 1.1 if the primary units contain the same number of secondary units.
- Sequential sampling, where the sample size is not fixed at the beginning of the sampling procedure. At first, a small sample with replacement is taken and analysed. Then it is decided whether the obtained information is sufficient, e.g. to reject or to accept a given hypothesis (see Chapter 3), or if more information is needed by selecting a further unit.

When a cluster or in two-stage sampling the clusters or primary units have different sizes (number of elements or areas), more sophisticated methods are used (Rasch et al. 2008, Methods 1/31/2110, 1/31/3100).

Both a random sampling (procedure) and arbitrary sampling (procedure) can result in the same concrete sample. Hence, we cannot prove by inspecting the concrete sample itself whether or not the sample is randomly chosen. We have to check the sampling procedure used instead.

In mathematical statistics random sampling with a replacement procedure is modelled by a vector $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n)^T$ of random variables $\boldsymbol{y}_i$, $i = 1, \dots, n$, which are independently distributed as a random variable $\boldsymbol{y}$, i.e. they all have the same distribution. The $\boldsymbol{y}_i$, $i = 1, \dots, n$ are said independently and identically distributed (i.i.d.). This leads to the following definition.

Definition 1.2 A random sample of size n is a vector $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n)^T$ with n i.i.d. random variables $\boldsymbol{y}_i$, $i = 1, \dots, n$ as elements.

Random variables given in bold print (see Appendix A for motivation).

The vector $Y = (y_1, y_2, \dots, y_n)^T$ is called a realisation of $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n)^T$ and is used as a model of a vector of observed values or values selected by a random selection procedure.

To explain this approach let us assume that we have a universe of 100 elements (the numbers 1–100). We like to draw a pure random sample without replacement (wor) of size $n = 10$ from this universe and model this by $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots , \boldsymbol{y}_{10})^T$. When a random sample has been drawn it could be the vector $Y = (y_1, y_2, \dots , y_{10})^T = (3, 98, 12, 37, 2, 67, 33, 21, 9, 56)^T = (2, 3, 9, 12, 21, 33, 37, 56, 67, 98)^T$. This means that it is only important which element has been selected and not at which place this has happened. All samples wor occur with probability $\frac{1}{\binom{100}{10}}$. The denominator $\binom{100}{10}$ can be calculated by R with the

`> choose()` command

```
> choose(100,10)
[1] 1.731031e+13
```

and from this the probability is $\frac{1}{1731031 \times 10^7}$.

We can now write

$$P\{(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots , \boldsymbol{y}_{10})^T = (2, 3, 9, 12, 21, 33, 37, 56, 67, 98)^T\} = \frac{1}{1731031 \times 10^7}.$$

In a probability statement, something must always be random. To write

$$P\{(y_1, y_2, \dots , y_{10})^T = (2, 3, 9, 12, 21, 33, 37, 56, 67, 98)^T\}$$

is nonsense because $(y_1, y_2, \dots , y_{10})^T$ as the vector on the right-hand-side is a vector of special numbers and it is nonsense to ask for the probability that 5 equals 7.

To explain the situation again we consider the problem of throwing a fair dice; this is a dice where we know that each of the numbers 1, …, 6 occurs with the same probability $\frac{1}{6}$. We ask for the probability that an even number is thrown. Because one half of the six numbers are even, this probability is $\frac{1}{2}$. Assume we throw the dice using a dice cup and let the result be hidden, than the probability is still $\frac{1}{2}$. However, if we take the dice cup away, a realisation occurs, let us say a 5. Now, it is stupid to ask, what is the probability that 5 is even or that an even number is even. Probability statements about realisations of random variables are senseless and not allowed. The reader of this book should only look at a probability statement in the form of a formula if something is in bold print; only in such a case is a probability statement possible.

We learn in Chapter 4 what a confidence interval is. It is defined as an interval with at least one random boundary and we can, for example, calculate with some small α the probability $1 - \alpha$ that the expectation of some random variable is covered by this interval. However, when we have realised boundaries, then the interval is fixed and it either covers or does not cover the expectation. In applied statistics, we work with observed data modelled by realised random variables. Then the calculated interval does not allow a probability statement. We do not know, by using R or otherwise, whether the calculated interval covers the expectation or not. Why did we fix this probability before starting the experiment when we cannot use it in interpreting the result?

The answer is not easy, but we will try to give some reasons. If a researcher has to carry out many similar experiments and in each of them calculates for some parameter a $(1 - \alpha)$ confidence interval, then he can say that in about $(1 - \alpha)100\%$ of all cases the interval has covered the parameter, but of course he does not know when this happened.

What should we do when only one experiment has to be done? Then we should choose $(1 - \alpha)$ so large (say 0.95 or 0.99) that we can take the risk of making an erroneous statement by saying that the interval covers the parameter. This is analogous to the situation of a person who has a severe disease and needs an operation in hospital. The person can choose between two hospitals and knows that in hospital A about 99% of people operated on survived a similar operation and in hospital B only about 80%. Of course (without further information) the person chooses A even without knowing whether she/he will survive. As in normal life, also in science; we have to take risks and to make decisions under uncertainty.

We now show how R can easily solve simple problems of sampling.

Problem 1.8 Draw a pure random sample without replacement of size $n < N$ from N given objects represented by numbers 1, …, N without replacing the drawn objects. There are $M = \binom{N}{n}$ possible unordered samples having the same probability $p = \frac{1}{M}$ to be selected.

Solution
Insert in R a data file y with N entries and continue in the next line with `>sample (y,n, replace = FALSE)` or `>sample (y,n, replace = F)` with $n < N$ to create a sample of size $n < N$ different elements from y; when we insert `replace = TRUE` we get random sampling with replacement. The default is `replace = FALSE`, hence for sampling without replacement we can use `>sample (y, n)`.

Example
We choose $N = 9$, and $n = 5$, with population values $y = (1,2,3,4,5,6,7,8,9)$

```
> y <- c(1,2,3,4,5,6,7,8,9)
> sample(y,5)
[1] 7 6 5 1 3
```

A pure random sampling with replacement also occurs if the random sample is obtained by replacing the objects immediately after drawing and each object has the same probability of coming into the sample using this procedure. Hence, the population always has the same number of objects before a new object is taken. This is only possible if the observation of objects works without destroying or changing them (examples are tensile breaking tests, medical examinations of killed animals, felling of trees, harvesting of food).

Problem 1.9 Draw with replacement a pure random sample of size n from N given objects represented by numbers 1, …, N with replacing the drawn objects. There are $M_{\text{rep}} = \binom{N+n-1}{n}$ possible unordered samples having the same probability $\frac{1}{M_{\text{rep}}}$ to be selected.

Solution
Insert in R a data file y with N entries and continue in the next line with `>sample (y, n, replace =TRUE)` or `>sample(y, n, replace=T)` to create a sample of size n not necessarily with different elements from y.

Examples
Example with $n < N$

```
> y<-c(1,2,3,4,5,6,7,8,9)
> sample(y,5,replace=T)
[1] 2 4 6 4 2
```

Example with $n > N$

```
    > y<-c(1,2,3,4,5,6,7,8,9)
    > sample(y,10,replace=T)
[1] 3 9 5 5 9 9 8 7 6 3
```

A method that can sometimes be realised more easily is systematic sampling with a random start. It is applicable if the objects of the finite sampling set are numbered from 1 to N, and the sequence is not related to the character considered. If the quotient $m = N/n$ is a natural number, a value i between 1 and m is chosen at random, and the sample is collected from objects with numbers $i, m + i, 2m + i, \dots, (n - 1)m + i$. Detailed information about this case and the case where the quotient m is not an integer can be found in Rasch et al. (2008, method 1/31/1210).

Problem 1.10 From a set of N objects systematic sampling with a random start should choose a random sample of size n.

Solution
We assume that in the sequence 1, 2, …, N there is no trend. Let assume that $m = \frac{N}{n}$ is an integer and select by pure random sampling a value $1 \le x \le m$ (sample of size 1) from the m numbers 1, …, m. Then the systematic sample with random start contains the numbers $x, x + m, x + 2m, \dots, x + (n - 1)m$.

Example
We choose $N = 500$ and $n = 20$, and the quotient $\frac{500}{20} = 25$ is an integer-valued. Analogous to Problem 1.1 we draw a random sample of size 1 from (1, 2, …, 25) using R.

```
> y<- c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,
  16,17,18,19,20,21,22,23,24,25)
> sample(y,1)
[1] 9
```

The final systematic sample with random start of size $n = 20$ starts with number $x = 9$ and $m = 25$: (9, 34, 59, 84, 109, 134, 159, 184, 209, 234, 259, 284, 309, 334, 359, 384, 409, 434, 459, 484).

Problem 1.11 By cluster sampling, from a population of size N decomposed into s disjoint subpopulations, so-called clusters of sizes $N_1, N_2, \dots, N_s$, a random sample has to be drawn.

Solution
Partial samples of size n_i are collected from the ith stratum ($i = 1, 2, \dots, s$) where pure random sampling procedures without replacement are used in each stratum. This leads to a random sampling without replacement procedure for the whole population if the numbers n_i/n are chosen proportional to the numbers N_i/N. The final random sample contains $n = \sum_{i=1}^{s} n_i$ elements.

Example
Vienna, the capital of Austria, is subdivided into 23 municipalities. We repeat a table with the numbers of inhabitants N_i^* in these municipalities from Rasch et al. (2011) and round the numbers for demonstrating the example to values so that N_i/N is an integer, where $N = 1\,700\,000$.

Now we select by pure random sampling without replacement, as shown in Problem 1.8, from each municipality n_i from the N_i inhabitants to reach a total random sample of 1000 inhabitants from the 1 700 000 people in Vienna.

While for the stratified random sampling objects are selected without replacement from each subset, for two-stage sampling, subsets or objects are selected at random without replacement at each stage, as described below. Let the population consist of s disjoint subsets of size N_0, the primary units, in the two-stage case. Further, we suppose that the character values in the single primary units differ only at random, so that objects need not to be selected from all primary units. If the desired sample size is $n = r\,n_0$ with $r < s$, then in the first step, r of the s given primary units are selected using a pure random sampling procedure. In the second step, n_0 objects (secondary units) are chosen from each selected primary unit, again applying a pure random sampling. The number of possible samples is $\binom{s}{r} \cdot \binom{N_0}{n_0}$, and each object of the population has the same probability $p = \frac{r}{s} \cdot \frac{n_0}{N_0}$ to reach the sample corresponding to Definition 1.1.

Problem 1.12 Draw a random sample of size n in a two-stage procedure by selecting first from the s primary units having sizes N_{i} ($i = 1, \ldots$, s) exactly r units.

Solution
To draw a random sample without replacement of size n we select a divisor r of n and from the s primary units we randomly select r proportional to the relative sizes $\frac{N_i}{N}$ with $N = \sum_{i=1}^{s} N_i$ (i = 1, …, s). From each of the selected r primary units we select by pure random sampling without replacement $\frac{n}{r}$ elements as the total sample of secondary units.

Example
We take again the values of Table 1.1 and select $r = 5$ from the $s = 23$ municipalities to take an overall sample of $n = 1000$. For this we split the interval (0,1] into 23 subintervals $\left(1000\frac{N_{i-1}}{N}, 1000\frac{N_i}{N}\right]$ $i = 1, \ldots, 23$ with $N_0 = 0$ and generate five uniformly distributed random numbers in (0,1]. If a random number multiplied by 1000 falls in any of the 23 sub-intervals (which can be easily found by using the 'cum' column in Table 1.1) the corresponding municipality has to be selected. If a further random number falls into the same interval it is replaced by another uniformly distributed random number. We generate five such random numbers as follows:

```
> runif(5)
 [1] 0.18769112 0.78229430 0.09359499 0.46677904 0.51150546
```

The first number corresponds to Mariahilf, the second to Florisdorf, the third to Landstraße, the fourth to Hietzing, and the last one to Penzing. To obtain a random sample of size 1000 we take pure random samples of size 200 from people in Mariahilf, Florisdorf, Landstraße, Hietzing, and Penzing, respectively.

Table 1.1 Number $N_i^*, i = 1, \ldots, 23$ of inhabitants in 23 municipalities of Vienna.

Municipality	N_i^*	N_i	$n_i = 1000\frac{N_i}{N}$	cum
Innere Stadt	16 958	17 000	10	10
Leopoldstadt	94 595	102 000	60	70
Landstraße	83 737	85 000	50	120
Wieden	30 587	34 000	20	140
Margarethen	52 548	51 000	30	170
Mariahilf	29 371	34 000	20	190
Neubau	30 056	34 000	20	210
Josefstadt	23 912	34 000	20	230
Alsergrund	39 422	34 000	20	250
Favoriten	173 623	170 000	100	350
Simmering	88 102	85 000	50	400
Meidling	87 285	85 000	50	450
Hietzing	51 147	51 000	30	480
Penzing	84 187	85 000	50	530
Rudolfsheim	70 902	68 000	40	570
Ottakring	94 735	102 000	60	630
Hernals	52 701	51 000	30	660
Währing	47 861	51 000	30	690
Döbling	68 277	68 000	40	730
Brigittenau	82 369	85 000	50	780
Floridsdorf	139 729	136 000	80	860
Donaustadt	153 408	153 000	90	950
Liesing	91 759	85 000	50	1 000
Total	N^*=1 687 271	N= 1 700 000	n = 1 000	

Rounded numbers N_i, n_i, and cumulated n_i.
Source: From Statistik Austria (2009) Bevölkerungsstand inclusive Revision seit 1.1. 2002, Wien, Statistik Austria.

References

Crawley, M.J. (2013). *The* R *Book*, 2nd edition, Chichester: Wiley.

Rasch, D. and Schott, D. (2018). *Mathematical Statistics*. Oxford: Wiley.

Rasch, D., Herrendörfer, G., Bock, J., Victor, N., and Guiard, V. (2008). *Verfahrensbibliothek Versuchsplanung und - auswertung*, 2. verbesserte Auflage in einem Band mit CD. R. Oldenbourg Verlag München Wien.

Rasch, D., Pilz, J., Verdooren, R., and Gebhardt, A. (2011). *Optimal Experimental Design with* R. Boca Raton: Chapman and Hall.