

## Data Analytics for Data Scientists

### Design of Experiments (DoE)

#### Suggested solutions for Exercise 05: Sampling

2025

Prof. Dr. Jürg Schwarz

MSc Adrian Bieri

<b>Suggested solution 01</b>	<b>2</b>
Definition of a population	2
<b>Suggested solution 02</b>	<b>3</b>
Sampling error & Variability of sample means	3
<b>Suggested solution 03</b>	<b>5</b>
Simple random sampling with R	5
<b>Suggested solution 04 – Voluntary</b>	<b>6</b>
Calculating a sample size with R	6

# **Suggested solution 01**

## **Definition of a population**

You want to conduct a survey that involves all residents of the city of Zurich.

Discuss the criteria you would choose in order to define the population ...

1. In terms of geographical aspects
2. In terms of temporal aspects
3. In terms of factual / content-related aspects

## **Suggested answers to the questions**

In principle, the elements of the population of interest can be found in a register, for example at the Residents' Registration Office.

For a clear definition of the population, you would need to consider further aspects.

For example:

- How would you define «residents»?  
Does the population consist of only adults or does it include children?
- When does the survey begin – in the preliminary study, in the main study?  
When does it end?
- How are weekend commuter treated?
- How do you treat people who are resident in the city of Zurich but have temporarily deregistered?
- Is there a theoretical reason for surveying only those who have been residents of Zurich for a certain minimum period?
- Should all foreign residents be surveyed including those with a temporary residence permit?
- What happens to those who live in collective households (e.g. a psychiatric clinic)?

## Suggested solution 02

### Sampling error & Variability of sample means

Evaluate the effect of the following sampling restrictions on the type of sampling error ...

1. Non-sampling error
2. Sampling error
3. Variability of sample means

Restriction	Solutions
You are using an incorrect equation to calculate the mean.	2 <sup>1</sup>
You are conducting a survey with a sample of $n = 5$ .	3 <sup>2</sup>
You take an online survey on cosmetic products and ask, among others, about the age of the interviewee at the beginning of the questionnaire.	1, 2 <sup>3</sup>
You conduct a telephone survey for a lifestyle product. 20% of the 20 to 39-year-olds can no longer be reached via landline phone.	1, 2 <sup>4</sup>
The city of Zurich is conducting a study on the topic of sustainability. Attitudes towards sustainability are also surveyed in a neighborhood with high unemployment, poverty, a high proportion of illegal migrants, etc.	1, 2, 3 <sup>5</sup>
A local television station conducts a telephone survey during the evening news, which is only announced in this broadcast.	1, 2, 3 <sup>6</sup>

**General assumption:** The research question refers to the general population.

#### Reason

##### <sup>1</sup>Reason

2. Sampling error  
Use of an inappropriate estimator → See also in script

##### <sup>2</sup>Reason

3. Variability of sample means.  
The smaller a sample, the larger the standard error of the sample mean → See also in script

##### <sup>3</sup>Reason

1. Non-sampling error  
If the population includes all genders, the survey topic "cosmetic products" is likely to lead to systematic nonresponse among some of the individuals contacted. That is, certain individuals do not even open the online questionnaire because of the topic.
2. Sampling error  
Asking for age at the beginning of the questionnaire can lead to certain people abandoning the survey already at the beginning, because they are pressured by this question, especially in the context of cosmetic products (age, beauty, ...)

Therefore, these persons do not have the same selection probability, as other elements of the defined population.

In general, questions about the person (age, gender, income, etc.) should be asked only at the end of a questionnaire.

#### <sup>4</sup>Reason

##### 1. Non-sampling error

The fact that 20% of 20-39 year olds can no longer be reached via the fixed network corresponds to a "classic". → Coverage error

##### 2. Sampling error

The above-mentioned non-sampling error results in 20- to 39-year-olds being underrepresented, conversely means that 40-year-olds and older are overrepresented.

Analogous to the example of the online survey on cosmetic products, the case could arise here that when contacted by telephone, the willingness to participate in a survey on a life-style product is dependent on the age group.

#### <sup>5</sup>Reason

##### 1. Non-sampling error

There is a tendency for residents of this neighborhood to have a higher refusal rate compared to the general population due to their specific circumstances (systematic nonresponse), especially because the topic of sustainability is not relevant to the residents of the neighborhood.

##### 2. Sampling error

Logistical problems in the distribution of questionnaires could result in many questionnaires being lost. People without an address, for example homeless people or illegals without an official residence, do not have the same selection probability as other people in the defined population. → Selection error

##### 3. Variability of sample means.

Residents of this neighborhood are likely to have lower average language skills, lower education levels, and lower interest compared to the general population, which is why larger (random) differences emerge in responses.

#### <sup>6</sup>Reason

##### 1. Non-sampling error

The data produced with the telephone survey only apply to the region, because other people do not watch this local TV station, and therefore cannot be generalized to the whole of Switzerland.

In addition, only a certain group of people in the television station's audience follow the telephone survey.

##### 2. Sampling error

Restricting recruitment to the evening means that not all elements of the viewership have the same probability of selection. For example, younger viewers do not watch television in the evening. → Selection error

##### 3. Variability of sample means

Given items 1. & 2. the sample is likely to become small, which would be associated with a large standard error.

## Suggested solution 03

### Simple random sampling with R

Given is an address list *address.csv*

Use R to draw a simple random sample with  $n = 50$  elements from this data set.

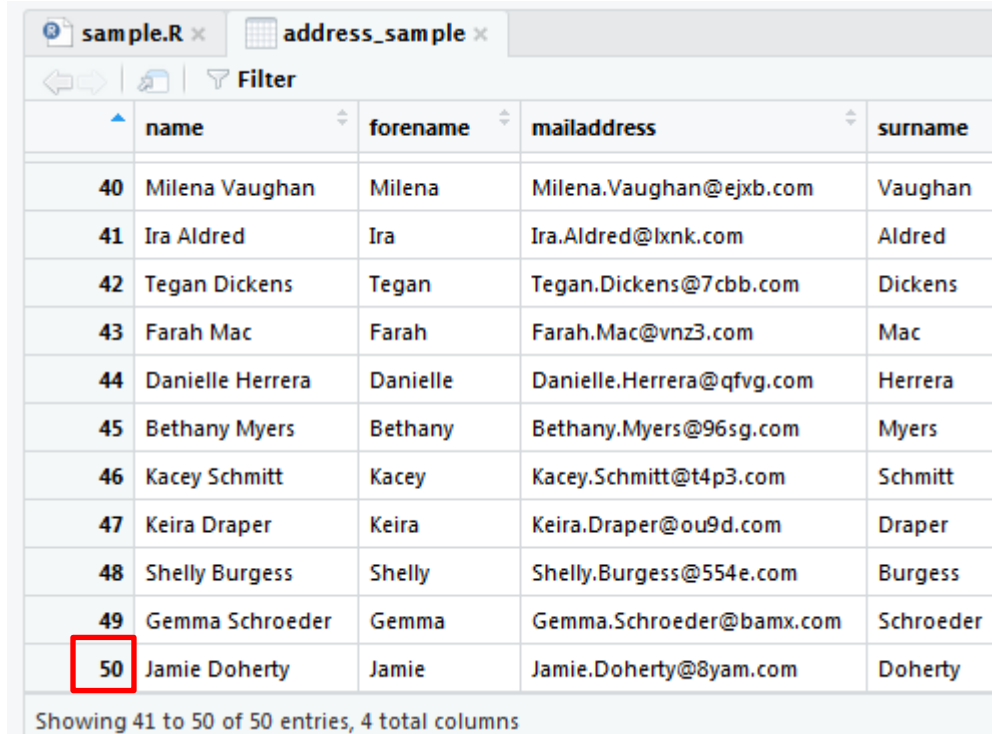
Insert also the R-code, the R-output and if necessary, R-plots in your answer.

### Suggested answers to the questions

R-Code → Collection of R-Code.r

```
library(readr)
address <- read_csv("address.csv")
View(address)

address_sample <- address[sample(599, 50, replace=FALSE),]
View(address_sample)
```



	name	forename	mailaddress	surname
40	Milena Vaughan	Milena	Milena.Vaughan@ejxb.com	Vaughan
41	Ira Aldred	Ira	Ira.Aldred@lxnk.com	Aldred
42	Tegan Dickens	Tegan	Tegan.Dickens@7cbb.com	Dickens
43	Farah Mac	Farah	Farah.Mac@vnz3.com	Mac
44	Danielle Herrera	Danielle	Danielle.Herrera@qfvg.com	Herrera
45	Bethany Myers	Bethany	Bethany.Myers@96sg.com	Myers
46	Kacey Schmitt	Kacey	Kacey.Schmitt@t4p3.com	Schmitt
47	Keira Draper	Keira	Keira.Draper@ou9d.com	Draper
48	Shelly Burgess	Shelly	Shelly.Burgess@554e.com	Burgess
49	Gemma Schroeder	Gemma	Gemma.Schroeder@bamx.com	Schroeder
50	Jamie Doherty	Jamie	Jamie.Doherty@8yam.com	Doherty

Showing 41 to 50 of 50 entries, 4 total columns

## Suggested solution 04 – Voluntary

### Calculating a sample size with R

Given is the research question:

How many German households use a Smart-TV?

Use R to describe the results – insert also the R-code, the R-output and if necessary, R-plots.

- 1) How large a sample do you need in order to answer the research question, in this case:

The error probability is 5% → **z = 1.96**

The error range is 3% → **e = 0.03**

No information is given about the proportion of German households with a Smart-TV

- 2) How large a sample do you need in order to answer the research question, in this case:

The error probability is 5% → **z = 1.96**

The error range is 3% → **e = 0.03**

From another, valid survey it is known that the proportion of German households with Smart-TVs is **35%**

### Suggested answers to the questions

R file *sample\_size.r*

Prerequisite: Germany → approx. 40,000,000 households<sup>1</sup>

```
# install.packages("sampler")
library(sampler)

rsampcalc(N=40000000, e=3, ci=95, p=0.5)
rsampcalc(N=40000000, e=3, ci=95, p=0.35)
rsampcalc(N=40000000, e=3, ci=95, p=0.65)
```

```
> rsampcalc(N=40000000, e=3, ci=95, p=0.5)
[1] 1068
> rsampcalc(N=40000000, e=3, ci=95, p=0.35)
[1] 972
> rsampcalc(N=40000000, e=3, ci=95, p=0.65)
[1] 972
```

General remark: Cochran's equation is symmetrical in terms of p

$$\text{Sample size} = \frac{z^2 \cdot p \cdot (1-p)}{e^2} \rightarrow \frac{1.96^2 \cdot 0.35 \cdot (1-0.35)}{0.03^2} = \frac{1.96^2 \cdot 0.65 \cdot (1-0.65)}{0.03^2} = 972$$

---

<sup>1</sup> Rule of thumb: In Europe, the number of households is roughly equal to half the population.