

Data Analytics for Data Scientists

Design of Experiments (DoE)

Lecture 05: Sampling

2025

Prof. Dr. Jürg Schwarz

Program: 16:15 until 17:55

16:15	Begin of the lesson
	Lecture: Jürg Schwarz <ul style="list-style-type: none">◦ Definition of population◦ Sampling procedure◦ Sample size → Appendix◦ Quality & Sources of error◦ Preview of Lecture 06
	Tutorial: Students / Jürg Schwarz / Assistants <ul style="list-style-type: none">◦ Working on the exercise<ul style="list-style-type: none">◦ Support by Jürg Schwarz / Assistants
17:55	End of the lesson

Population

Set of all (potentially explorable) elements that have a common characteristic or a common combination of characteristics

The observation units (individuals, households, etc.) must be defined before the research start.

In order to define the population, clear differentiation criteria must be formulated, so that for each observation unit it can be determined whether it is part of the population.

1. Differentiation in **geographical** aspects
Generally unproblematic
2. Differentiation of **temporal** aspects (point in time or period)
May cause problems during the implementation (sampling plan, etc.)
3. Differentiation in **factual / content-related** aspects
Often leads to definition problems

Example – Students at the University of Zurich

A survey is to be conducted among students at the University of Zurich.

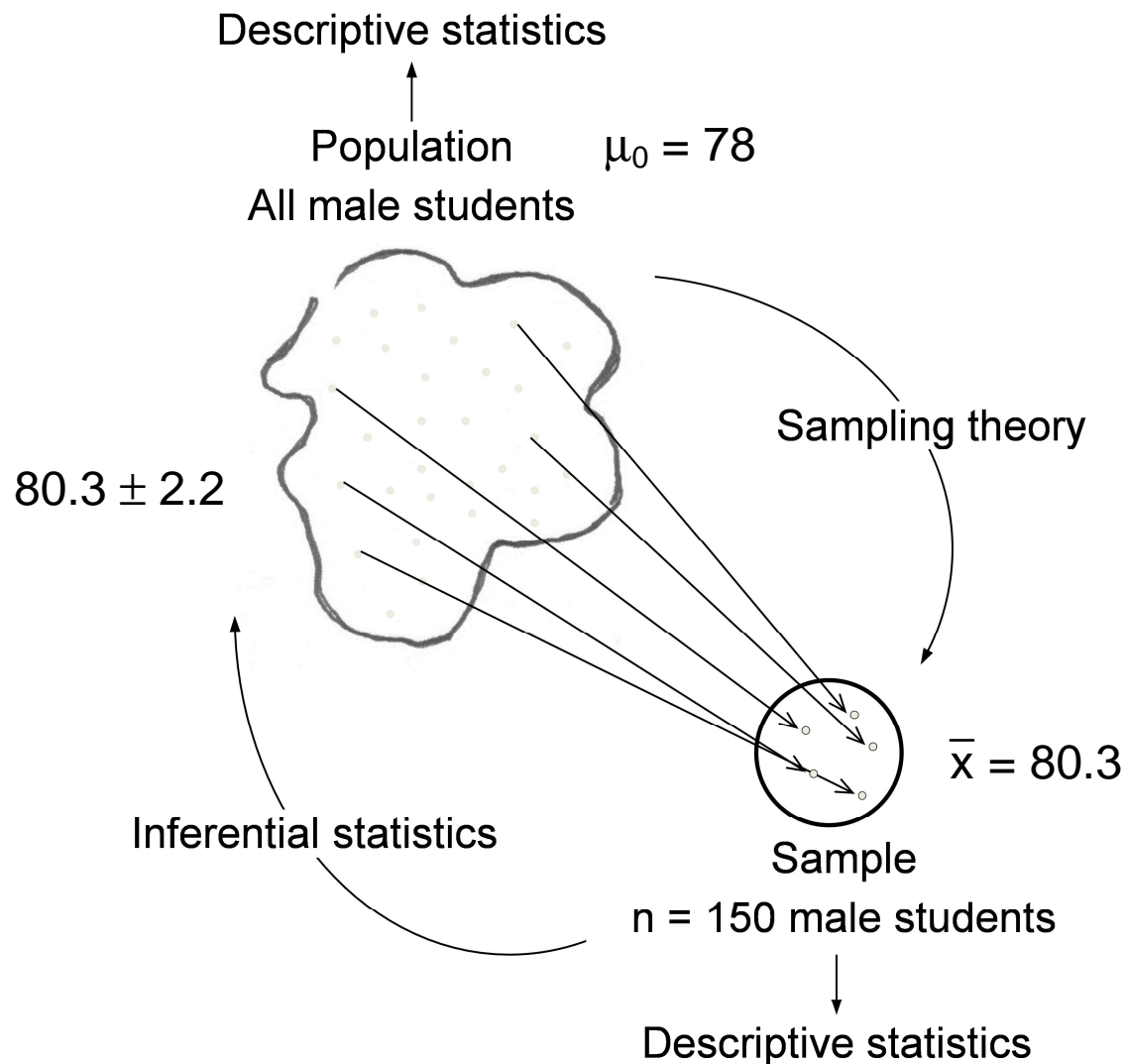
Population:

Persons at the University of Zurich (**1.**), in the spring semester (**2.**), currently enrolled (**3.**)

Further aspects have to be taken into account or questions have to be clarified, such as:

- A Should all persons with a valid enrollment be considered or should a specific group be selected (e.g. only newly enrolled persons)?
- B Are persons included who are enrolled but are not actively studying at that time (e.g. stay abroad, illness, internship)?
- C Are doctoral students part of the population?
- D Should foreign students be considered who study at the University of Zurich for one semester as part of an exchange program?

Generalizing from a sample to the population



$\bar{x} = 80.3$ Mean of the sample
 80.3 ± 2.2 Inference on the population

Research question

What is the weight of Swiss students (men)?

Sampling theory

Drawing a random sample from a population.

Inferential statistics

Estimating the value μ_0 in the population, based on the mean of the sample \bar{x} .

Every estimate has a confidence interval.

Point estimate: 80.3

Confidence interval: ± 2.2

Each statement must be evaluated while considering the risk of making a false statement.

Point estimate of the mean

- \bar{x} Mean of the sample
- μ_0 True mean in the population (generally unknown)

An "estimator" is a function that calculates a value.

$E(\dots)$ calculates an expected value.

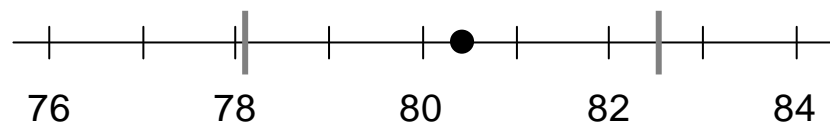
Can \bar{x} be used for an estimation of μ_0 ? Does $\mu_0 = E(\bar{x})$ apply?

The **mean value \bar{x}** of a sample is an unbiased, efficient and consistent estimator of the true mean value in the population: $\mu_0 = E(\bar{x})$

Confidence interval

```
> CI(weight$weight, ci=.95)
      upper      mean      lower
82.49780 80.33784 78.17789
```

95% confidence interval: **[78.2, 82.5]**

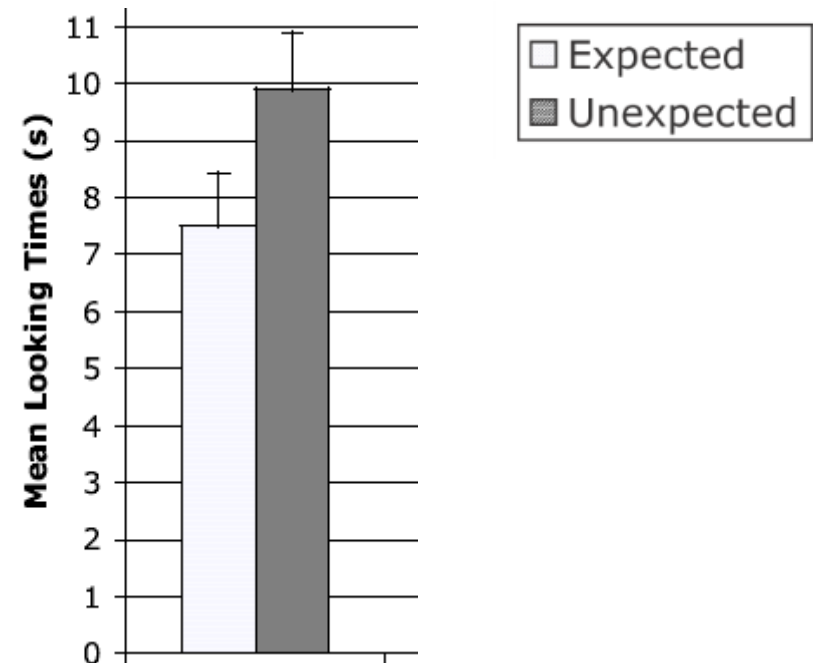
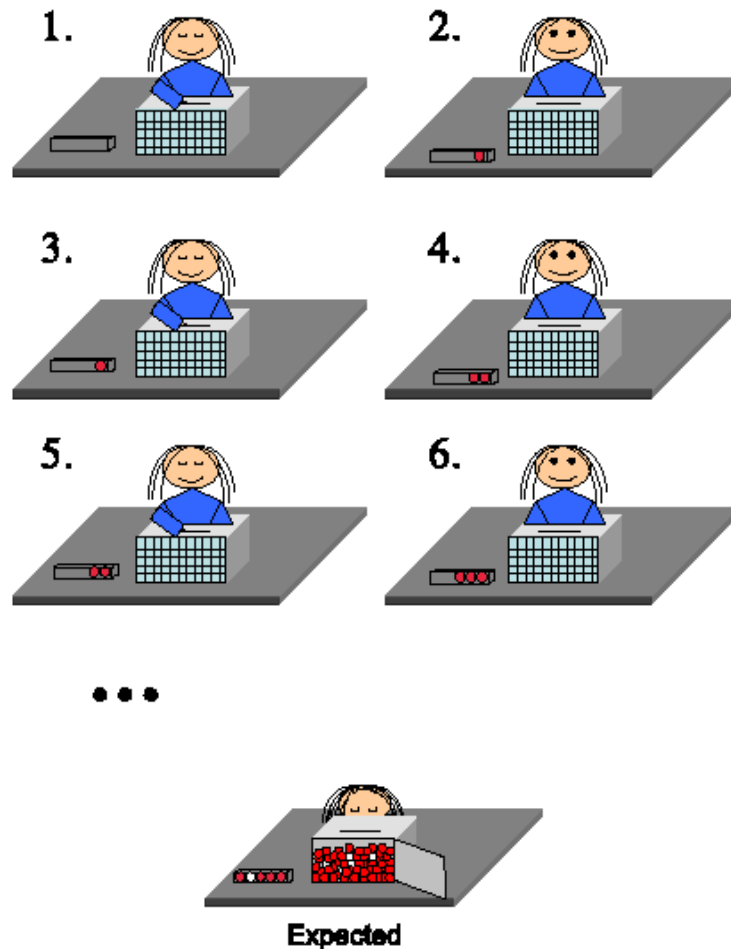


Point estimate 80.3

The mean 80.3 is a point estimate of the true value μ_0 of the weight in the population. With a probability of 95%, the true value μ_0 of the population lies within the interval [78.2, 82.5].

Sampling

Intuitive sampling knowledge of babies



The expected event
(Expected: 7.5 seconds)
takes significantly less time than
the unexpected event
(Unexpected: 9.9 seconds).

Sample: Background and characteristics

Why draw samples?

The population is often very large or not fully accessible.

→ A full survey (census) requires too much effort (time, money, personnel)

The population can be **defined** but not **identified**.

→ For example, there is no directory of non-smokers

Characteristics of a sample

A sample is a subset of all observation units and should reflect the relevant aspects of the population as accurately as possible.

When is a sample "representative"? → See also "Representativeness" from → [Slide 19](#)

Note: Representativeness is not a concept defined in statistics!

Three elements contribute to creating or describing "representativeness"

- The sample is drawn randomly.
- Estimation procedure for generalizing from the sample to the population is reported.
- Accuracy is reported, which is influenced by the sample size, among other things.

Sampling methods – Categories

Probabilistic (random) sampling procedures

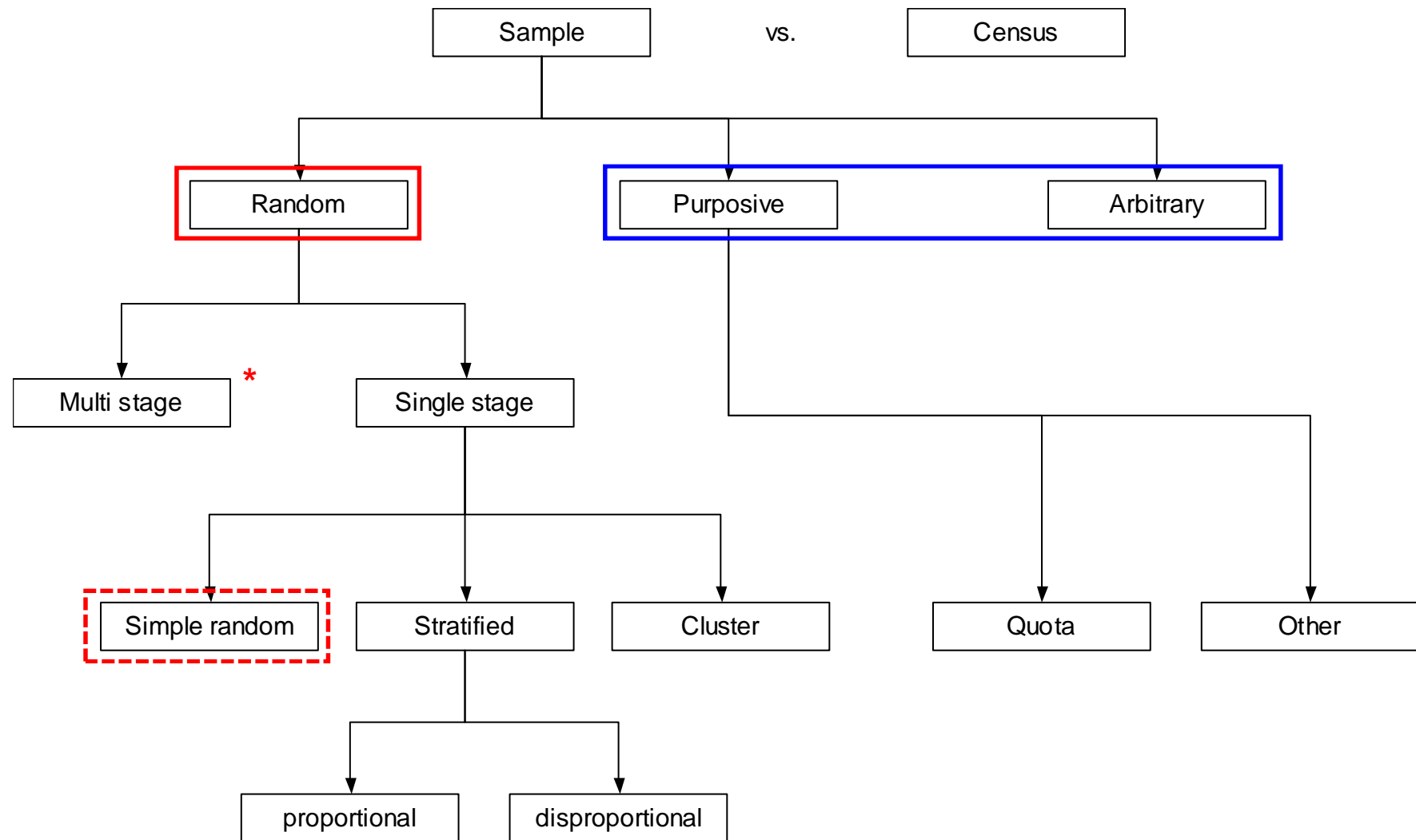
- Selecting elements based on a random mechanism
(Theory: Urn model / Practice: Random number table, random number generator).
- Each element has a positive, non-zero probability of being included in the sample.
The selection probability of the elements is known or can be calculated.
- Drawing conclusions from the sample to the population is permitted.

Non-probabilistic (non-random, purposive, arbitrary) sampling procedures

- Selection of the elements is not based on a random mechanism, but is made by certain decisions (purposive or arbitrary selection of elements).
→ Part of the elements has a selection probability of zero.
- Selection probability of the individual elements is not known and cannot be calculated.
- Drawing conclusions from the sample to the population is **not** permitted.
It is often unclear to which population the sample refers.

A probabilistic sampling is a necessary prerequisite for drawing conclusions from the sample to the population.

Sampling procedure – Overview



*Multi stage sampling consists of combinations of single-level procedures with different selection units.

Additional information in the Appendix starting on → [Slide 27](#)

Details on the sampling procedures

General requirement: Sampling frame

- Best case: A list of all elements of the population is available
Example Students at the University of Zurich in spring semester
- Worst case: No list is available
Example Customers of a shopping center

Simple random sampling (SRS)

Features

- Random selection of n elements from the N elements of the population
- Each element has the same probability of being included in the sample

Procedure

- Drawing with random generator from a list
- Option: Random route sample
Random start number (1 to N/n) and step size N/n

Simple random sampling with R – Example with data set *address** and R file**

address is an address list with N = 699 entries (*sampling frame*)

For to conduct an online survey, a random sample with n = 300 addresses is drawn.

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

address_sample <- address[sample(699, 300, replace=FALSE),]
View(address_sample)
```

replace=FALSE

The element drawn from the data set is not "put back" (replaced). Each element in the data set can occur only once in the sample.

address.csv

	_01_name	_02_forename	_03_surname	_04_mailaddress
693	Iye Cooke	Iye	Cooke	Iye.Cooke@gavv.com
694	Rosie Booth	Rosie	Booth	Rosie.Booth@uyge.com
695	Connie Lyon	Connie	Lyon	Connie.Lyon@9edx.com
696	Bianca Burris	Bianca	Burris	Bianca.Burris@zvyj.com
697	Ray Cabrera	Ray	Cabrera	Ray.Cabrera@tk8u.com
698	Clare Oconnell	Clare	Oconnell	Clare.Oconnell@wnlw.com
699	Vivian Mcmillan	Vivian	Mcmillan	Vivian.Mcmillan@xrea.com

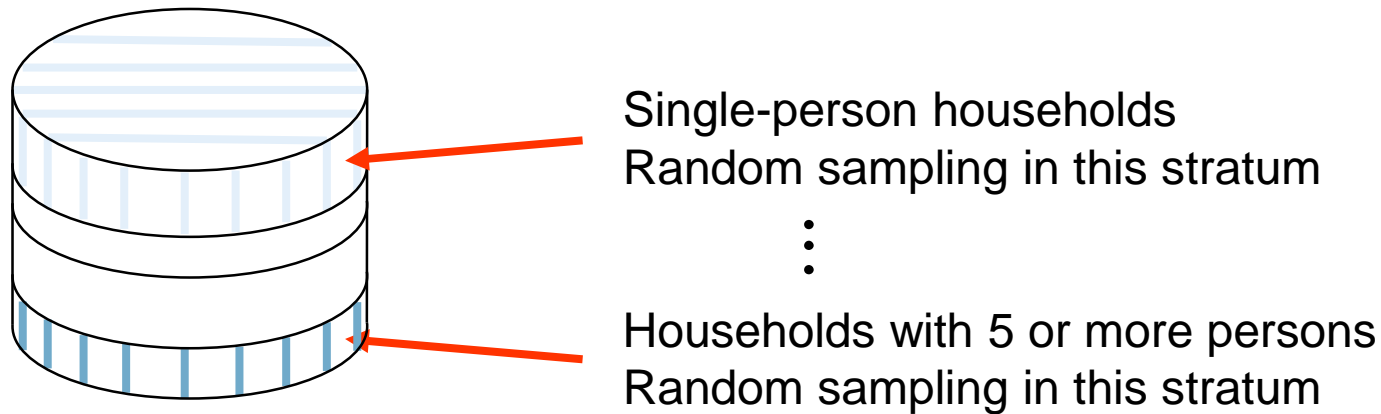
address_sample.csv

	_01_name	_02_forename	_03_surname	_04_mailaddress
294	Vicki Whitehouse	Vicki	Whitehouse	Vicki.Whitehouse@wqirg.com
295	Micah Ashton	Micah	Ashton	Micah.Ashton@7rs2.com
296	Iram Velez	Iram	Velez	Iram.Velez@bokj.com
297	Elaine William	Elaine	William	Elaine.William@3avy.com
298	Jamie Doherty	Jamie	Doherty	Jamie.Doherty@8yam.com
299	Riya May	Riya	May	Riya.May@siwt.com
300	Martyna Johnson	Martyna	Johnson	Martyna.Johnson@alq5.com

Stratified random sampling

Example

- Population = Private households in the canton of Lucerne
- Stratification by household size



Features

- Structure of the population regarding certain characteristics is known in advance.
- The population is stratified in accordance with these characteristics.

Procedure

- Random sampling per stratum

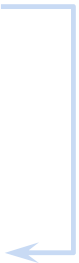
Proportional: Selection set in each stratum is the same (self-weighting)

Disproportional: Selection rate per stratum is different (samples will be weighted)

Stratified random sampling – Examples of variants

Population = Private households in the canton of Lucerne in the year 2023*

	Total	1 Person	2 Persons	3 Persons	4 Persons	5+ Persons
Absolute values	190'682	66'916	64'851	23'166	24'302	11'447
Proportions	100%	35.1%	34.0%	12.1%	12.7%	6.0%
Sample proportional	1'000	351	340	121	127	60
Proportions (rounded)	100%	35%	34%	12%	13%	6%
Sample disproportional A	1'000	200	200	200	200	200
Proportions (rounded)	100%	20%	20%	20%	20%	20%
Sample disproportional B	1'000	302	304	113	121	160
Proportions (rounded)	100%	30%	30%	11%	12%	16%



Cases A & B

The shares are weighted according to the population (Design weight → [Slide 29](#)).

Case **A** → Example of a disproportional stratified sample with **equal percentages**.
→ this allocation facilitates the between-strata analysis

Case **B** → Number "5+ Persons" is small ($n = 60$), because share in the population is small.
→ to improve the quality of the estimate, the proportion in the sample is increased
→ this process is called **oversampling**

Cluster sampling

Example

- Study of pupils in the 6th grade in German-speaking Switzerland
- Random selection of schools / classes
(SRS or stratified, proportional, disproportional, etc.)
- Survey of all pupils in the selected classes

Features

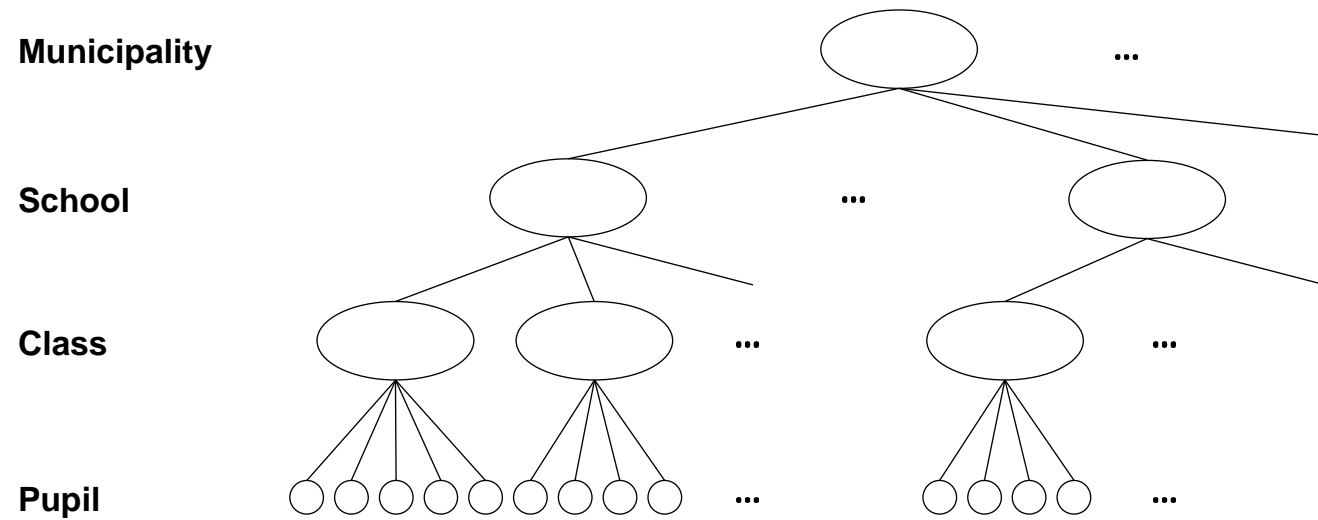
- The elements are selected at a higher level.
- Cluster sampling mainly solves logistical problems in terms of efficiency of the analysis as well as feasibility and cost.
- If there are no lists of the population's elements, cluster sampling is generally the only option.

Special

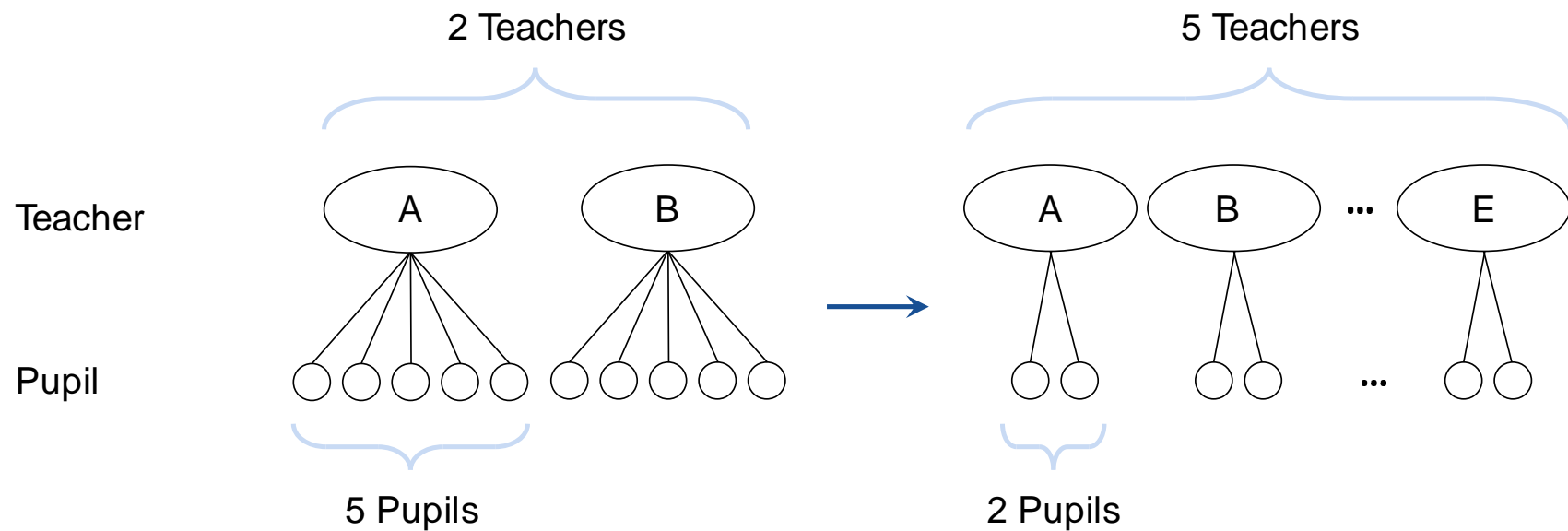
Cluster effect =
$$\frac{\text{Variability **within** clusters is **small** (cluster elements are very similar)}}{\text{Variability **between** clusters is **large** (clusters differ greatly)}}$$

Many clusters with few objects will reduce the cluster effect.

Cluster sample – Basic structure



Cluster sample – Cluster effect



Arbitrary sampling

Example snowball sampling

- Study of female sex workers (street and escort)

Snowball sampling: First contact via the "Flora-Dora-Bus" (mobile service of the Flora-Dora women's counselling center of the City of Zurich). There, sex workers found flyers with information about the purpose of the study and a contact number.

Many women took part who had been informed about the study by colleagues.

Features

- Application in populations where it is difficult to find the elements.
- Social networks are used to recruit the sample.

Other methods

Targeted sampling: For populations that are difficult to reach and whose members are not closely networked. Preferred locations or places of residence of the members are identified and then systematically recruited on site.

Respondent Driven Sampling: Multiple waves of peer-to-peer recruitment with statistical adjustments are conducted to approximate a random sample. Recruited individuals are only allowed to recruit a limited number of other individuals and are only rewarded for each person actually recruited.



Representativeness

Representativeness

Statements in the media:

Voting poll ...

The sample is [...] representative of Swiss voters.

Often an empty phrase to emphasize the seriousness of the survey.

But: The degree of representativeness is not measurable!

There is no equation (such as standard error) for calculating representativeness.

There is no such thing as a 'representative', 'unbiased', 'fair' (...). A sample can be judged only in relation to the process that produced it. The central concepts of selection bias and sampling error have no meaning except in this context.

Representativeness for all possible features is impossible.

The sample should be representative with regard to key characteristics of the study.

See also the elements indicated by the Federal Statistical Office (→ [Slide 8](#))

Sample size? / Representative sample?

Election forecast by *The Literary Digest*

Presidential election in the U.S. in 1936 – the eighth year of the Great Depression

Two candidates were in the race

- Franklin D. Roosevelt (Democrat, President-in-Office, inventor of "New Deal")
The New Deal was a series of economic and social reforms launched under Roosevelt to face the global economic crisis. It was a radical change in economic, social and political history.
- Alf Landon (Republican, Challenger)

The US magazine *The Literary Digest* conducted a political survey.

About 10 million questionnaires were mailed and 2.3 million were returned.

The prediction by *Literary Digest* was notably different from the actual election result.

	Literary Digest	Election result	Gallup
Roosevelt	40%	62%	56%
Landon	60%	38%	44%

Sampling errors & Variability of sample means

Types of sampling errors

Non-sampling error

Definition: Difference in the mean value between the defined ideal population and the real population that cannot be attributed to deficiencies in the random selection of the sample.

Forms

- Coverage error ("coverage error"): Part of the population cannot be identified.
Example: The telephone directory for the landline network serves as the basis for sampling.
- Systematic non-response ("non-response error"):
Lack of information on certain individual elements.
Example: People systematically refuse to answer a question in a survey.

Sampling error

Definition: Difference between the estimated mean value from a randomly drawn sample and the real mean value of the population.

Forms

- Selection error: Not all elements of the population have the same selection probability.
Example: Working people answer the phone less often during the day.
- Use of an unsuitable estimator.
Example: The empirical variance is calculated with $1/n$ instead of $1/(n-1)$.

Variability of sample means

Factors influencing sample mean variability

In randomly drawn equal samples, the sample means vary depending on the ...

- Attribute → The more heterogeneously an attribute is distributed in the population, the greater the variability of the sample means among many samples.
- Sample size → The smaller the sample size, the greater the variability of sample means among many samples.

Reasoning about the standard error of the sample mean

The standard error ...

- is a measure of the variability of the sample means among many samples.
- says something about the quality of the estimated mean.
- quantifies the spread of sample means from repeated random samples of the same size around the population mean μ_0 .

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

$\hat{\sigma}_{\bar{x}}$ = Standard error of the mean

s = Standard deviation of the sample

n = Sample size

Conclusion

The larger the sample size, the smaller the standard error, and therefore, the larger the sample size, the more precise the sample mean is as an estimator of the population mean

Preview of Lecture 06

What has happened so far

The relationship between population and sample

Sampling procedures

- Drawing a simple **random sample** by using RStudio.
- Known is an **equation** for calculating the **sample size** and using RStudio to do so.
- Concepts on **representativeness** / **sampling error** / **error sources** are understood.

What follows in Lecture 06

Sample size & power analysis

- How large must a sample be in order to test a certain effect?
Answer: Power analysis → with G*Power and R

Effect size

- What is that? → Concept by Jacob Cohen



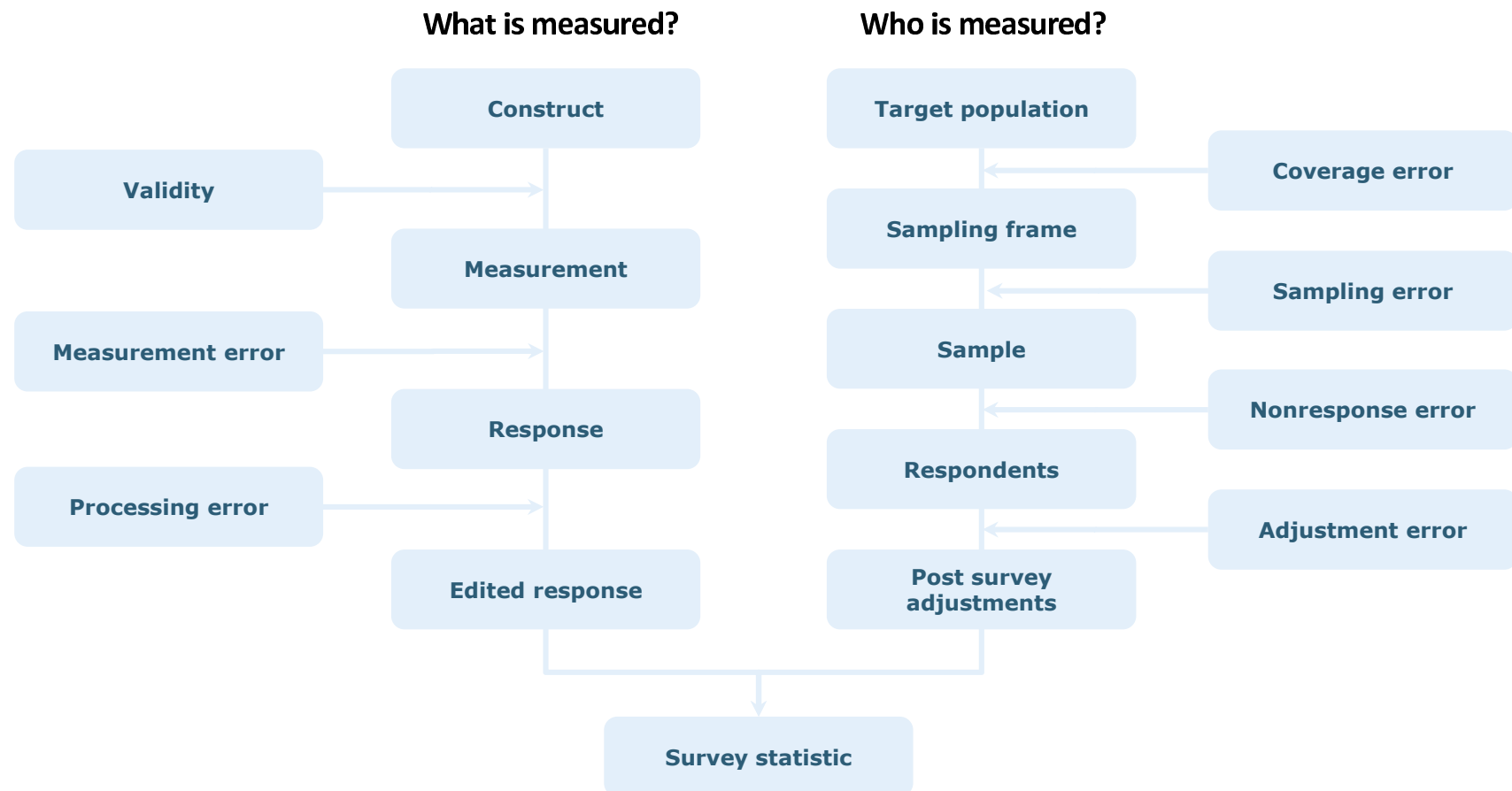
Details about the sampling errors

Concept for surveys: Total survey error

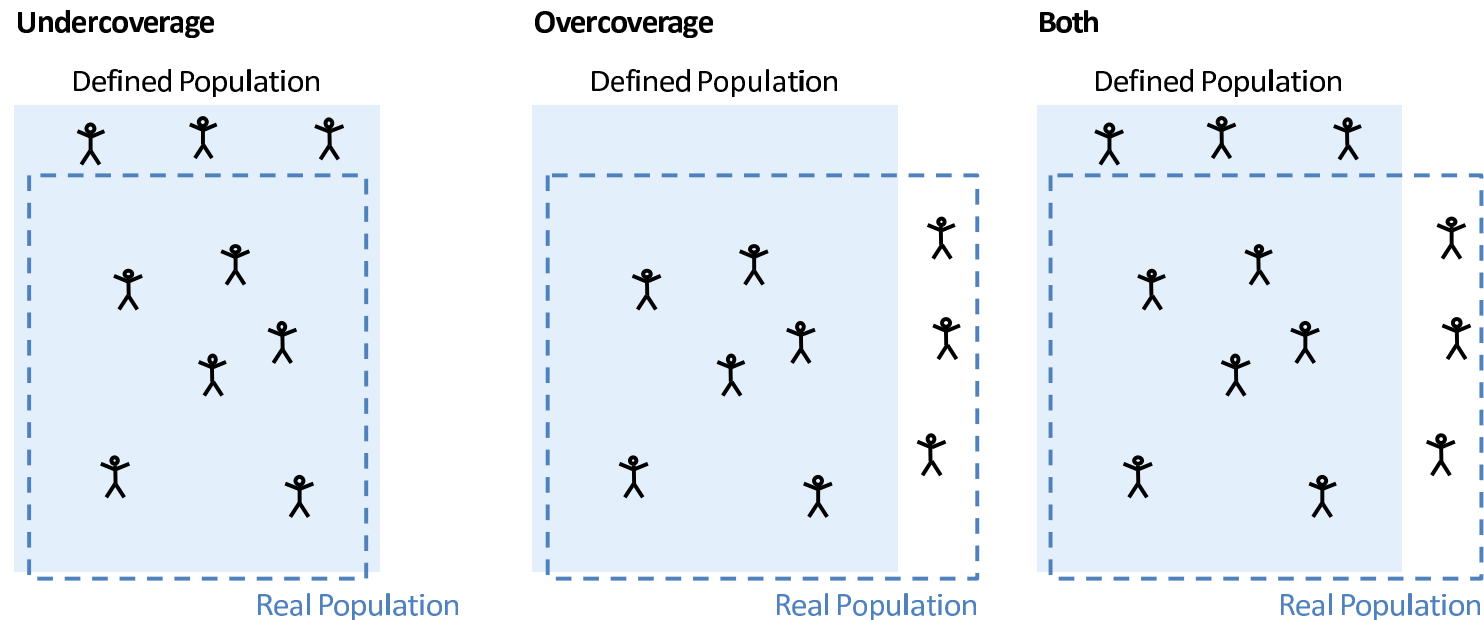
Total survey error is a concept that purports to describe statistical properties of survey estimates, incorporating a variety of error sources.

Unfortunately, the term "total survey error" is not well defined. Different researchers include different components of error within it, and a number of typologies exist in the literature.

Robert M. Groves, Director of the U.S. Census Bureau until 2012



Example of an Non-Sampling Error → Coverage Error



Example of undercoverage

- Defined population Inhabitants of a city
 - Real population Entries in local phone book
- Persons with suppressed phone numbers are missing in the sample.

Example of overcoverage

- Defined population Inhabitants of a city
 - Real population Entries of doctors in the local phone book
- Persons who are not inhabitants are included in the sample.

Appendix

Comparison of sampling methods

Random sampling	Purposive sampling	Arbitrary sampling
◦ Inference to the population possible	◦ Limited inference to the population possible	◦ No inference to the population possible
◦ Quantitative research design	◦ Based on quantitative research design	◦ Explorative (qualitative) research design
◦ Methods of inferential statistics	◦ In principle, methods of inferential statistics	◦ Creation of theory or hypothesis possible
◦ Sample size must be estimated	◦ Structure of the population must be known	◦ Simple access to the elements of the sample
◦ Random error can be estimated	◦ Design weighting necessary	◦ Example: convenience sample

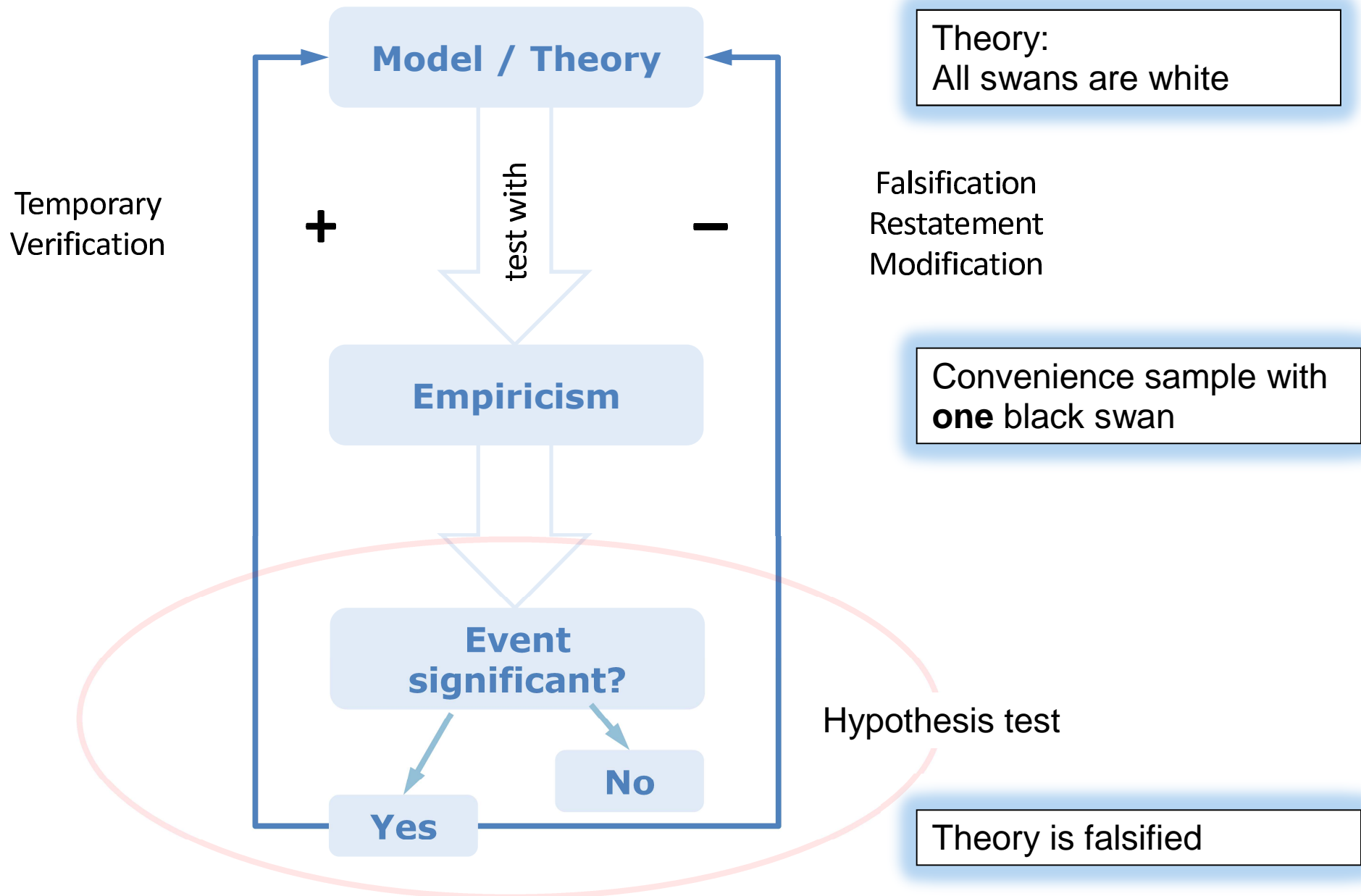
Possible statement

- Swans weigh
12 kg \pm 5% on average

Possible statement

- Not all swans are white.
(falsification)

Arbitrary sampling in the light of critical rationalism (Karl Popper)



Design weight

The term "design weight" (also known as "sampling weight") refers to weighting factors used in statistical analyses to adjust the data to more accurately represent the population from which the sample was drawn. These weights compensate for inequalities in sampling caused by the sample design, such as an uneven probability of selection.

The design weight is typically determined after sampling, but the process and methodology needed to calculate it **must be carefully planned prior to sampling**.

Design weights are used in a variety of research fields, particularly in survey research and epidemiologic studies where samples are weighted to be representative of a larger population. They are crucial when analyzing data from complex sampling designs, such as stratified, multi-stage or lumped samples, and help to obtain reliable estimates of population parameters.

Sample size

Equation of Cochran

$$\text{Sample size} = \frac{z^2 \cdot p \cdot (1-p)}{e^2}$$

p → proportion in the population

Assumption about the proportion of the variable to be examined in the population.

Example: 61% of Swiss households have christmas trees → **p = 0.61**

e → error range

Error range for the estimates obtained from the sample (means, etc.)

Typical value for the error range: 3% → **e = 0.03**

z → z value

The z value is associated with a general error probability:

z indicates how likely the calculated value of the sample size is wrong.

Typical value for the error probability: 5% → (...) → **z = 1.96**

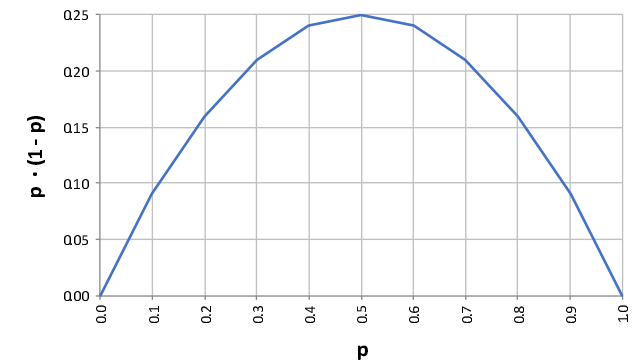
An example:

Research question: How many Swiss households have christmas trees?

How large a sample do you need in order to answer the research question?

Prerequisites

- In the absence of information, it is assumed that 50% of households have christmas trees → **p = 0.50**
With p = 0.50, the sample size becomes the **maximum**, with given z and e. It is therefore an upper estimate of the sample size.
- The error probability is 5% → **z = 1.96**
- The error range is 3% → **e = 0.03**



$$\text{Sample size} = \frac{z^2 \cdot p \cdot (1-p)}{e^2} = \frac{1.96^2 \cdot 0.50 \cdot (1-0.50)}{0.03^2} = \frac{3.842 \cdot 0.25}{0.0009} = 1067.1$$

With a sample of size $n = 1067^*$, the proportion of christmas trees in Swiss households can be determined.

The margin of error is 3%; the error probability is 5%.

*Rounding up or down to the nearest integer

Correction for small populations

The equation must be adjusted if there are only a few elements in the population.

$$\text{Sample size} = \frac{\frac{z^2 \cdot p \cdot (1-p)}{e^2}}{\text{Correction term}} = \frac{\frac{z^2 \cdot p \cdot (1-p)}{e^2}}{1 + \frac{z^2 \cdot p \cdot (1-p)}{e^2 \cdot N}}$$

z = z value

p = Proportion in the population

e = Error range

N = Number of elements in the population

Number of elements N in the population	Correction term
100	11.67
1'000	2.07
10'000	1.11
50'000	1.02
100'000	1.01
1'000'000	1.00
4'000'000	1.00
8'000'000	1.00

Example to find in *Collection of R code*

Research question: How many Swiss households have christmas trees?

How large a sample do you need in order to answer the research question, for the whole of Switzerland (4'000'000 households) and for Rheinfelden (10,000 households)?

Prerequisites

- In the absence of other information, you choose **p = 0.50**, which leads to an upper estimate of the sample size.
- The error probability is 5% → **z = 1.96**
- The error range is 3% → **e = 0.03**

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

rsampcalc(N=4000000, e=3, ci=95, p=0.5)
rsampcalc(N=10000, e=3, ci=95, p=0.5)
```

The 5% error probability is the complement of the confidence level (100% - 5%) = 95%
→ ci=95

```
> rsampcalc(N=4000000, e=3, ci=95, p=0.5)
[1] 1067
> rsampcalc(N=10000, e=3, ci=95, p=0.5)
[1] 965
```

Calculating the sample size with Cochran's equation

Survey in Switzerland N = 4,000,000

Correction term = $1.0003 \approx 1.0$

$$n = \frac{\frac{z^2 \cdot p(1-p)}{e^2}}{1 + \frac{z^2 \cdot p(1-p)}{e^2 \cdot 4'000'000}} =$$

$$n = \frac{\frac{1.96^2 \cdot 0.50 \cdot (1-0.50)}{0.03^2}}{1 + \frac{1.96^2 \cdot 0.50 \cdot (1-0.50)}{0.03^2 \cdot 4'000'000}} =$$

$$n = \frac{1067.1}{1 + 0.0003} = 1066.8$$

Survey in Rheinfelden N = 10,000

Correction term = $1.1070 \approx 1.1$

$$n = \frac{\frac{z^2 \cdot p(1-p)}{e^2}}{1 + \frac{z^2 \cdot p(1-p)}{e^2 \cdot 10'000}} =$$

$$n = \frac{\frac{1.96^2 \cdot 0.50 \cdot (1-0.50)}{0.03^2}}{1 + \frac{1.96^2 \cdot 0.50 \cdot (1-0.50)}{0.03^2 \cdot 10'000}} =$$

$$n = \frac{1067.1}{1 + 0.1070} = 964.2$$

Table of contents

Population	3
Generalizing from a sample to the population	5
Sampling	7
Intuitive sampling knowledge of babies	7
Sample: Background and characteristics	8
Details on the sampling procedures	11
Representativeness.....	19
Representativeness	19
Sampling errors & Variability of sample means.....	21
Types of sampling errors	21
Preview of Lecture 06.....	23
What has happened so far	23
What follows in Lecture 06	23
Details about the sampling errors.....	25
Appendix	27
Comparison of sampling methods	27
Design weight	29
Sample size.....	30
Equation of Cochran.....	30

