Sampling

Applied data science with R

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

Learning Goals

- Understand the difference between a **sample** and a **population**
- Learn how to do a **Monte Carlo simulation** in R and understand its usefulness
- Understand the **Central Limit Theorem** and its practical importance for inferential modelling

Motivation

Why sampling?

- The goal of scientists (and managers) is often to learn something about phenomena that involve a great number of subjects
  - **Marketing**: how do customers respond to certain ads?
  - **Sociology**: how do attitudes on climate change relate to citizen's socio-economic background?
  - **Economics**: what makes a firm competitive?
  - **Political science**: whom do people vote for and why?
- In all these cases, the subjects from a very large (or even unknown) **population**
- We cannot study the entire population → study subsets of it, and then make inferences about the whole population

→ These subsets are called **samples**, and when and how the **inference** from a sample to a population works will be the subject of the upcoming sessions

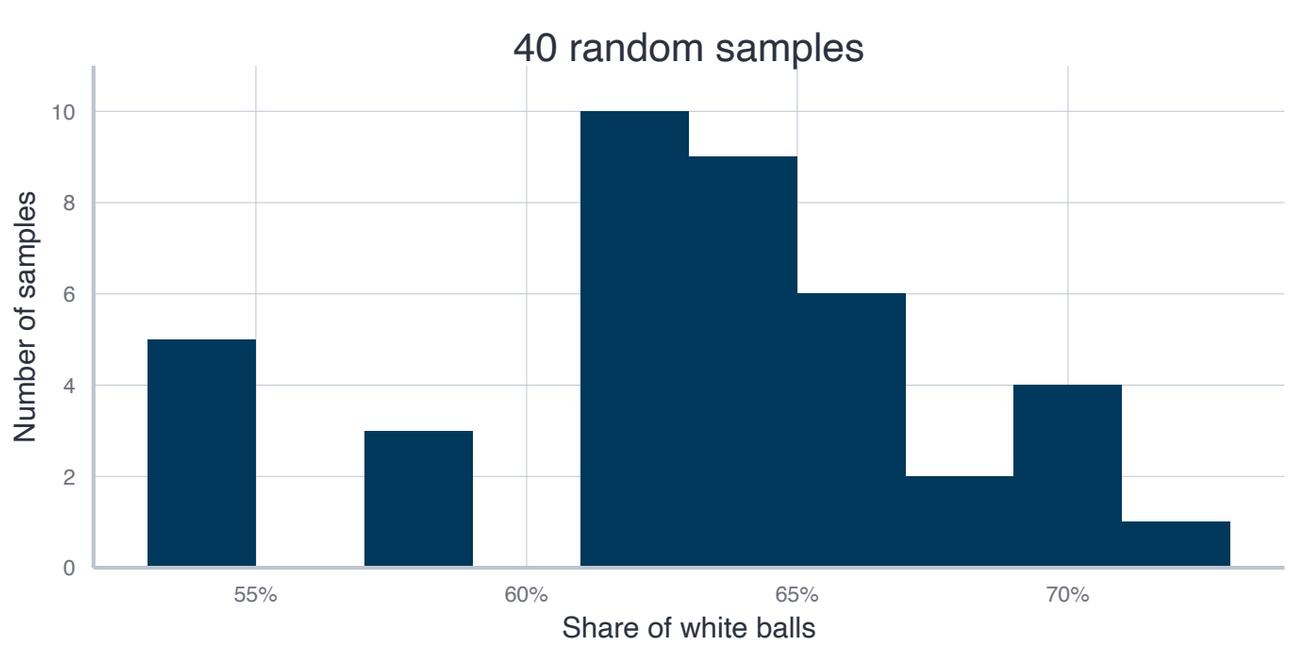
Motivating example

- Suppose we have bought a ball pid with grey and white balls
- How many of the balls are grey? How many are white?
- We could either do an exhaustive count
  - But if the seller is correct, the ball pid contains 5.000 balls → 🤯🥵

- Alternatively, take a sample of 50 balls, count them, and make an inference about the original ball pid 😎

Motivating example

- Suppose we sample 50 balls and find that 64% of the balls were white
  - Does this mean that 64% of all balls are white?
- Not really, our sample was drawn randomly → **random sample**
  - Repetition of the process → different share of white balls
  - Suppose we draw 40 such random samples and each time report the share of white balls



The fact that the different random samples differ from each other is referred to as the concept of **sample variation**

Monte Carlo simulations

Studying sampling via simulations

- Idea: simulate the act of drawing samples from a population computationally
  - The act of drawing a random sample is a **random process**
  - Simulations used to study properties of random processes by repeating them many times are called **Monte Carlo simulations** (MCS)
- MCS help us understand determinants & implications of sampling variation

- In an MCS we create the population ourselves and know everything about it
  - **In reality** we do not know the true properties of the population
  - In reality we only draw **one single sample**
  - The stylised MCS context is still useful to study determinants & implications of sampling variation

Monte Carlo Simulations

- Key idea: create **artificial population** about which we know everything
- Then draw samples from this population and study questions such as:
  - Are properties of samples similar to that of the population?
  - What determines sample variation?
  - What is the effect of different sample sizes or sampling iterations?
- Then hope that the strategies to answer these question behave similarly in the real world when we do not know the true population and draw only one sample

For more details on the practical implementation see the tutorial!

Detour: for-loops in R

- A for-loop is a tool that ‘loops over’ an object and implements the same operation during each iteration

- Example for such a ‘base object’:

```
base_list <- list(
  "element_1" = c(1, 4, 5, 6),
  "element_2" = c(9, 2, 2, 8),
  "element_3" = c(4, 3, 0, 7)
)
```

- A for-loop consists of three parts:

```
result <- rep(NA, length(base_list)) # The output container
for (i in seq_along(base_list)) { # The looping sequence
  result[[i]] <- mean(base_list[[i]]) # The action body
}
```

Detour: for-loops in R

- Some hints:
  - Start by writing a very short loop with 2-3 iterations only and make sure everything works
  - Add `print()` statement into the action body to see whether the looping keyword works as intended
  - Always use `seq_along()` or `seq_len()` in your looping sequence

Exercise

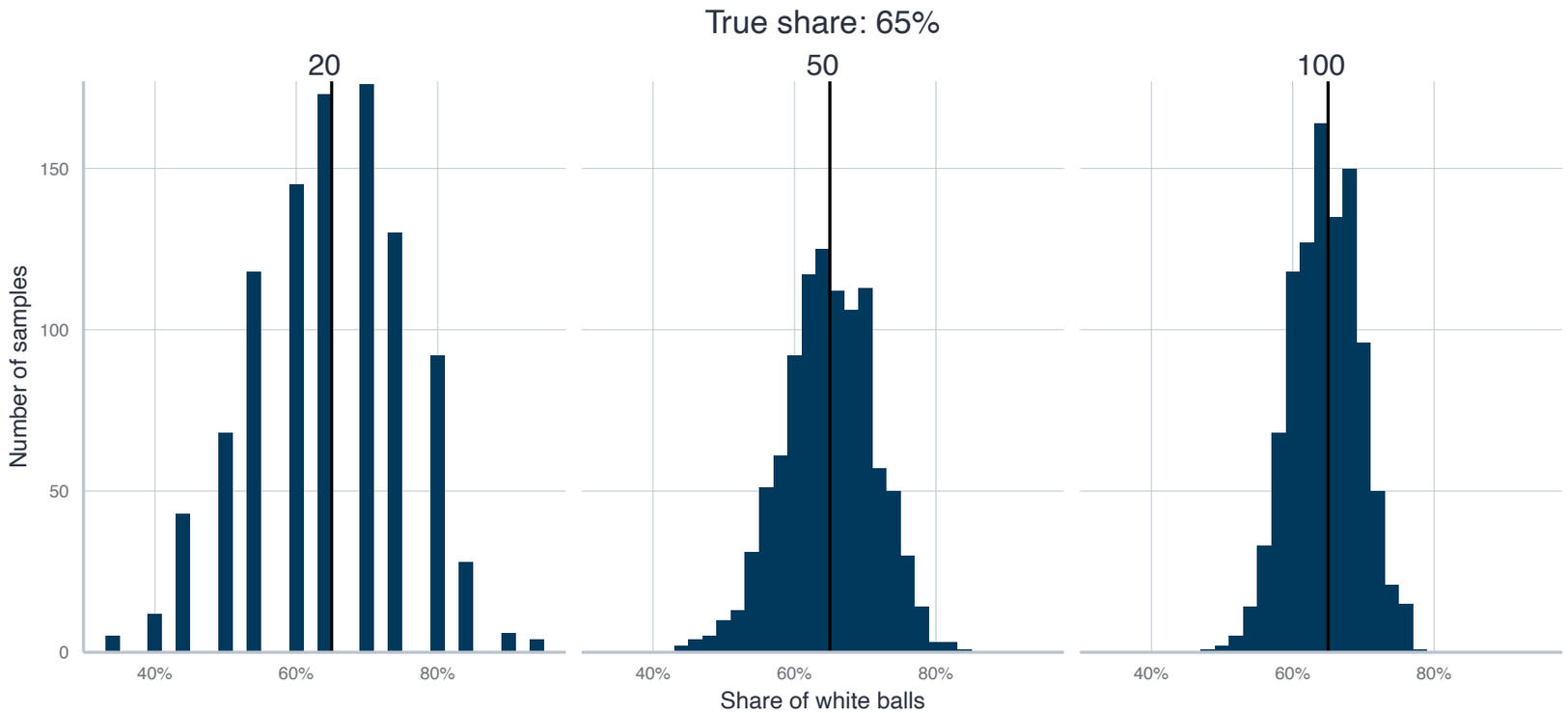
Write a for-loop that loops over the vector `c(1,2,3,4,5)` and computes the square root for each element.

Monte Carlo Simulations

- Back to our ball pid
- **Question: what sample size would be good?**
- Simulation:
  - Artificial pid with 5000 balls, 65% white
  - Action to be taken: draw a sample of size `sample_n`
  - Compute the share of white balls and record this value
  - Iterate this act 1000 times and look at the sample variation
  - Answer the question: how does the sample size impact our estimate?
- If the sampling in reality works according to the same mechanisms, conclusions from this artificial example carry over to real cases!

Monte Carlo Simulation - central results

- Here is a summary of our central results:



Sample size <int>	Mean share <dbl>
20	0.65155
50	0.64786
100	0.65108

Sample size <int>	Variation <dbl>
20	0.10801280
50	0.06418364
100	0.04806604

We measure the variation via the standard deviation

- And these are the central take-aways:

I. All distributions have a very similar mean of about 65%

II. The larger the sample, the smaller the sample variation

Exercise: Monte Carlo Simulation

- Assume you want to compute the average height of students of the Europa-University Flensburg

- Assume that the data set `DataScienceExercises::EUFstudents` contains the result of a census among EUF students

- Study the process of sampling by conducting an MCS in which you draw random samples from this population of sizes 10 or 50.

- For your MCS, set the number of repetitions to 1000

- What do you observe for the differente sample sizes?

  - Note: a quick-and-dirty way to represent your results is the function `hist()`

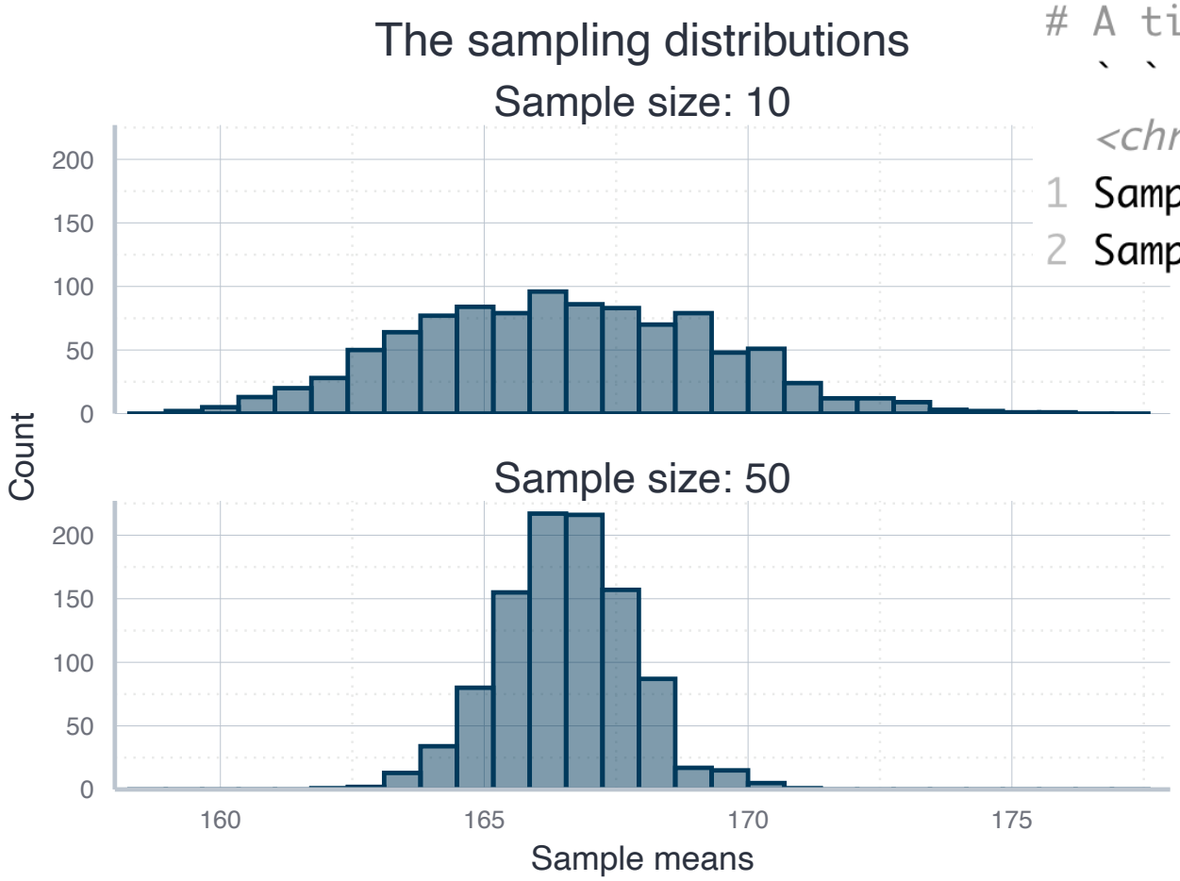
Exercise - MCS

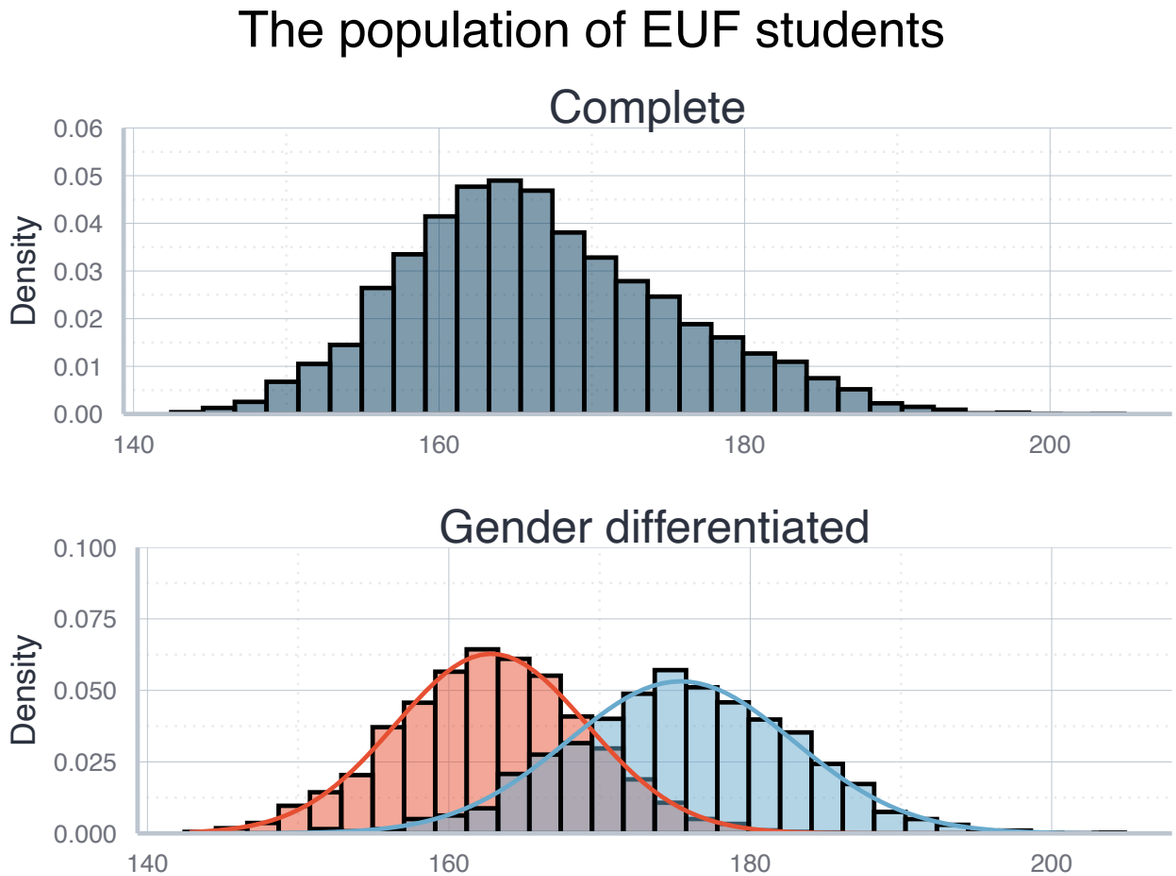
Sample statistics

```
# A tibble: 2 × 3
  ``                   Mean Variation
  <chr>               <dbl>     <dbl>
1 Sample size: 10      166.      2.83
2 Sample size: 50      167.      1.24
```

Population statistics

```
# A tibble: 3 × 3
  Gender  Mean    SD
  <chr>  <dbl> <dbl>
1 female  163.  6.35
2 male    175.  7.50
3 total   167.  8.85
```





- Note: the student population of the EUF is asymmetric in terms of gender
  - While the population is not normally distributed, the sampling distributions tend to be normal → take up later

Terminology

Terminology

- In the following we introduce the fundamental terminology that we use when talking about anything that has to do with sampling

- We will cover the following areas:

Populations	Samples	Methodology

- Most of these concepts are also of prime importance in the context of estimation and inference

Population terminology

A **population** is a collection of individuals or objects that are of interest.
Population size N: the number of individuals making up the population

A **population parameter** is a statistical property of the population that is of interest.

A **census** is the act of studying each member of the population to determine the population parameter of interest exactly.

- **Example:** We are interested in the average height of all German women.
  - Population: all German women ($N \approx 42$M)
  - Population parameter: population mean
  - Census: measure all German women and compute the mean height

Sample terminology

A **sample** is a subset of the population. If the elements of the sample were selected randomly, we speak of a **random sample**.

The **sample size** is the number of its elements and denoted as $n < N$.

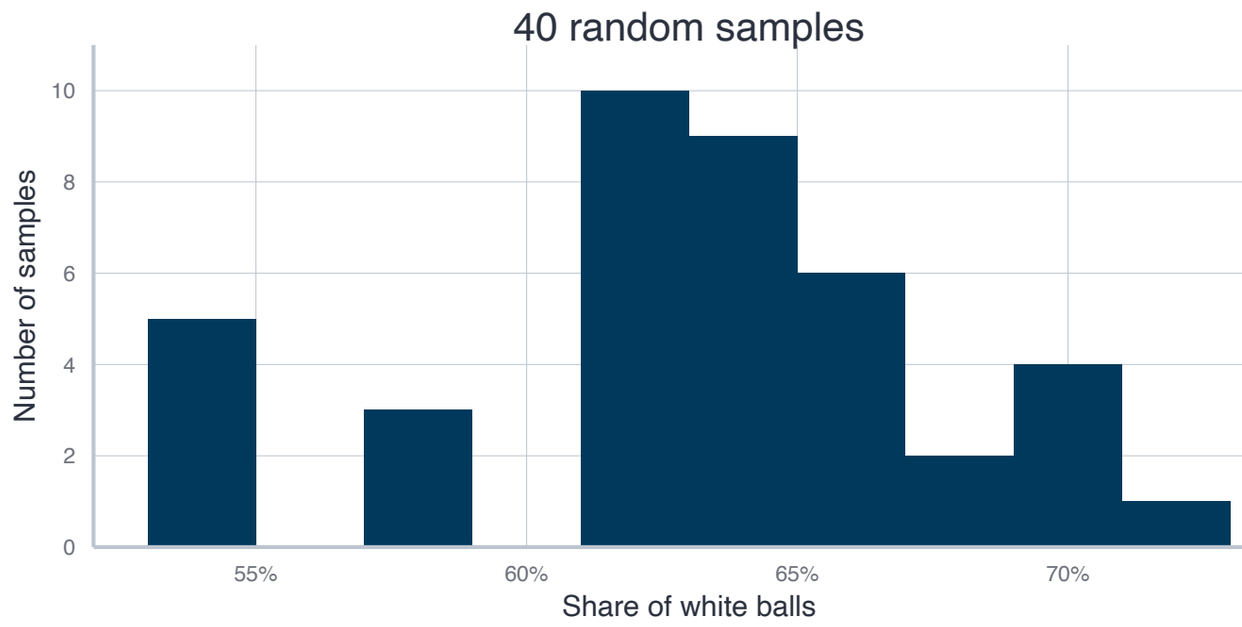
A **point estimate** or **sample statistic** is a statistic computed for the sample and that is to be used to **estimate** the population parameter of interest. It is written with a ^ on the symbol (e.g. $\hat{\beta}$).

- **Example:** We are interested in the average height of all German women.
  - (Random) sample: a group of (randomly selected) women in Germany
  - Sample statistic: the mean height of the women in the sample

Sample terminology

A **sampling distribution** is the distribution of a point estimate.

It formalises the effect of **sampling variation**, which originates from the random element of drawing a sample.



- **Note:** We considered the artificial case in which we drew many samples from the population. The distribution of the estimates is the sampling distribution.
  - In reality we draw only one sample → no direct access to the sampling distribution
  - We can still get information doubt the sampling distributions via **bootstrapping**

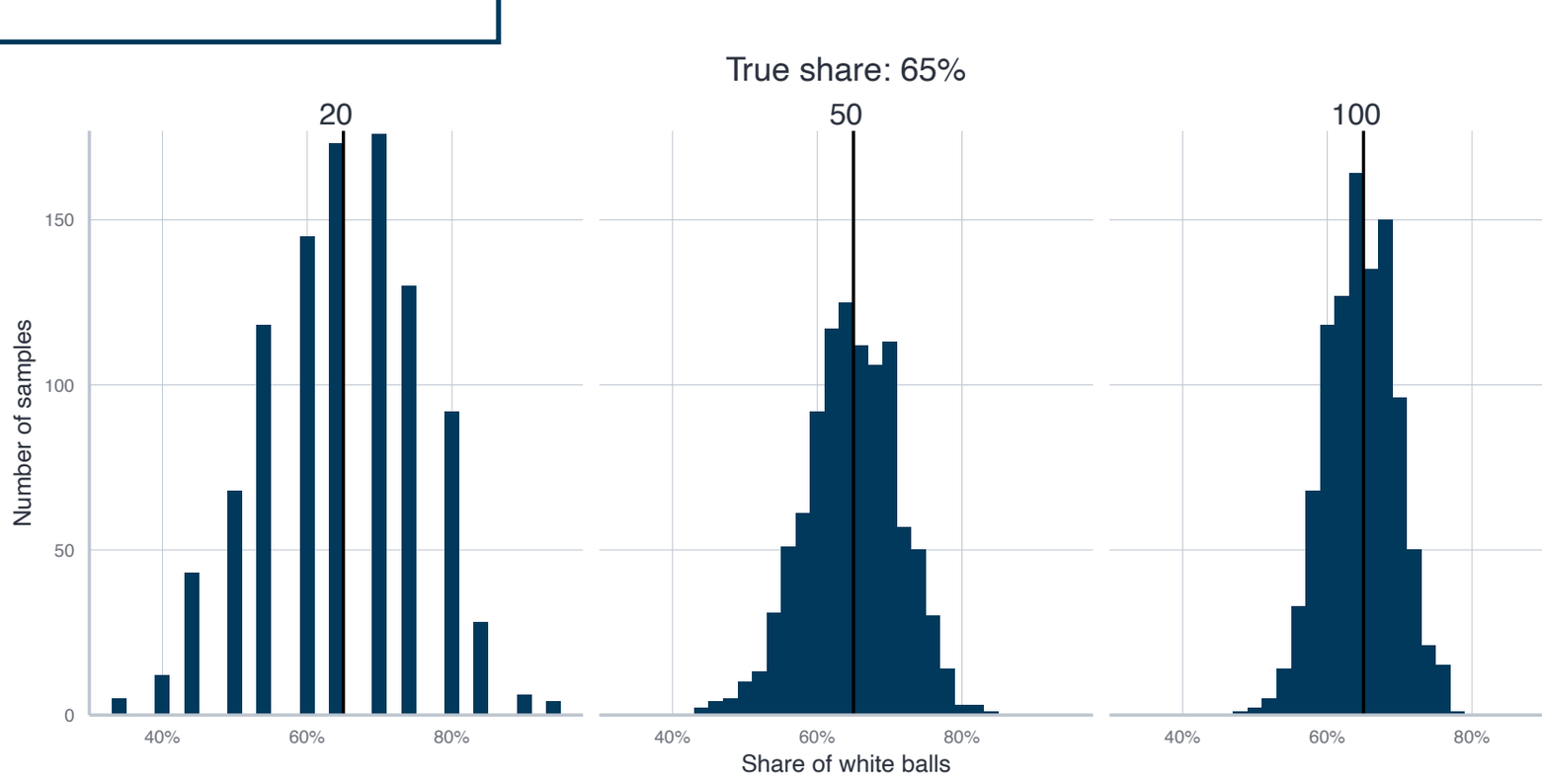
Sample terminology

A **standard error** of a point estimate is the standard deviation of its sampling distribution.

It can be used as a measure for the precision of our estimation, and it decreases with sample size.

Sample size <int>	Standard deviation <dbl>
20	0.10439990
50	0.06794096
100	0.04554750

- **Example:** The standard error of our estimate for the share of white balls $\hat{p}$ is...
  - 0.1, 0.07, and 0.05 for sample sizes of 20, 50, and 100, respectively



Methodological concepts

- Building upon the notion of a sample, here are important sample properties:
  - A sample is **representative** for a population if it resembles the relevant properties of the latter
  - A sample is **unbiased** if each member of the population has the same probability to become a member of the sample
- To ensure that a sample is representative and unbiased we usually aim to do **random sampling**
  - Hope that this produces a sample that is **generalisable**: results for the sample can be generalised into statements about the population
- The act of inferring statistical properties of a population by using statistical properties of a sample is called **statistical inference**

Wraping up the terminology

- Based on our methodology, we can summarise the process of statistical inference as follows:

1. Draw a sample of size n from the study population of size N
2. If the sample is a **random** sample...
3. ...is is usually **unbiased** and **representative** of the population
4. Then results based on the sample can be **generalised** to the population
5. This implies that **sample statistics** are good **estimators** for the respective **population parameters** → no **census** necessary

Exercise - applying the terminology

- Consider the previous example where you studied the height of selected EUF students to make a statement about average height of all EUF students
- Describe the various elements using the terminology we have introduced above. Make sure you make use of the following concepts:
  - Population
  - Sample and sample size
  - Point estimate
  - Sampling distribution
  - Standard error
  - Properties of the sample and inference

The central limit theorem

The Central Limit Theorem

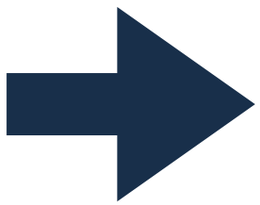
- Many of the experiments we did in this session were somehow artificial:
  - In reality we can only draw a single sample...
  - ...from a population to which we have no direct epistemic access
- Statistical inference in practice: one sample → statement about population
- But how come that statistical inference **is (often) possible**? The reason lies in the famous Central Limit Theorem

Central Limit Theorem (informal)

When a sample becomes larger, its sampling distribution becomes narrower and more normally distributed (regardless of the population distribution).

The Central Limit Theorem

- The CLT links a single sample to the population:
  - The **point estimate** based on our sample can be considered a draw from a **normal distribution** with the **mean being the true population parameter**...
  - ...and the **standard deviation** of this distribution corresponding to the **standard error** of our point estimate
- This is why sample size is so important: it makes our estimates more precise and leads to normal sampling distributions
- Again: even if the underlying distribution is not normal, the sampling distribution of the point estimates will still be normal!



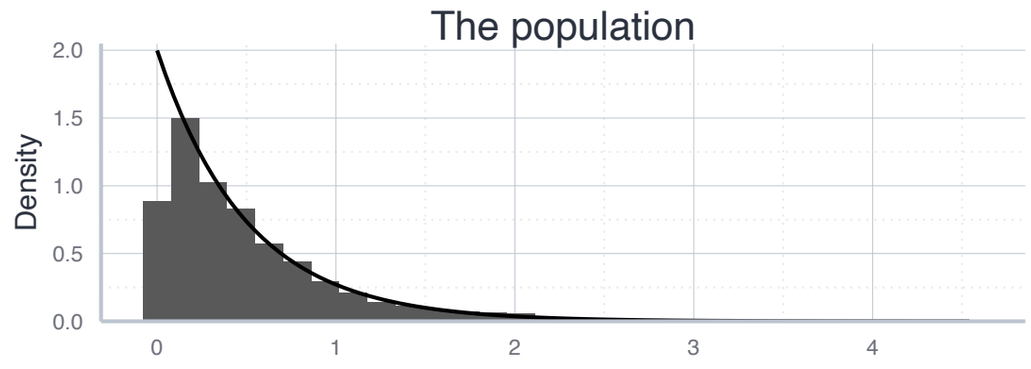
Check this out yourself!

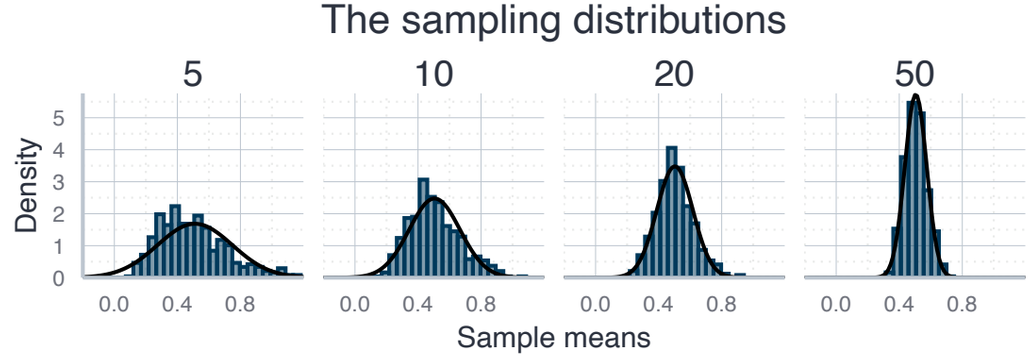
Exercise: Illustration of the CLT

- Choose a distribution for your artificial population with $N = 5000$:
  - Conduct a MCS where you draw larger and larger samples from your population
  - Visualize the sampling distribution of the sample means
- Upload your visualisations via Moodle - next week we will compare them and thereby appreciate the practical implications of the CLT more clearly!

Illustration of the Central Limit Theorem

- Here is my example with a exponentially distributed population:





- This is why we can assume a normal sampling distribution if our sample is 'large enough'

- But be aware: the CLT does not apply universally!

Summary & outlook

Summary

- Sampling theory provides tools to draw conclusions about unknown populations of interest by analysing only a sub-sample of this population
- The process of inferring population parameters of interest from samples using statistical techniques is called **statistical inference**
- We introduced all the necessary terminology to discuss the process of sampling and the methods of inference to be used
- To study how estimates are effected by sample variation we used **Monte Carlo Simulations** (MCS)
- This is a more general simulation tool to study random processes
  - Here is was useful to consider the artificial cases of drawing many samples from a known population → helps understanding how sampling works

Recap questions

- Explain the relationship between a population and a sample.
- What is a point estimate/sample statistic? How des it relate to a population parameter?
- Explain the concept of sample variation and the sampling distribution. What's the difference?
- What is a standard error?
- What properties do samples have if they are the result of random sampling?
- What is a Monte Carlo Simulation?
- Explain the Central Limit Theorem in your own words. Why is it so relevant for inferential data analysis?