universität wien

BAK3: Introduction to Quantitative Methods

Week 8: Inference

Leonardo Carella

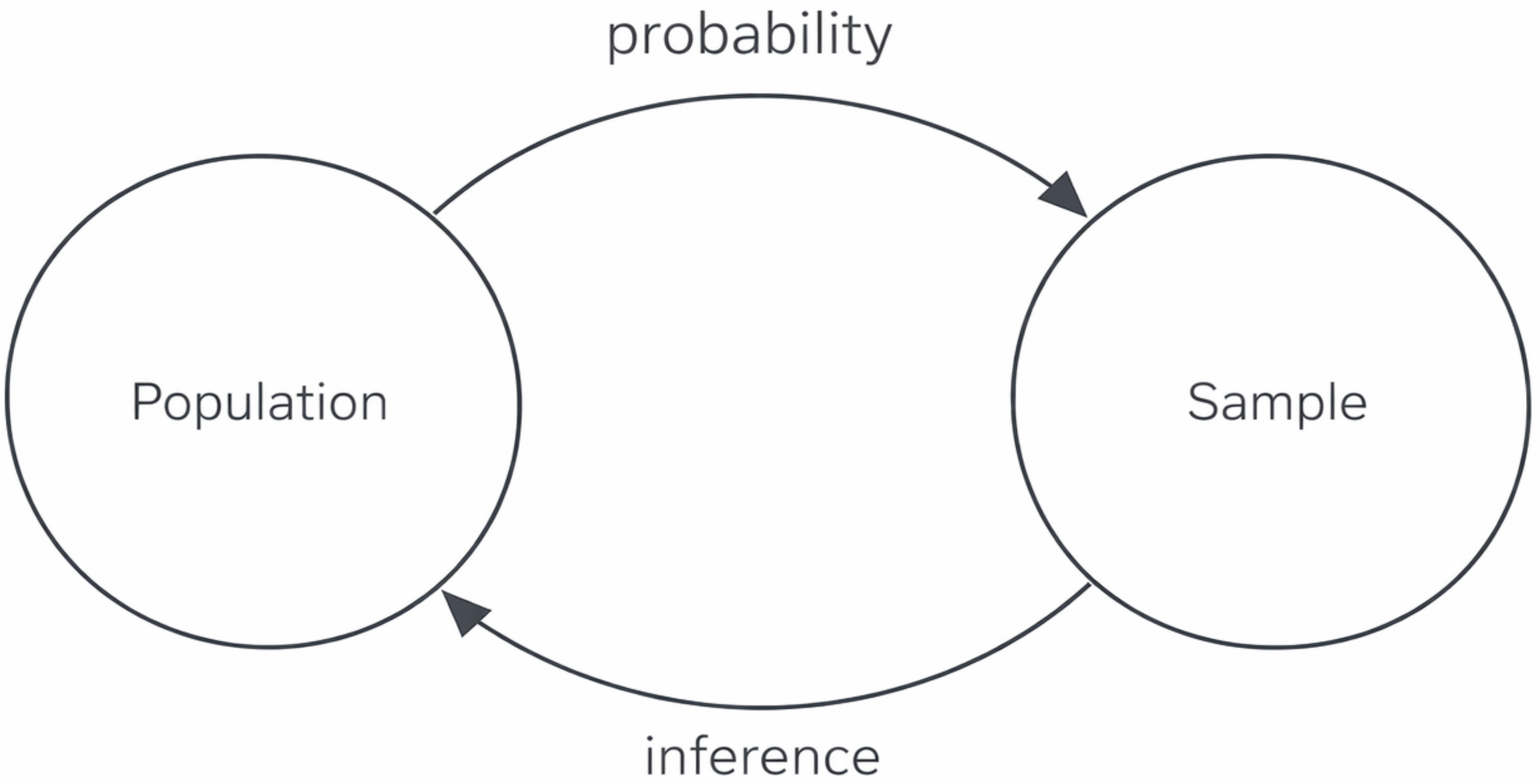
universität wien

The Plan for today

- Statistics:
  - **Review**: the standard error and normal distribution.
  - New stuff: the **confidence interval**.
- R: back to Fulton county, Georgia…
  - Mainly aimed at understanding the confidence interval better.
  - A little bit of new coding: C.I.s with the `t.test()` function.

universität wien

The Big Picture



universität wien

Key Terms

- **Population**: the universe of cases to which we want to generalise.
- **Sample**: a subset of units from the population we actually observe.
- **Parameter**: statistic that describes a variable in the population.
- **Estimate**: statistic that describes a variable in a sample.
- **Inference**: the process of using sample data to draw conclusions about population parameters

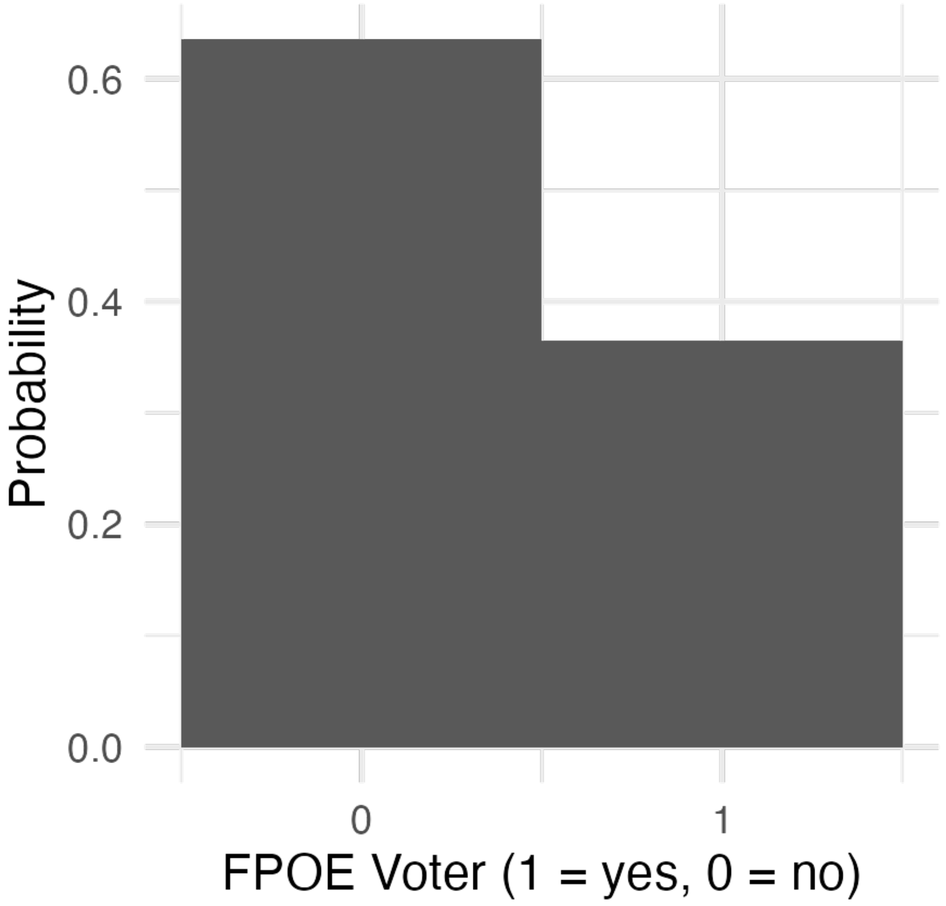
universität wien

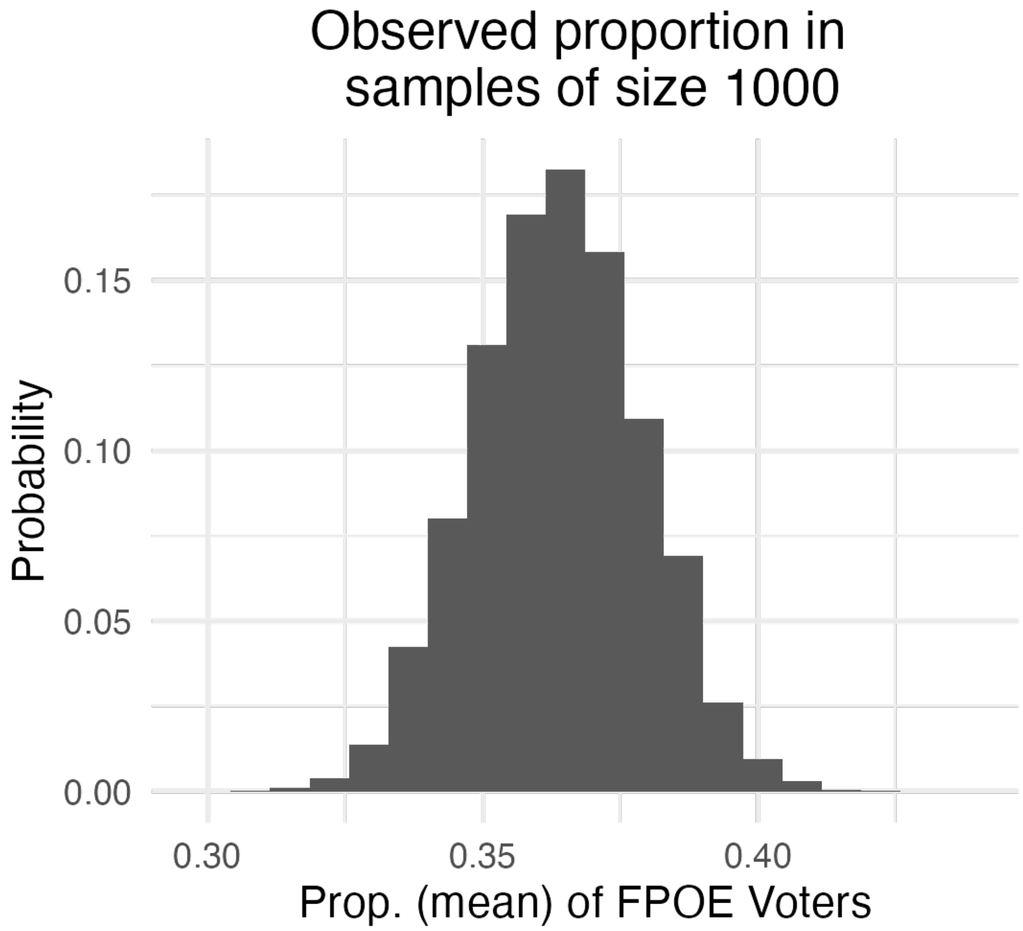
The Central Limit Theorem

- The CLT is a **probabilistic** law about repeated random sampling: it tells us that if we take lots and lots of random samples…
- …the distribution of **the sample means** (our estimate)…
    1. Approximates a **normal distribution**, and…
    2. Has for its mean the the **population mean** (our parameter)
- Provided sufficiently large sample size (conventionally, $N > 30$).

universität wien

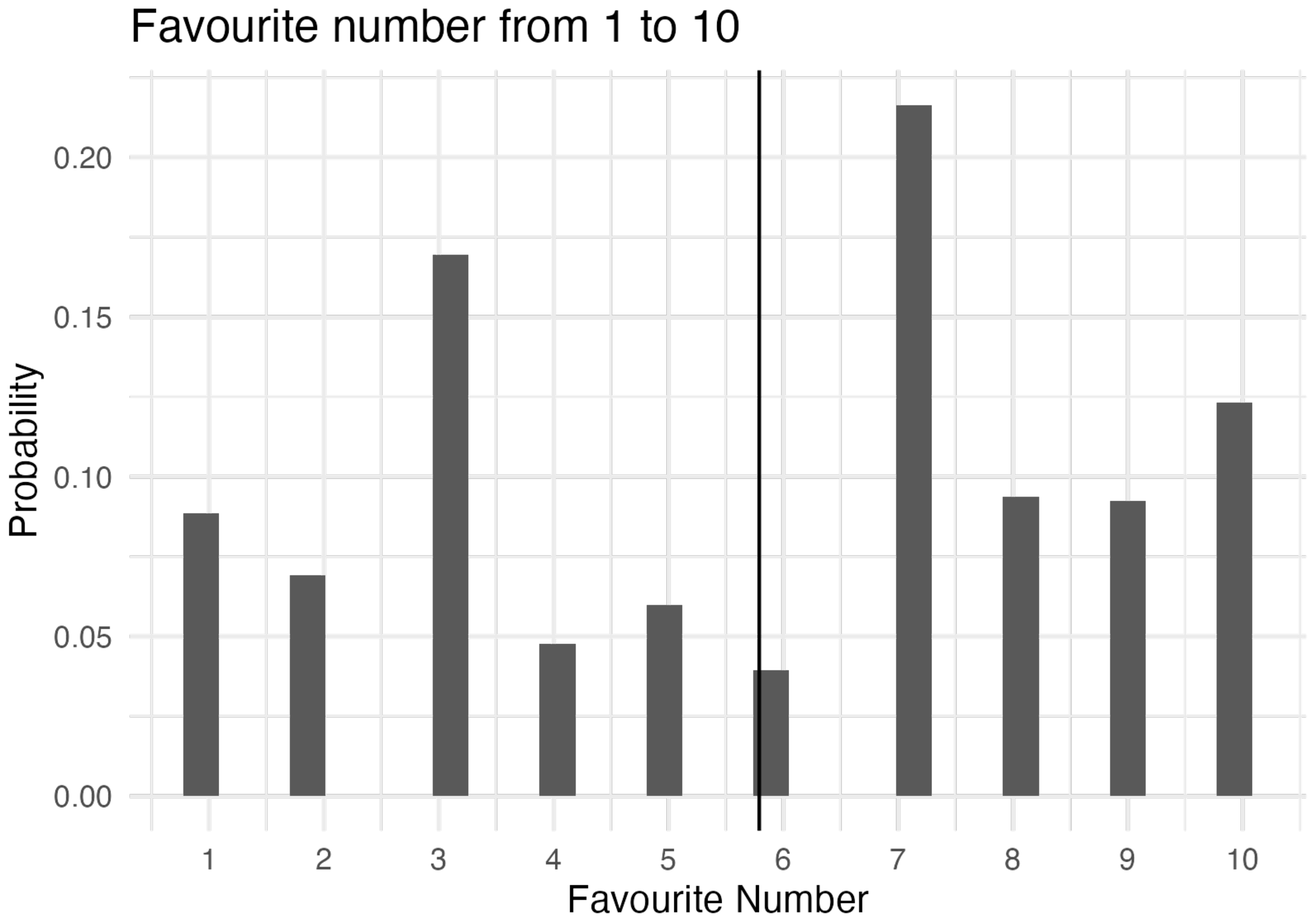
The Central Limit Theorem

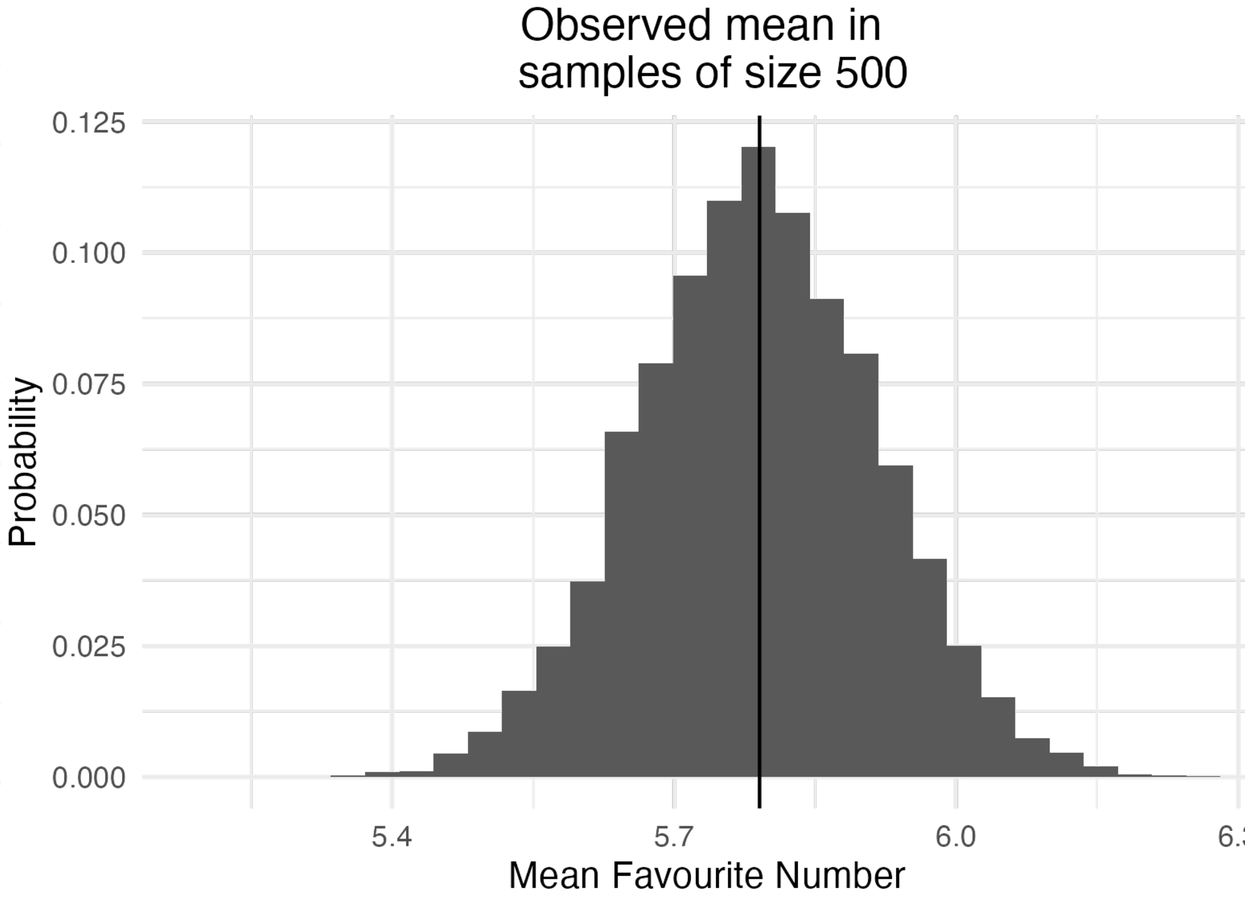




universität wien

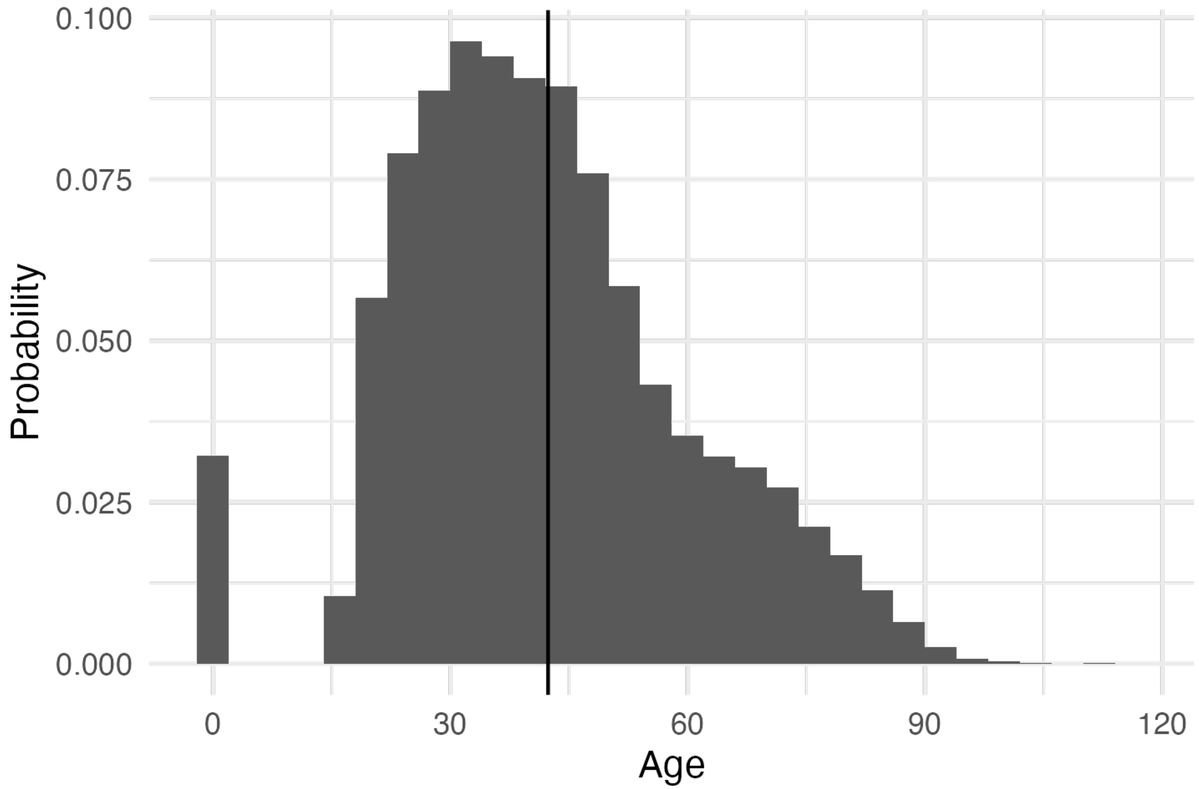
The Central Limit Theorem

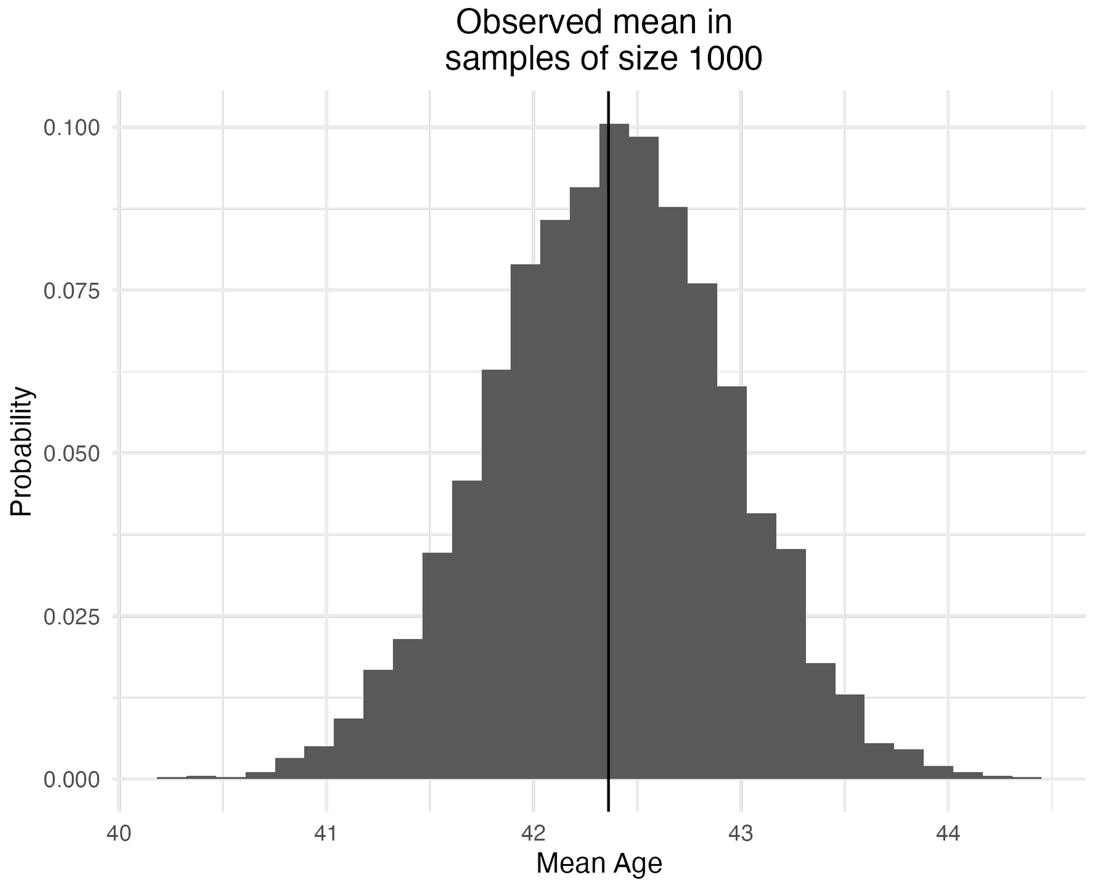




universität wien

The Central Limit Theorem

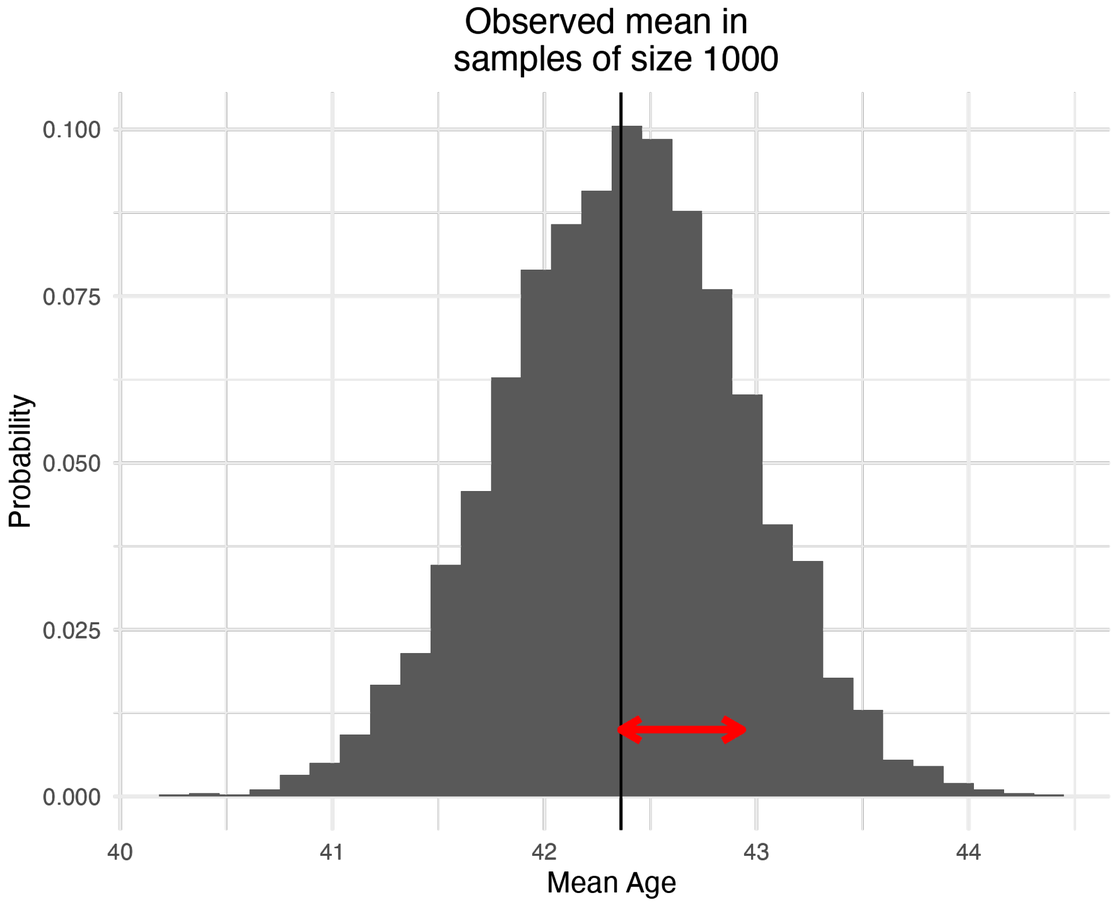





CENTRAL LIMIT THEOREM
IT'S ALWAYS NORMAL?
ALWAYS HAS BEEN.
ALL DISTRIBUTIONS

The Standard Error

The standard error is the **standard deviation** of the sampling distribution.



universität wien

The Standard Error

- We can approximate it from a **single sample** with the formula…

$$\mathrm{SE}_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

  - **Correct interpretation:**
    - "It reflects how much the **sample mean** would be expected to change if the analysis were repeated many times under identical conditions."
  - **Wrong interpretation**:
    - "It measures the spread/variability of the variable." That would be the…

universität wien

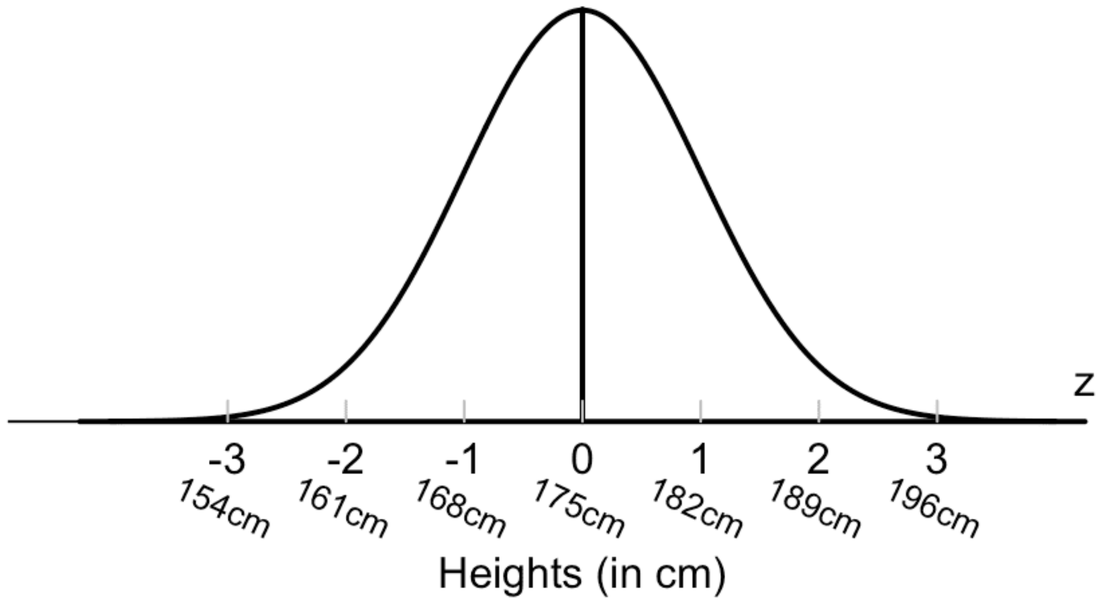
The Standard Error

- Back to the definition: the standard error is the **standard deviation** of the sampling distribution.
- Today, we'll make use of the fact that it's therefore the standard deviation of a **normal distribution** (we know this thanks to the CLT).
- But first let's remind ourselves of what a **normal distribution** is.

universität wien

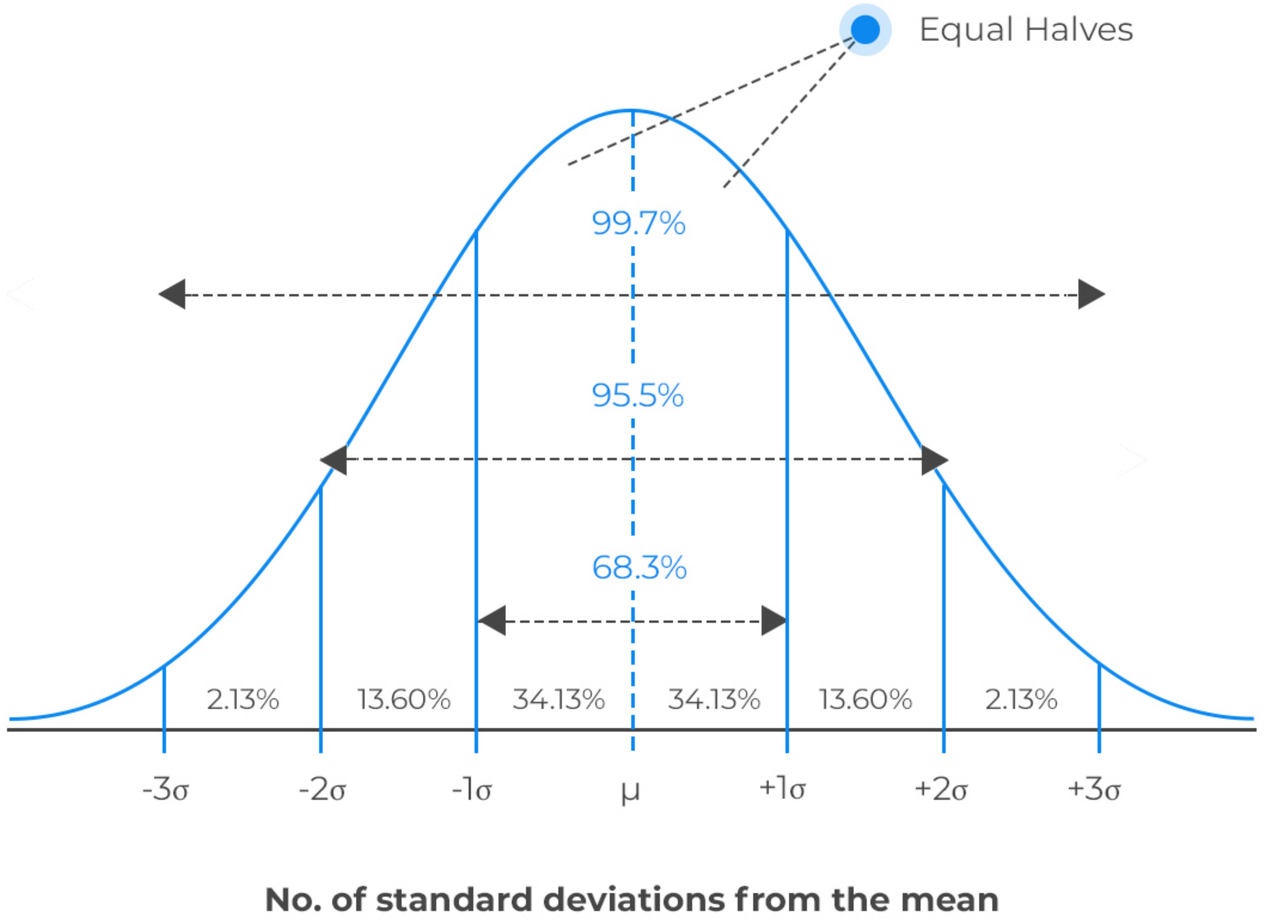
Normal Distribution

- The normal distribution is a **probability density function**: it assigns probability values to a continuum of numbers.
- It has a **single peak**, which is equal to the **mean** and the **median**.
- The distribution is **symmetric** and looks like a “bell curve”.



Normal Distribution

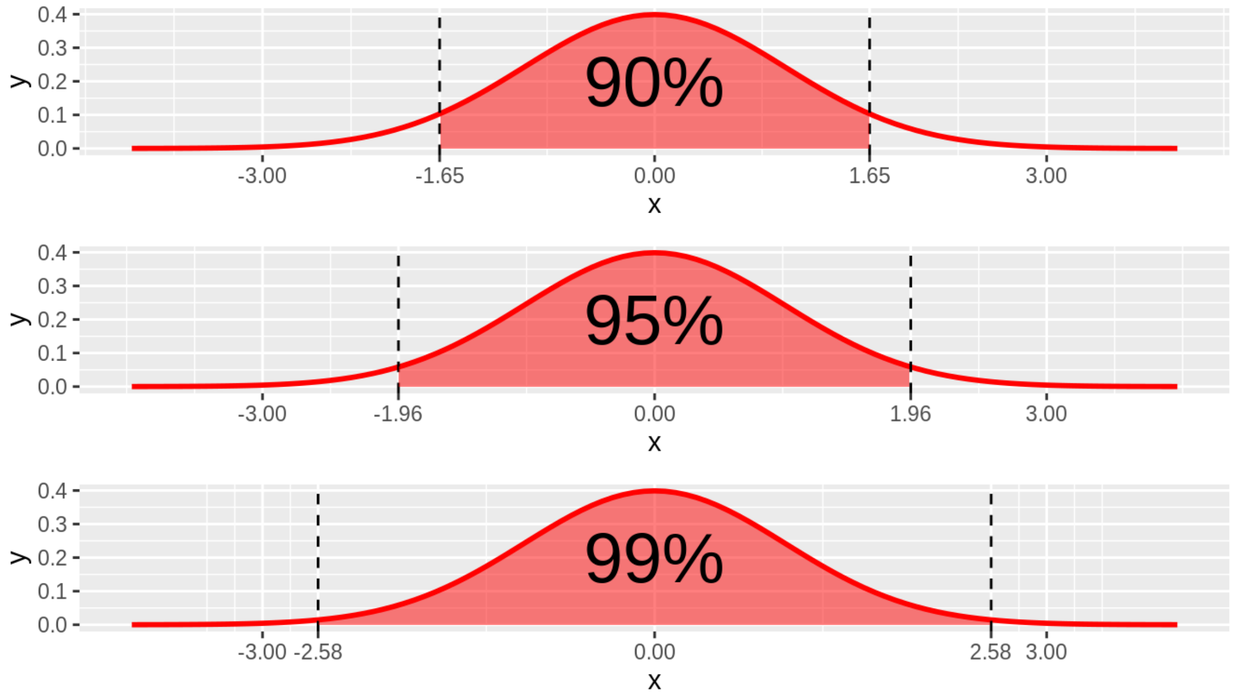
- Standard deviations of a normally distribution describe what percentage of the values fall within a certain distance of the mean (area under the curve).



universität wien

Normal Distribution

- 90% of the outcomes fall within 1.64σ of the mean.
- **95% of the outcomes fall within 1.96σ of the mean.**
- 99% of the outcomes fall within 2.58σ of the mean.
- Values known as *z*-scores.



universität wien

The Confidence Interval

- Now, if…
  1. The distribution of sample means is normal, meaning that 95% of the sample means fall within 1.96 standard deviations from the ‘peak’, and…
  2. the ‘peak’ of this distribution is the population mean, and…
  3. the standard deviation of this distribution is the standard error…
- It follows that **in 95% of the samples, the sample mean will be within 1.96 standard errors of the population mean.**

universität wien

The Confidence Interval

- The confidence interval builds on these insights by giving us a **range of values** around the sample mean that is 'likely' to contain the population mean.
- For instance, if we want a confidence interval that will contain the population mean **in 95% of samples**, we look at the interval that is 1.96 standard errors above or below the sample mean we observe:

$$\text{C.I.}_{0.95} = \bar{x} \pm 1.96 \cdot \text{SE}_{\bar{x}}$$

universität wien

The Confidence Interval

- More generally, we can build *any* confidence interval using the 'fixed' z-scores that we discussed before:

$$\text{C.I.}_{0.90} = \bar{x} \pm 1.64 \cdot \text{SE}_{\bar{x}}$$

$$\text{C.I.}_{0.99} = \bar{x} \pm 2.58 \cdot \text{SE}_{\bar{x}}$$

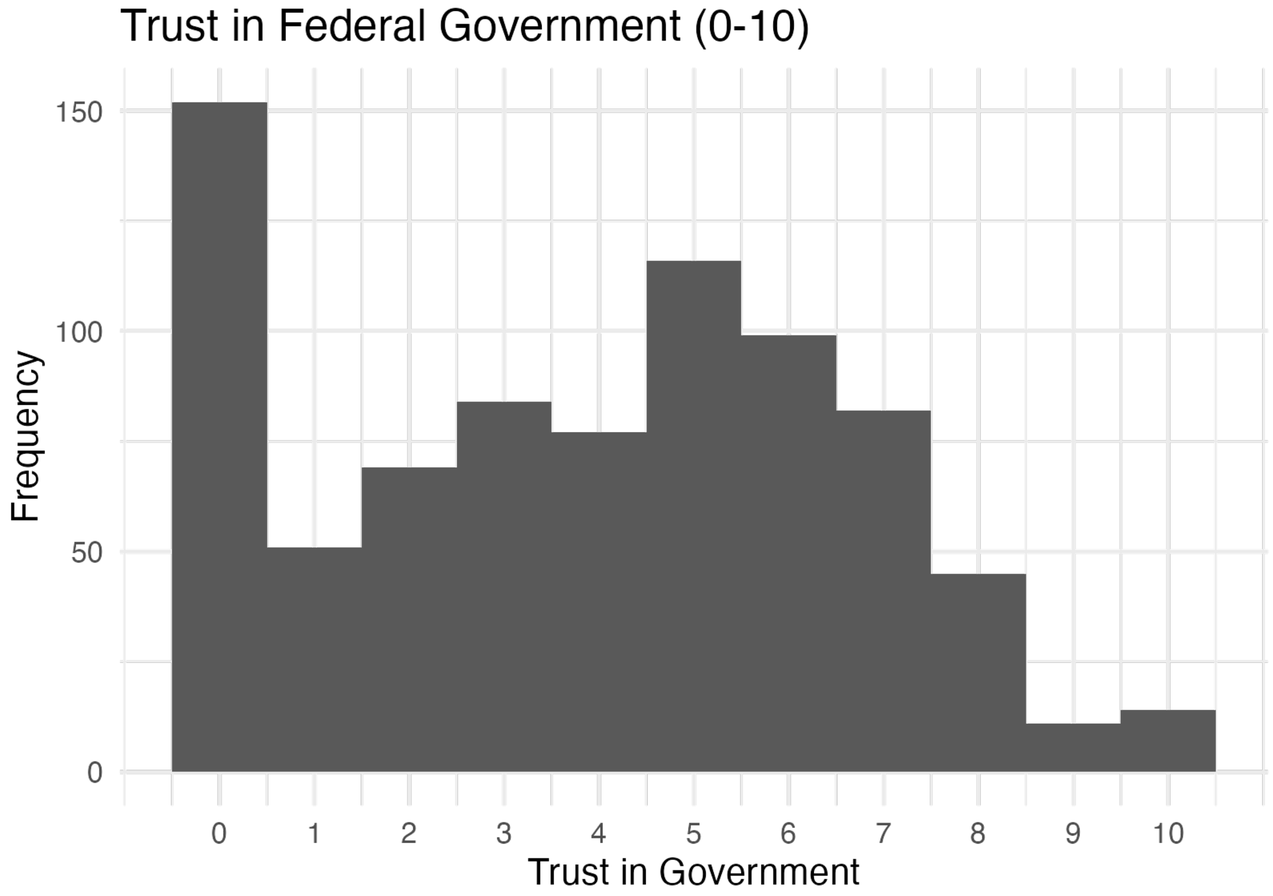
$$\text{C.I.}_{0.68} = \bar{x} \pm 1 \cdot \text{SE}_{\bar{x}}$$

- In practice, we most often use the 95% confidence interval ($z = 1.96$).

universität wien

Confidence Interval, in Action

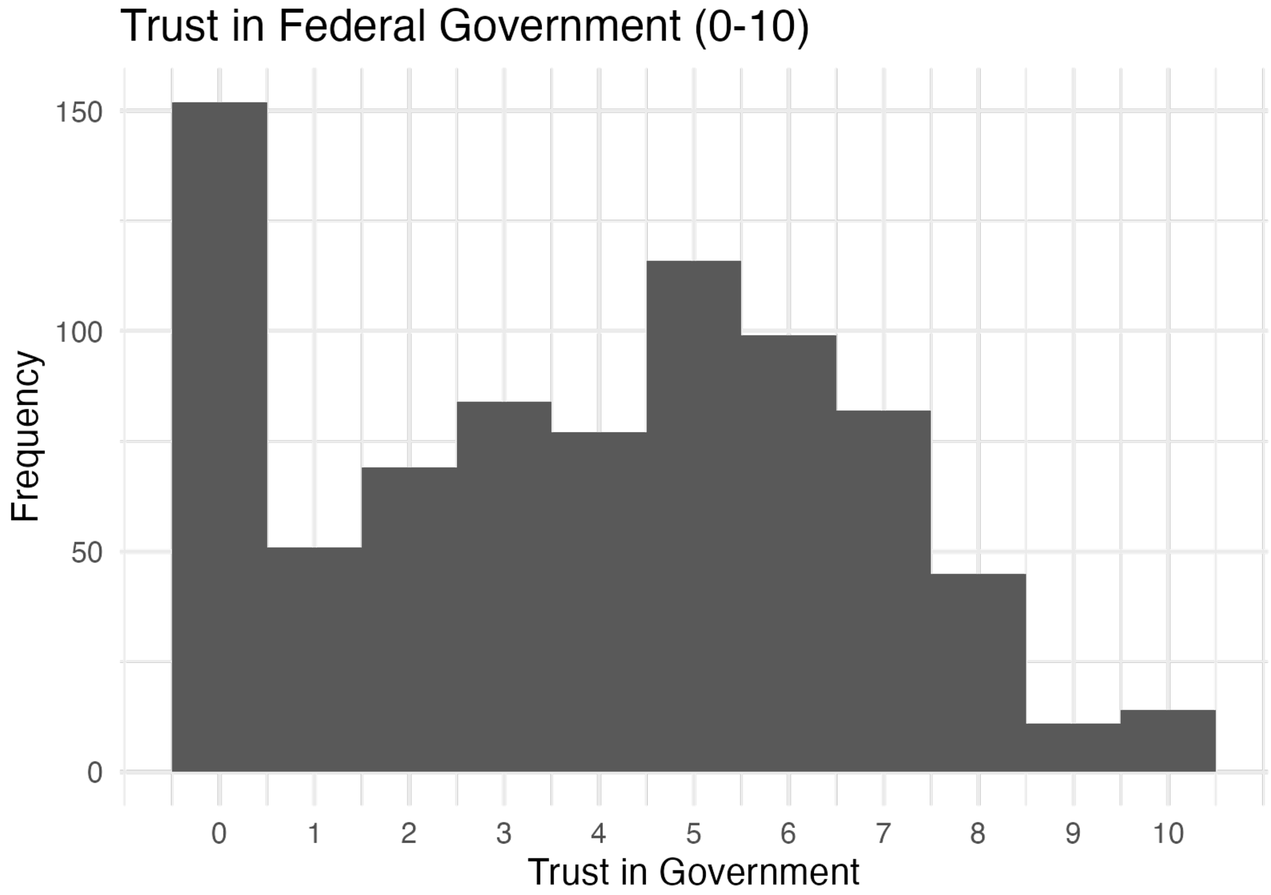
- I have a **random** sample of 800 Austrian adults, and I ask them how much they trust the government on a scale from 0 to 10.
- I obtain a sample mean of **3.87** and a sample standard deviation of **2.74**.



Confidence Interval, in Action

- With my sample standard deviation of 2.74 and sample size of 800, I can compute the **standard error**…

$$\text{SE}_{\bar{x}} \approx \frac{2.74}{\sqrt{800}} = 0.097$$



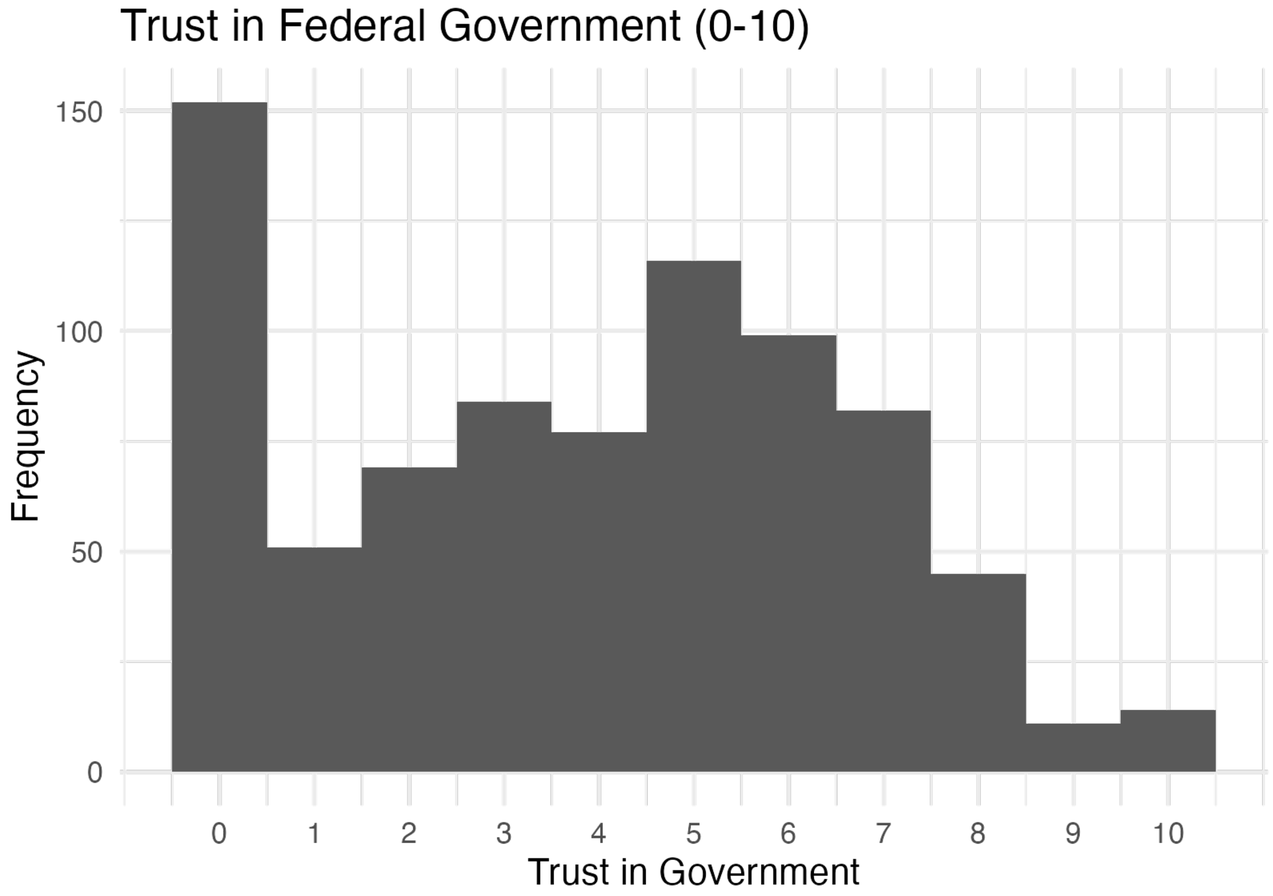
Confidence Interval, in Action

- With the standard error (0.097) and the sample mean (3.87), I can compute the confidence interval…

$$\text{C.I.}_{0.95} = 3.87 \pm 1.96 \cdot 0.097 =$$

$$[3.68,\ 4.06]$$

- Where 3.68 is the ‘lower bound’ and 4.06 is the ‘upper bound’ of the CI.



universität wien

Confidence Interval, in Action

- ***Reporting the results:***
  - The mean trust in government is 3.87 (95% CI [3.68, 4.06]).
- ***Correct interpretations:***
  - If we sampled from the population many times, **95% of the confidence intervals constructed this way** would contain the true population mean."
  - "I am **95% confident** that the population mean is within this interval."

universität wien

Confidence Interval, in Action

- ***Wrong interpretation***: “There is a 95% probability that the true mean lies in this specific interval.”
- It’s almost correct, but **still wrong**, because…
  - the population mean doesn’t vary across samples, it’s fixed, so the probability that it’s in the interval is either 0 or 1…
  - …and because in different samples you will get different confidence intervals (the sample mean and approximate SE vary).

universität wien

Confidence Interval, in Action

- Imagine you magically knew that the ‘true’ trust in government among all Austrian adults is **exactly 4**.
- Now imagine we have tons of research money and so we can take lots and lots of samples of size 800 from the population, and compute the mean and the standard error of the sample mean….

universität wien

Confidence Interval, in Action

- Imagine you magically knew that the ‘true’ trust in government among all Austrian adults is **exactly 4**.
- now we take lots and lots of samples of size 800 from the population…

Sample	Mean	S.E.	95% C.I.	4 is in CI?
1	4.09	0.1023	[3.89–4.29]	YES
2	3.97	0.1012	[3.77–4.17]	YES
3	4.08	0.0980	[3.89–4.27]	YES
4	3.94	0.1000	[3.75–4.14]	YES
5	4.04	0.0977	[3.85–4.23]	YES
6	3.96	0.0986	[3.77–4.16]	YES
7	4.08	0.1031	[3.87–4.28	YES
8	4.19	0.0994	[4.00–4.39]	YES
9	3.88	0.1010	[3.68–4.08	YES
10	4.05	0.1000	[3.86–4.25	YES

Sample	Mean	S.E.	95% C.I.	4 is in CI?
11	4.05	0.0995	[3.86—4.25]	YES
12	4.18	0,1011	[3.98–4.38]	YES
13	3.84	0.1012	[3.65–4.04]	YES
14	**3.79**	**0.0995**	**[3.60–3.99]**	**NO**
15	4.02	0.0987	[3.83–4.21]	YES
16	4.00	0.0988	[3.81–4.19]	YES
17	4.02	0.1023	[3.82–4.22]	YES
18	3.93	0.0991	[3.73–4.12	YES
19	3.93	0.0972	[3.73–4.12]	YES
20	3.94	0.0987	[3.74–4.13]	YES

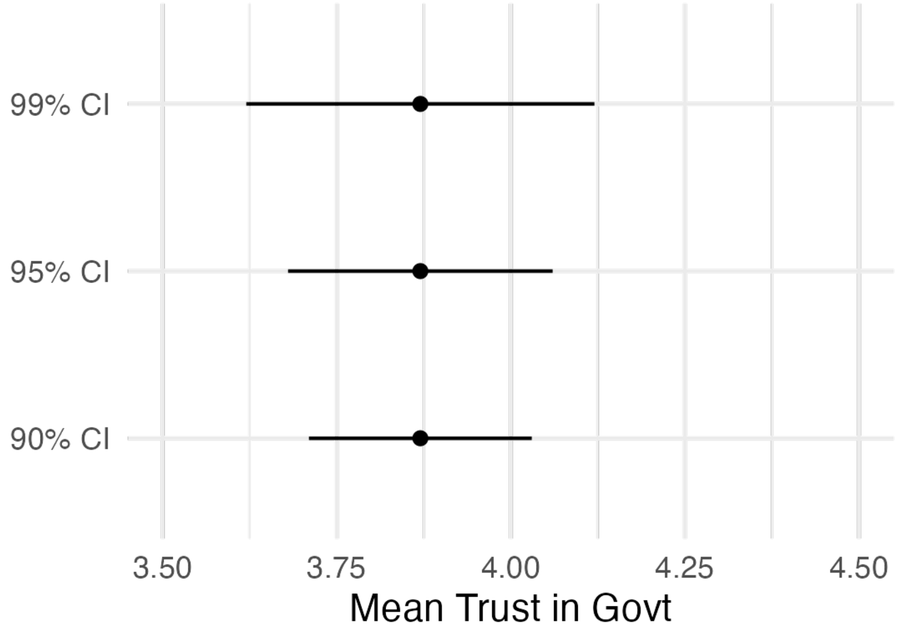
universität wien

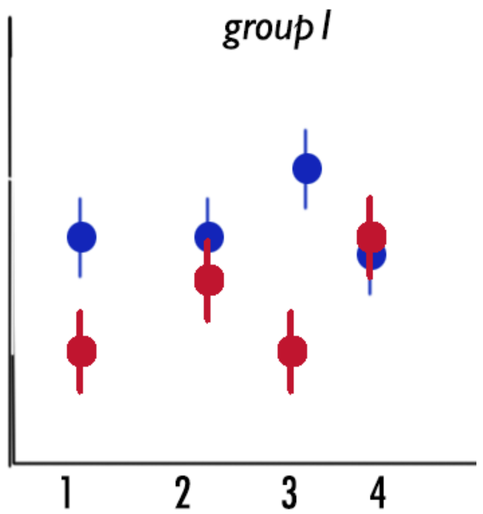
Confidence Interval: Recap

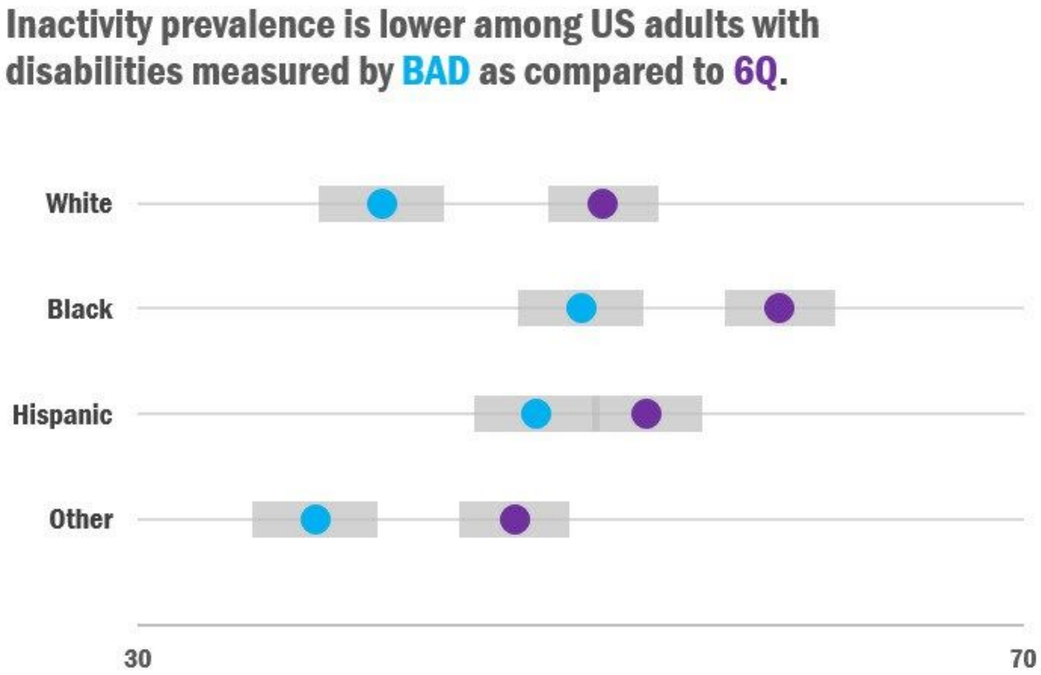
- A confidence interval gives a range of plausible values for a population parameter (e.g., a population mean), based on a sample estimate (e.g. a sample mean).
- It expresses the uncertainty around the sample estimate. The wider the confidence interval, the less precise our inference from the sample to the population.
- It's computed **from the sample mean and standard error**, as C.I. $= \bar{x} \pm z \cdot \text{SE}_{\bar{x}}$ where z is a fixed value (z-score) associated with a certain confidence level. For instance, for the 95% confidence interval, the C.I.$_{0.95} = \bar{x} \pm 1.96 \cdot \text{SE}_{\bar{x}}$.

Dot-and-Whisker Plots

- Point estimates and confidence intervals are usually visualised like this…







universität wien

A few more things…

- Confidence interval of a **sample proportion**: normally, it's fine to treat a proportion as a mean, which represents the proportion of 1s.
- ***BUT*** when you have small samples or the proportion is close to 1 (100%) or to 0 (0%), you may get a confidence interval with negative lower bounds or upper bounds above zero.
- Adjusted intervals (R function `prop.test`) account for this. Most of the times, you don't need to worry about this.

universität wien

A few more things…

- We can now compute the standard error and the confidence interval for sample means and proportions.
- In the next classes, we will also learn about the SE and CI of two other estimators: difference-in-means and regression coefficients.
- ***BUT*** you could theoretically get these statistics for other sample estimates: medians, first quartile, standard deviations etc. Some of these have no ‘clean’ formulas for the standard error though.
- One powerful, computation-based method for all these things: the **bootstrap**. Beyond the scope of this class. Just be aware you can do it.

universität wien

Last but not Least

- Next class: hypothesis testing (*yuck*).
- We'll be back to work in RStudio. I also want you to learn how to work with data you find 'in the wild', so no dataset will be provided.
- I will post instructions on how to download the AUTNES Online Panel Study, a survey of the Austrian National Election Study (AUTNES).
- So keep an eye on Moodle and try to download this before class.
- And now, let's go back to Fulton county, Georgia…