

# Lecture 2: Statistical Inference & Uncertainty

Paul Röttger

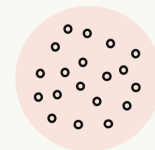
Applied Analytical Statistics

27<sup>th</sup> of January 2026

# Plan for today | Statistical inference and uncertainty

Today we move **from description to inference**

We go beyond describing our sample towards learning about populations.



1. **Statistics as random variables:** the i.i.d. assumption and sampling distributions
2. **Large sample theory:** the Law of Large Numbers and the Central Limit Theorem
3. **Quantifying uncertainty:** standard errors and probability statements about statistics.
4. **Analytical inference:** construction and interpretation of confidence intervals
5. **Computational inference:** bootstrap confidence intervals, pros and cons



We will finish with a **class activity** to help you prepare for your summative.

# Inductive inference | Parameters and statistics

Goal of inductive inference: use data from a **sample** to learn about an unknown **population**.

**RQ:** What proportion of **UK adults** uses AI chat assistants at least once a week?

**Data:** Self-reported AI usage data from a representative survey of **1,000 UK adults**.

The population is described by fixed but unknown **parameters**.

Examples: **population mean**, **population proportion**, **population variance**

$\mu$

$p$

$\sigma^2$



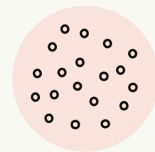
Based on our sample, we compute **statistics** to estimate the population parameters.

Examples: **sample mean**, **sample proportion**, **sample variance**

$\bar{x}$

$\hat{p}$

$s^2$



# Inductive inference | Statistics as random variables

Last week, we learned about **random variables**, which take a numerical value for each possible outcome in the sample space (e.g. a coin, with heads/tails mapped to 1/0).

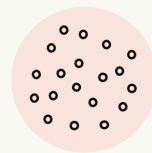
A **sample** (i.e. a specific set of observations) is the result of a random sampling process. Different samples result in different values of our statistic (e.g. the sample mean).

**Data:** Self-reported AI usage data from a representative survey of 1,000 UK adults.

Therefore, **a statistic is itself a random variable**.

The sample proportion  $\hat{p}$  varies across samples.

The population proportion  $p$  is fixed but unknown.



# Inductive inference | The i.i.d. assumption

To reason about samples and populations, we typically have to assume that our data is **independent and identically distributed**  $\rightarrow$  i.i.d.

**Independent:** Each observation does not influence or inform any other observation.

**Identically distributed:** All observations are drawn from the same population distribution.

Formally, when  $X_1, X_2, \dots, X_n$  are i.i.d. draws from the same population, then:

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n P(X_i \in A_i)$$

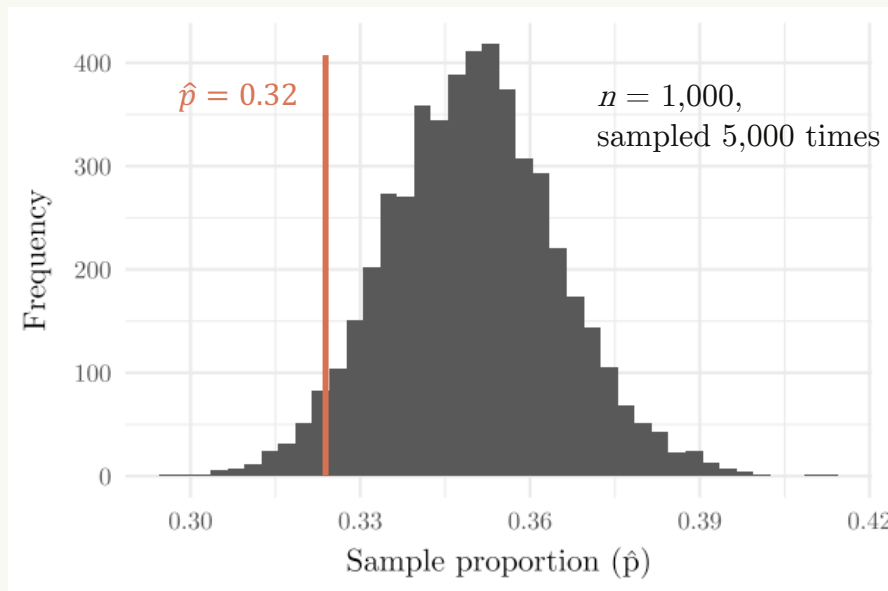
and  $X_i \sim F$  for all  $i$



What kinds of data would likely violate these assumptions?

Without (approximate) i.i.d. assumptions, standard statistical inference becomes unreliable!

# Sampling distributions | Linking samples to populations



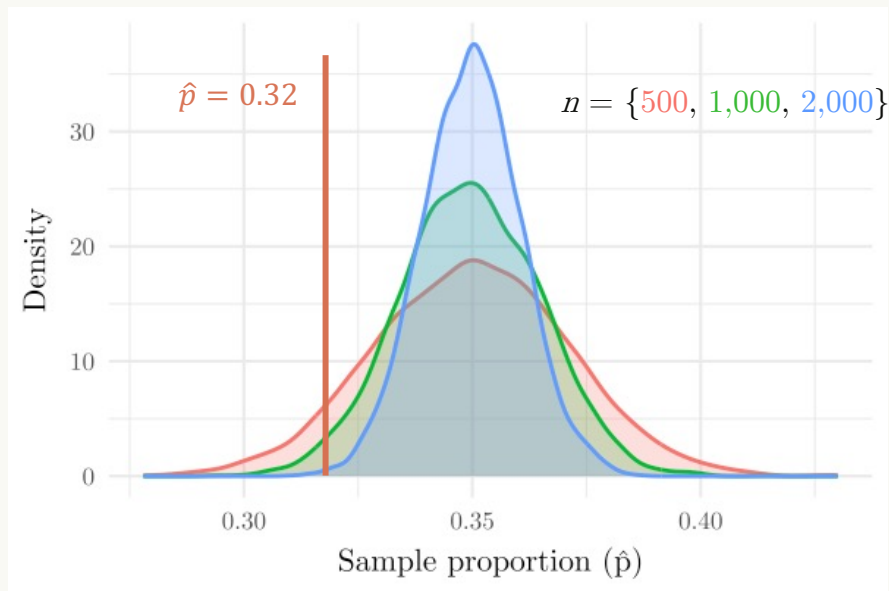
The **sampling distribution** is the distribution of a **statistic** across repeated i.i.d. samples from the same population.

This is a **hypothetical construct**. We do not usually observe this distribution, just one sample = **one statistic**.



How do we expect the shape of this distribution to change depending on sample size?

# Sampling distributions | Variability and sample size



The **larger** our sample, the **less variability** there is in our sampling distribution.



How does the probability of observing  $\hat{p} = 0.32$  change as we increase sample size?

Larger samples make it more likely that the statistic we observe is **closer to the centre of the sampling distribution**.

# Large sample theory | Two key results for enabling inference

**Random sampling introduces variability in our data.** For sampling distributions, we saw that different samples from the same population produce different statistics.

The **Law of Large Numbers** (LLN) and **Central Limit Theorem** (CLT) describe what happens to this variability **as sample size grows large**, assuming i.i.d. samples.

→ **asymptotic behaviour**, i.e. what happens as  $n \rightarrow \infty$

Most of classical statistical inference relies on these two results!

They help us today because they allow us to **approximate sampling distributions**.



# Law of Large Numbers | Formula and intuition

Let  $X_1, X_2, \dots, X_n$  be i.i.d. draws from the same population,

where the population mean  $\mu = \mathbb{E}[X]$

and the sample mean  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , meaning that  $\bar{X}_n$  is a random variable.

Then the **Law of Large Numbers** (LLN) states that  $\bar{X}_n \rightarrow \mu$  when  $n \rightarrow \infty$

## Intuition:

- As sample size increases, the sample mean converges to the population mean.
- Random sampling variability averages out in large samples.



Now we are suddenly talking about means. Does the LLN apply to sample proportions?

# Law of Large Numbers | Proportions as means

A **sample proportion** is just a sample mean of a Bernoulli random variable.

Let the random variable  $X_i = \begin{cases} 1 & \text{if some specified event occurs} \\ 0 & \text{otherwise} \end{cases}$

Then  $X_i \sim \text{Bernoulli}(p)$  and  $\mathbb{E}[X_i] = p$ , where  $p$  is the probability of the event.

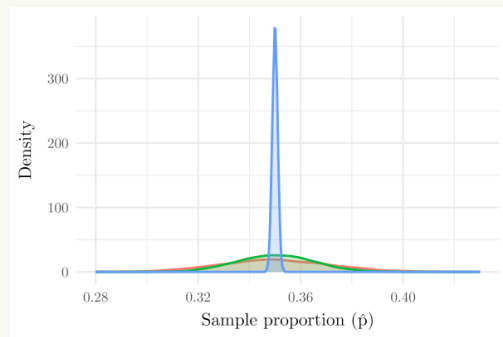
Then the sample proportion  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$  is exactly the sample mean of the observed  $X_i$ 's.

Therefore, the LLN applies to sample proportions!

# Law of Large Numbers | Convergence is key

The LLN tells us that **sample means converge to the true population mean**.

Estimates become **stable** as sample size increases, meaning that  $P(|\bar{X}_n - \mu|) \rightarrow 0$  as  $n \rightarrow \infty$ . Large deviations between our statistic and the true parameter become increasingly unlikely.



For large  $n$ , our sampling distribution is extremely narrow, centered around the true population mean (in this case  $p$ ).

However, the LLN **does not tell us how much variability remains for finite samples**, i.e. the data we are working with.

We need to describe the **shape of the sampling distribution** to quantify uncertainty.

# Central Limit Theorem | Formula and intuition

Let  $X_1, X_2, \dots, X_n$  be i.i.d. draws from the same population, where the population mean  $\mu = \mathbb{E}[X]$  and population variance  $\sigma^2 = \text{Var}(X)$ .

As before, the sample mean  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  is a random variable.

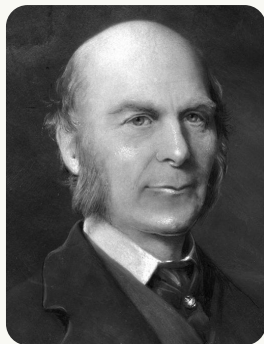
Then the **Central Limit Theorem** (CLT) states that  $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$  as  $n \rightarrow \infty$

Equivalently, for large  $n$ , we can say that  $\bar{X}_n \approx N(\mu, \frac{\sigma^2}{n})$

## Intuition:

- For large  $n$ , the sampling distribution of the sample mean is approximately normal.
- The sampling distribution is centered at the true population mean  $\mu$ .
- The sampling distribution is narrower for larger samples, with a spread of  $\sigma^2/n$ .

# Central Limit Theorem | An amazing law of nature



Francis Galton



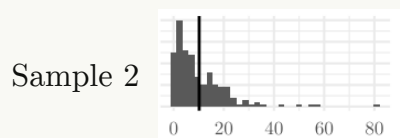
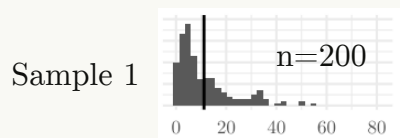
Godfather of modern mathematical statistics  
but also **originator of eugenics** (further reading [here](#)).

I know of scarcely anything so apt to impress the imagination as the **wonderful form of cosmic order expressed by the Central Limit Theorem**. The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the **supreme law of Unreason**.

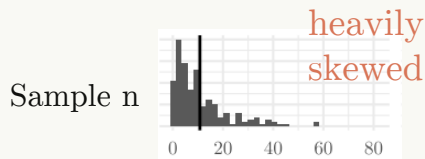
# Central Limit Theorem | The great normaliser

Notice that we did not make many assumptions for the CLT. Most importantly:

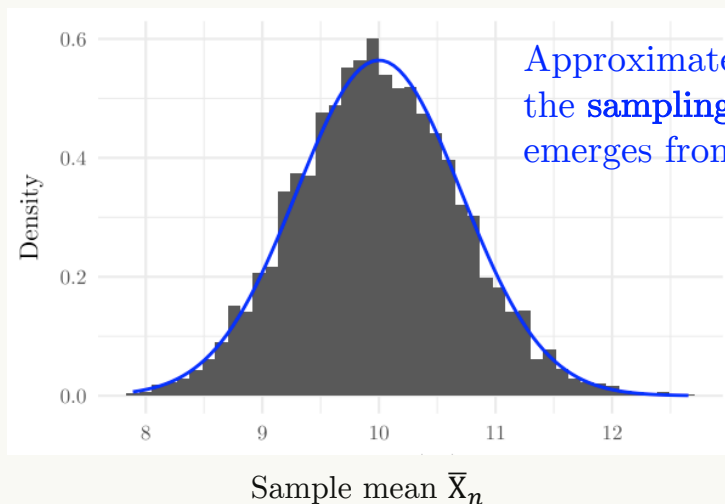
**The CLT does not require the sample itself to be normally distributed.**



...



X



# Central Limit Theorem | Approximate normality

In some cases, we know the exact sampling distribution of a statistic.

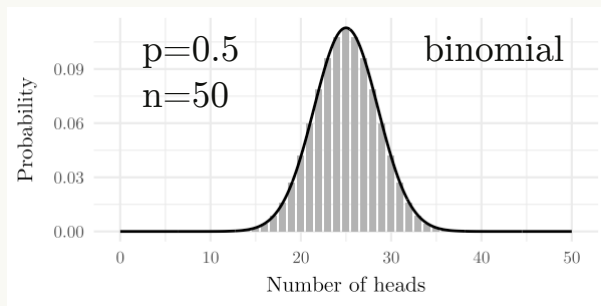


**Example:** Number of heads in a sequence of  $n$  independent coin tosses.



What distribution does this statistic follow?

The larger  $n$ , the better the normal distribution approximates the sampling distribution.



There is **no universal rule** for what  $n$  is large enough so that we can use the CLT approximation.

For sample means, even  $n \geq 40$  can be large enough.  
For sample proportions, a rule of thumb is  $np \geq 10$ .

# Standard errors | Quantifying sampling uncertainty

The standard error SE is the standard deviation of the sampling distribution of a statistic.

↖  
specific to the sampling distribution

↖  
general measure of variability

Standard errors are defined by **population parameters** and **sample size**:

By CLT:  $\bar{X}_n \approx N(\mu, \frac{\sigma^2}{n})$

For the sampling distribution of the **sample mean**, we have  $SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$  →

For the **sample proportion**, we have  $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$  🤔 Why is this the case?

In practice, we **replace unknown population parameters** by their **sample estimates** (e.g.  $\hat{p}$ ).  
As we will see, this is **justified for large samples**.



## Standard errors | Exercise: SE of sample proportion

For a Bernoulli random variable:  $\text{Var}(X_i) = p(1 - p)$

For independent random variables:  $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$

For any constant  $a$ :  $\text{Var}(aY) = a^2 \text{Var}(Y)$

Use the above to show that  $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

$$SE(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i)} = \sqrt{\frac{1}{n^2} np(1-p)} = \sqrt{\frac{p(1-p)}{n}}$$

# Standard errors | Estimating population standard deviation

To compute standard errors, we need the **population standard deviation**  $\sigma$ .

In practice,  $\sigma$  is unknown and must be estimated from the sample.

$$SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

The **sample standard deviation** is defined as  $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$

Compare this with the **population standard deviation**  $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2}$

The key difference is the **bias correction term**, dividing by  $n - 1$  instead of  $n$ .

We make this correction because the sample mean  $\bar{X}_n$  is itself an estimate for  $\mu$ .

The naïve variance estimate (dividing by  $n$ ) would be too small.



See: consistency of sample variance

However, for large samples,  $s \rightarrow \sigma$ , and the correction becomes negligible.

# Sampling distributions | Probability statements about statistics

We now know, for any statistic, for large  $n$  and i.i.d. samples:

- The **shape** of the sampling distribution is approximately normal, by CLT
- The **center** of the sampling distribution is the true parameter, by LLN
- The **spread** of the sampling distribution is the standard error

We can use this knowledge to make **probability statements about statistics**.

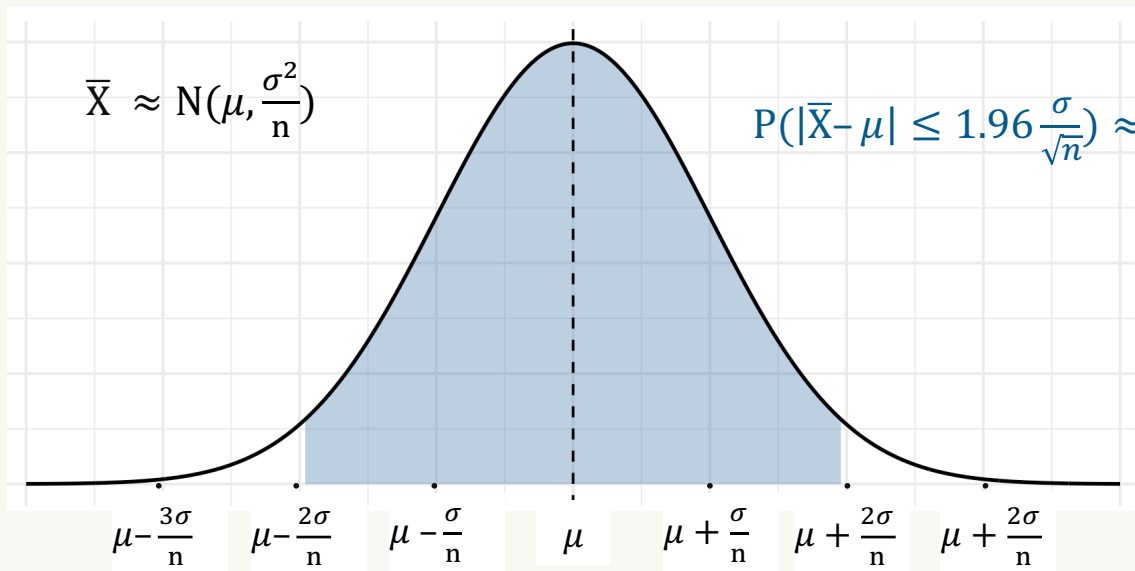
In other words: The sampling distribution tells us how a statistic varies across samples.

We can approximate the sampling distribution based on the assumptions above.

This allows us to **quantify how unusual any observed value of the statistic would be**.

# Sampling distributions | Probability of a sample mean

In 95% of repeated samples,  $\bar{x}$  falls within  $\pm 1.96$  standard errors of the true  $\mu$ .



Where do we get the 1.96 from?

We can standardise the normal distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

In a standard normal:  
 $P(|Z| \leq 1.96) \approx 0.95$

# Confidence intervals | Inverting the sampling distribution

Problem: We **cannot observe** the sampling distribution across all  $\bar{X}$ . We **do not know**  $\mu$ .

Solution: We **can observe** one realisation of the statistic  $\bar{x}$ .

By the CLT, we can make **claims about the probability of statistics**  $\bar{X}$ :

$$P(|\bar{X} - \mu| \leq 1.96 \cdot SE(\bar{X})) \approx 0.95 \text{ where } SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

We can now invert this statement to make **claims about the population mean**  $\mu$ :

$$P(\mu \in [\bar{X} \pm 1.96 \cdot SE(\bar{X})]) \approx 0.95$$



What is the verbal interpretation of this equation?

# Confidence intervals | Definition

**Definition:** A 95% confidence interval (CI) is a **procedure** that, in repeated sampling, produces intervals that contain the true population parameter 95% of the time.

CIs are random because they depend on the observed statistic.  
The population parameter is fixed but unknown.

In general, for large  $n$ :  $[\text{estimate} \pm (\text{critical value}) \cdot (\text{standard error})]$

For sample means:  $\left[ \bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}} \right]$  where  $1.96 = z_{0.975}$

For sample proportions:  $\left[ \hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$  where  $1.96 = z_{0.975}$

# Confidence intervals | Interpretation

**Definition:** A 95% confidence interval (CI) is a **procedure** that, in repeated sampling, produces intervals that contain the true population parameter 95% of the time.

A CI that we construct based on a specific sample either contains the parameter or not.

✗ ~~“There is a 95% probability the true parameter lies in this interval.”~~

✓ “If we repeated the study many times, 95% of the CIs would contain the true parameter.”

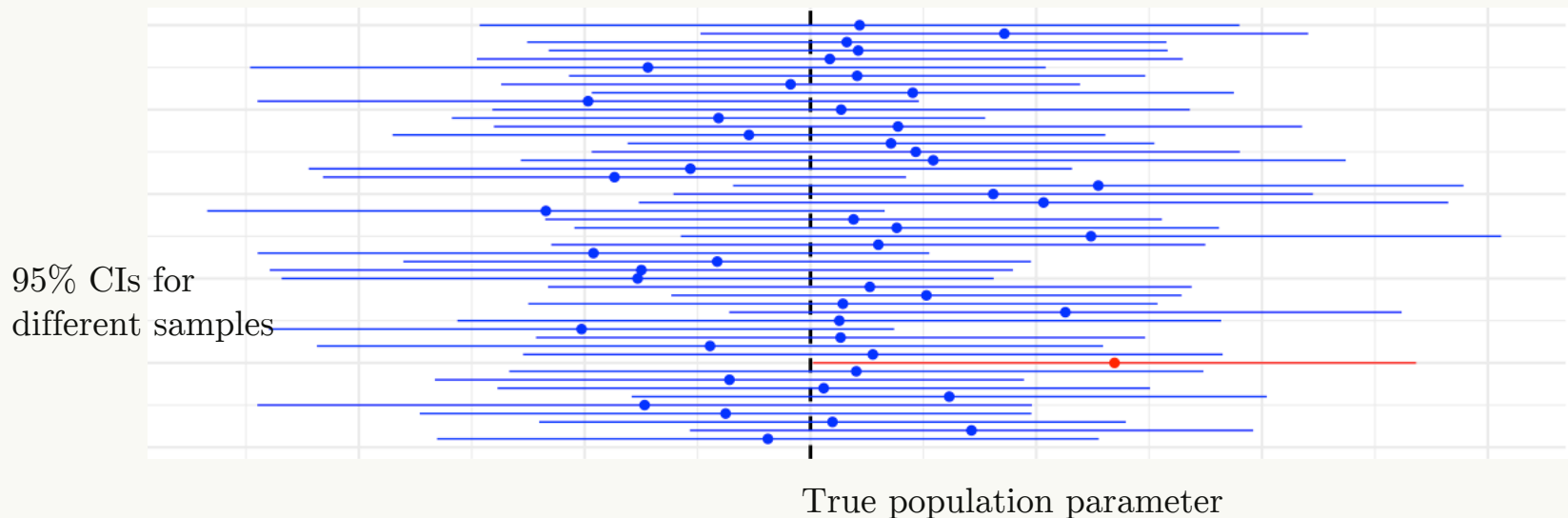
✗ ~~“95% of the data lie in this interval.”~~

✓ “CIs describe uncertainty about point estimates, not dispersion of observations.”

# Confidence intervals | Coverage

The **coverage** of a confidence interval describes the long-run proportion of times that the **confidence interval procedure** contains the true population parameter under repeated sampling.

→ A 95% CI has 95% coverage





# Confidence intervals | Scaling

Let's say we construct a 95% CI for a sample mean  $\left[\bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}\right]$



How does CI width scale with sample size  $n$ ?

CI width  $\propto \frac{1}{\sqrt{n}}$ . Diminishing returns! 4x the sample size = 0.5x CI width



How does CI width scale with sample standard deviation  $s$ ?

CI width  $\propto s$ . Linear returns! 2x the variability = 2x uncertainty about your estimate.

# Transition | From analytical to computational inference

So far, we relied on **analytical inference**, grounded in probability theory, LLN and CLT.

This **works well** when:

- Sample sizes are large
- Sampling distributions are approximately normal
- Standard errors have simple, known formulas

However, **analytical SEs and CIs become difficult** when:

- Sample sizes are small or moderate
- Sampling distributions are skewed or heavy-tailed
- Statistics are complex (e.g. medians, quantiles, complex estimators)

In these cases, theoretical approximations may be unreliable.

Therefore, we now introduce a **computational alternative** that requires **fewer assumptions**.

# Bootstrap | Intuition

Goal: Approximate the **sampling distribution** of a statistic (e.g. sample mean)  
Challenge: We **cannot repeatedly sample** from the population.

→ Same goal and challenge as before!



- We treat our sample as a **stand-in** for the population.
- We repeatedly draw samples **with replacement**.
- We **recompute the statistic** for each sample.
- We use the resulting “**bootstrap distribution**” of the statistic as a stand-in for the unobserved sampling distribution.

We no longer assume a theoretical distribution but estimate it directly from the data!

# Bootstrap | Algorithm

Let  $X_1, X_2, \dots, X_n$  be an observed random sample from an unknown population, and let  $\hat{\theta} = T(X_1, \dots, X_n)$  be a statistic estimating a population parameter  $\theta$ .

Define a **bootstrap sample**  $X_1^*, \dots, X_n^*$  as a sample drawn **with replacement** from  $\{X_1, \dots, X_n\}$

Let  $\hat{\theta}^* = T(X_1^*, \dots, X_n^*)$  be the estimated statistic for a given bootstrap sample.

Repeat the sampling procedure  $B$  times to obtain  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ .

Then the empirical “bootstrap distribution” of  $\hat{\theta}^*$  approximates the sampling distribution of  $\hat{\theta}$ .

# Bootstrap | Confidence intervals

Once we have the bootstrap distribution, we can use it to construct confidence intervals.

The most common approach is the **percentile bootstrap CI**:

For a given sample, the 95% bootstrap confidence interval is the interval between the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentile of the bootstrap distribution.

The **interpretation** of bootstrap CIs is the **same as for analytical CIs**:

A 95% bootstrap CI is a procedure that, in repeated sampling, produces intervals that contain the true parameter 95% of the time.



Are bootstrap CIs symmetric?

# Bootstrap | Strength = flexibility

## Minimal distributional assumptions.

The bootstrap does not require normality of the sampling distribution, closed-form variance formulas, or large-sample approximations (e.g. CLT).

## Applicable to many statistics.

The bootstrap works even for complex or nonlinear statistics (e.g. medians, quantiles), where analytical approximations of the sampling distribution are difficult to come by.

## Naturally capturing skewness and asymmetry.

Bootstrap distributions can be asymmetric, skewed, reflecting whatever shape is implied by the sample data, whereas analytical CIs often assume symmetry around the point estimate.

# Bootstrap | Limitation = sample quality

## Observations must be independent from each other.

This assumption fails for time series, network data, clustered or panel data...

Regular bootstrapping would severely underestimate uncertainty for dependent data.

## The sample must be sufficiently informative.

Since resampling produces many duplicate observations, the bootstrap distribution can be unstable with very small or imbalanced samples, especially if there are outliers.

## The sample must be representative.

The observed sample is a stand-in for the population. If the sample is biased, unrepresentative, or systematically distorted, **the bootstrap will reproduce this bias.**

(+ practical limitation: computational costs can be high for complex statistics)

# Comparison | Analytical vs. bootstrap inference

	Analytical inference	Bootstrap inference
Sampling assumption	i.i.d. sample	i.i.d. sample
Distributional assumption	Sampling distribution $\approx$ normal distribution (CLT)	Sample distribution $\approx$ population distribution
Required sample size	Large	Moderate
Sensitivity to skewness	High	Lower
Sensitivity to bias	High	High
Computational cost	Lower	Higher

**Analytical inference** is fast, elegant, theory-grounded. Best when assumptions clearly hold.  
**Bootstrap inference** is robust and flexible. Best when sampling distributions are complex.

**Neither method fixes bad data or bad research design!**



## Recap | Key takeaways from week 2

**Statistical inference = learning from samples under uncertainty.**

Statistics are random variables and estimates vary across samples from the same population.

**The sampling distribution links samples and populations.**

We approximate its shape to quantify uncertainty in our sample-specific estimates.

**Large-sample theory lets us quantify uncertainty analytically.**

For large  $n$ , LLN explains convergence. CLT states sampling distributions are  $\approx$  normal.

**When analytical assumptions are strained, computation offers an alternative.**

Bootstraps approximate the sampling distribution directly from a single sample of data.

## Next week | Hypothesis testing

So far, we used sampling distributions to **quantify uncertainty around estimates**.

Next week, we will take the next steps and start **drawing conclusions**:

Are our observed results compatible with a specific hypothesis about the population?

We will learn how to:

- Formulate **null and alternative hypotheses**
- Use **test statistics** and **reference distributions**
- Interpret **p-values** correctly (and avoid common mistakes)
- Connect **hypothesis tests** and **confidence intervals**

# Class activity | Group assignment based on your RQs

- G1: **Language, Communication, and Bias in AI & Media** – Caleb Agoha, Noha Mahgoub, Yunjia Qi
- G2: **AI, Generative Models, and Evaluation** – Max Davy, Howard Leong, Audrey Yip
- G3: **Media, Platforms, and Audience Response** – Sophie Bair, Charlotte Peart, Michi Wong
- G4: **Political Economy, Policy, and Institutions** – Celikhan Baylan, Graham Gaydos, Caleb Tan
- G5: **Social Behaviour, Trust, and Adoption** – Min Jung, Mia Kussman, Isaac Backer
- G6: **Health, Medicine, and Neuroscience** – Amelia Mercado, Laura Wegner, Ines Trichard
- G7: **Education, Labor, and Socioeconomic Outcomes** – Rehmat Arora, Yilin Qian, Yue Zhang
- G8: **Culture, Mobility, Lifestyle** – Teo Canmetin, Alena Tsvetkova, Fucheng Wang, Nesma Hammouda

Everyone not named: please get together in groups of 3.

# Class activity | Overview

Please access the [Week 2 Class Activity Google Doc](#) on Canvas.

We will do this activity in two Phases.

In each Phase, you will first work individually and then discuss in your assigned group.

In **Phase 1** you will:

- Write down some basic facts about your RQ.
- Introduce your RQ to the rest of your group.

In **Phase 2** you will:

- Think through potential challenges in your own project and your group's projects.
- Discuss these challenges and, collectively, work towards solutions.

**Goal:** Help you prepare for the summative (W4 proposal deadline) and your own research.

# Class activity | Anticipated challenges in answering your RQ

What are the **key challenges** you anticipate in **your own project**, and those in **your group**?

Data access

Sample size

Representativeness

Construct validity

Sampling bias

Ecological validity

Measurement reliability

Missing data

Dependence of observations

Temporal variation

Causality

... or any other challenge you think is relevant!