Lecture 0: Foundations

What Statistics Is and Why It's Hard

How much should you trust a number?

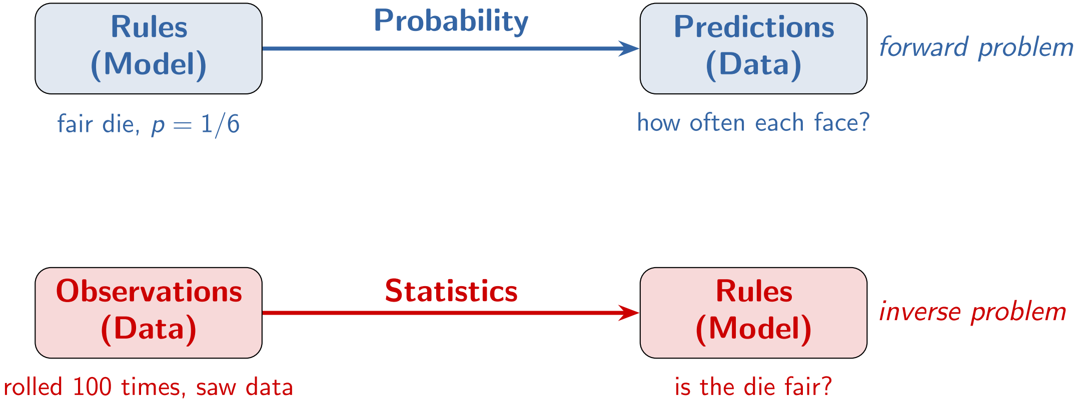
A poll says: "52% support candidate A" $(n = 1{,}000)$

A clinical trial says: "Drug B reduces symptoms by 15%" $(n = 200)$

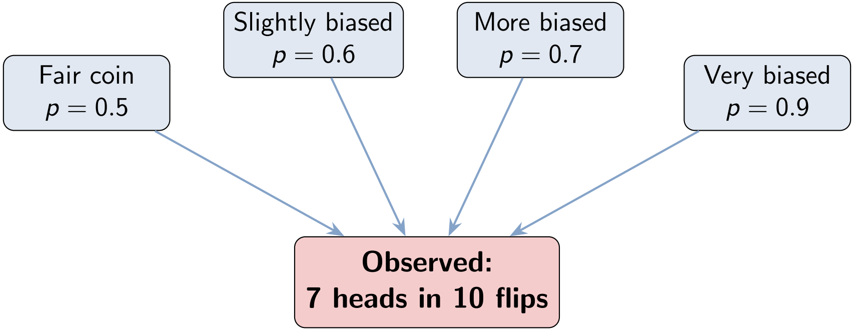
How confident should we be?

This entire course is about answering this question rigorously.

Probability vs Statistics



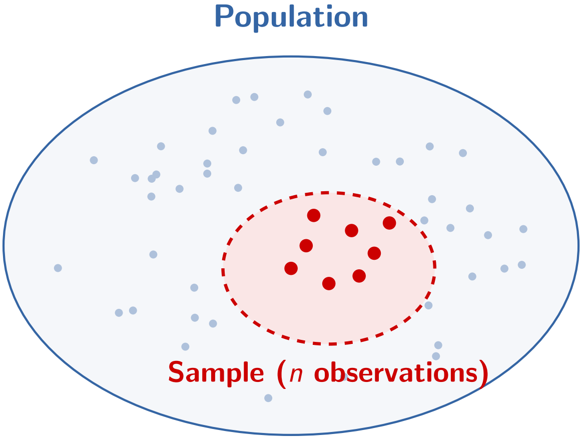
Why the inverse problem is harder



Many different models could have produced this data!

The inverse problem is **ill-posed** — statistics gives us tools to navigate this.

Population vs Sample



Population:
All units of interest

Can be finite or
conceptually infinite

Sample:
The subset we
actually observe

Parameter vs Statistic

Parameter θ

Fixed, unknown number
Describes the **population**

Examples:
μ = true mean lifetime
p = true approval rate
σ^2 = true variance

we estimate this ⟷ **using this**

Statistic $T(X_1, \ldots, X_n)$

Random variable, computable
Computed from the **sample**

Examples:
$\bar{X}$ = sample mean
$\hat{p}$ = sample proportion
S^2 = sample variance

A **parameter** is a fixed number. A **statistic** is a random variable.
Confusing these is the source of most beginner mistakes.

The Triple: Estimand / Estimator / Estimate

Estimand

The *target*

θ (true value)

"true mean height"

approximated by $\longrightarrow$

Estimator

The *recipe*

$\bar{X} = \frac{1}{n} \sum X_i$

"sample mean formula"

applied to data $\longrightarrow$

Estimate

The *number*

$\bar{x} = 172.4$ cm

"the value we got"

Discussion

A polling agency surveys 1,000 people and reports:

"62% support policy X"

Identify each:

1. What is the **population**?
2. What is the **parameter**?
3. What is the **sample**?
4. What is the **statistic**?
5. What is the **estimate**?

The i.i.d. Assumption

Classical statistics assumes our sample $X_1, X_2, \ldots, X_n$ is **i.i.d.**:

Independent

Knowing X_1 tells you nothing about X_2

Each observation is a fresh draw

Identically Distributed

Every X_i comes from the same distribution F

Same process generates each one

When does i.i.d. hold?

- ✓ Random sampling from a large population
- ✓ Repeated independent measurements of the same quantity
- ✓ Controlled experiments with proper randomization

i.i.d. is an **idealization** — it's approximately true in many practical settings, and most of what we'll do this course assumes it.

When does i.i.d. break?

Time dependence
stock prices, weather

Non-response bias
who refuses the survey?

Spatial correlation
neighboring sensors

Distribution shift
training data $\neq$ deployment

Selection bias
hospital-only patients

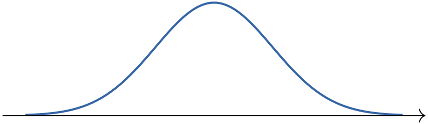
Clustering
students within schools

Not a disaster — just means you need different tools.
But if you *pretend* non-i.i.d. data is i.i.d.,
your conclusions can be **wildly wrong**.

The Plug-in Principle

Idea: We don't know the true distribution F, so replace it with the **empirical distribution** $\hat{F}_n$.

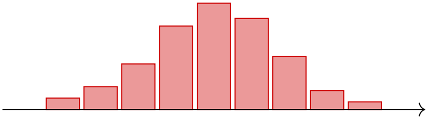
True distribution F
(unknown)



smooth, continuous

replace with →

Empirical distribution $\hat{F}_n$
(computable from data)



mass $1/n$ on each point

Plug-in in Action

Replace the **population quantity** with its **sample analogue**:

Want	**Population**	**Plug-in**
Mean	$\mu = \mathbb{E}_F[X]$	$\hat{\mu} = \bar{X} = \frac{1}{n}\sum X_i$
Variance	$\sigma^2 = \mathrm{Var}_F(X)$	$\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$
CDF	$F(t) = P(X \leq t)$	$\hat{F}_n(t) = \frac{\#\{X_i \leq t\}}{n}$

Glivenko–Cantelli theorem: $\hat{F}_n \to F$ uniformly as $n \to \infty$.
(The "fundamental theorem of statistics" — connects to LLN from Module 20.)

The Summarization Problem

You must summarize a distribution with a **single number** a.

How do you choose?

It depends on what "error" means to you.

This is formalized by a **loss function** $L(\theta, a)$.

Three Losses, Three Optimal Summaries

Squared Error

$$L = (\theta - a)^2$$

Penalizes large errors heavily

$$\Downarrow$$

Mean

Absolute Error

$$L = |\theta - a|$$

Linear penalty, robust to outliers

$$\Downarrow$$

Median

0–1 Loss

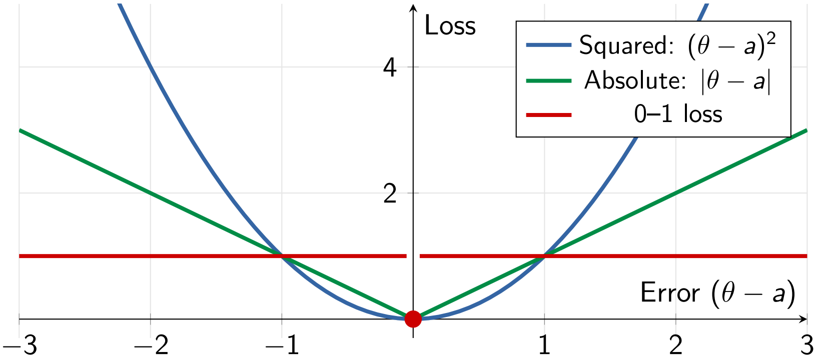
$$L = \mathbf{1}[\theta \neq a]$$

Wrong or right, nothing in between

$$\Downarrow$$

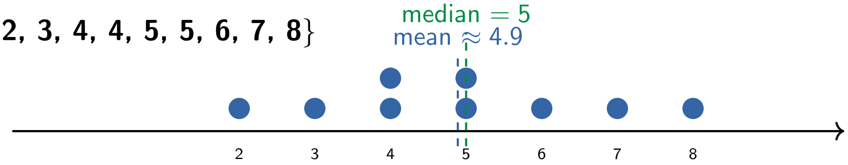
Mode

Visualizing the Losses

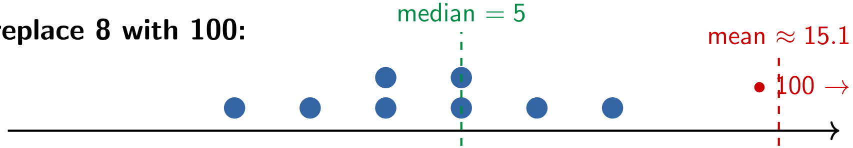


Mean vs Median: Sensitivity to Outliers

Dataset: $\{$**2, 3, 4, 4, 5, 5, 6, 7, 8**$\}$



Now replace 8 with 100:



One outlier moved the mean from 4.9 to 15.1.
The median didn't budge.

Risk and Empirical Risk

Risk (theoretical)

$$R(\theta, \hat{\theta}) \;=\; \mathbb{E}\big[L(\theta, \hat{\theta})\big]$$

Average loss over
all possible samples

(unknown — depends on F)

approximate $\dashrightarrow$

Empirical Risk

$$\hat{R} \;=\; \frac{1}{n}\sum_{i=1}^{n} L(X_i, a)$$

Average loss on the
data we actually have

(computable!)

Empirical Risk Minimization (ERM): choose the estimator that minimizes $\hat{R}$.
This principle unifies least squares, maximum likelihood, and most learning algorithms.

The Choice of Loss Reflects Your Values

Medical dosing
Overdose vs underdose have very different consequences
$\Rightarrow$ asymmetric loss

House prices
Mansions in the data?
MAE $\neq$ MSE answer
$\Rightarrow$ median vs mean

Spam filter
Blocking real email vs letting spam through
$\Rightarrow$ different error costs

Flood planning
Under-predicting flood level is catastrophic
$\Rightarrow$ conservative (high quantile)

There is no "correct" loss — it depends on context.

What Statistics Will Give Us

Today
Foundations

Estimation
What's our best guess for θ?
Lectures 2–3

Uncertainty
How confident are we?
Lectures 4–5

Decisions
Is the effect real or noise?
Lectures 6–7

Models
How do variables relate?
Lecture 8

Practical: Loss, Risk, and Robustness

Given a dataset (e.g., city temperatures, exam scores, household incomes):

1. Compute the sample **mean**, **median**, and **mode**
2. Compute empirical risk under each loss for each summary
3. Verify: mean minimizes squared-error risk, median minimizes absolute-error risk
4. Add one extreme outlier and repeat
5. Observe: how much does each summary move?

Discuss: Which summary would you report, and why?
Does the answer depend on context?

Questions?

Next lecture: Descriptive Statistics & Empirical Distributions