Statistics: Learning from Data under Uncertainty

Ilmun Kim

Department of Mathematical Sciences
KAIST

Overview

- Introduction to Statistics
- Estimation
- Hypothesis Testing
- Modern Topics
- Research Interest

Statistics: Learning from Data under Uncertainty

At its core, it combines

- Math + CS + Domain Expertise

  Foundation of modern statistical practice

Statistics: Learning from Data under Uncertainty

At its core, it combines

- Math + CS + Domain Expertise

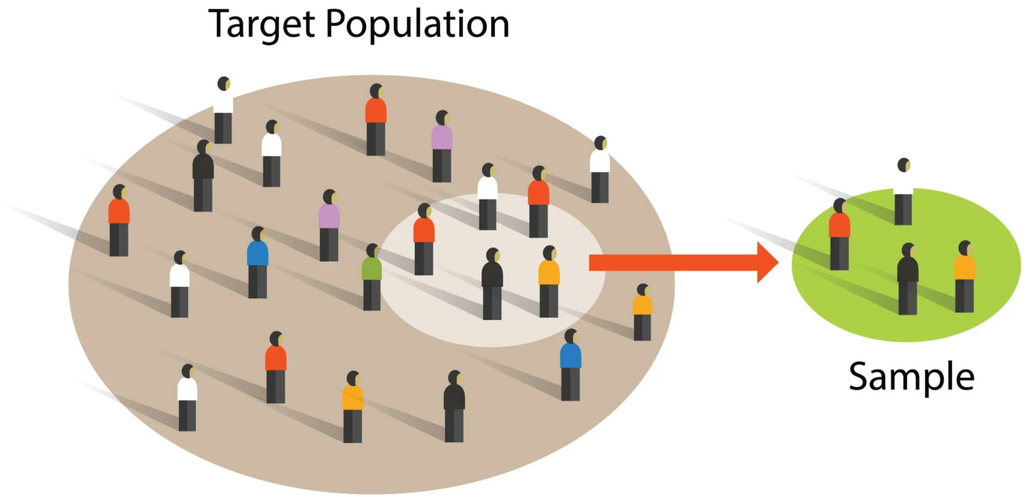
  Foundation of modern statistical practice

A central idea is the distinction between

- Population vs. Sample

  Full group of interest

  A representative (but not perfect) subset



Statistics: Learning from Data under Uncertainty

At its core, it combines

- Math + CS + Domain Expertise

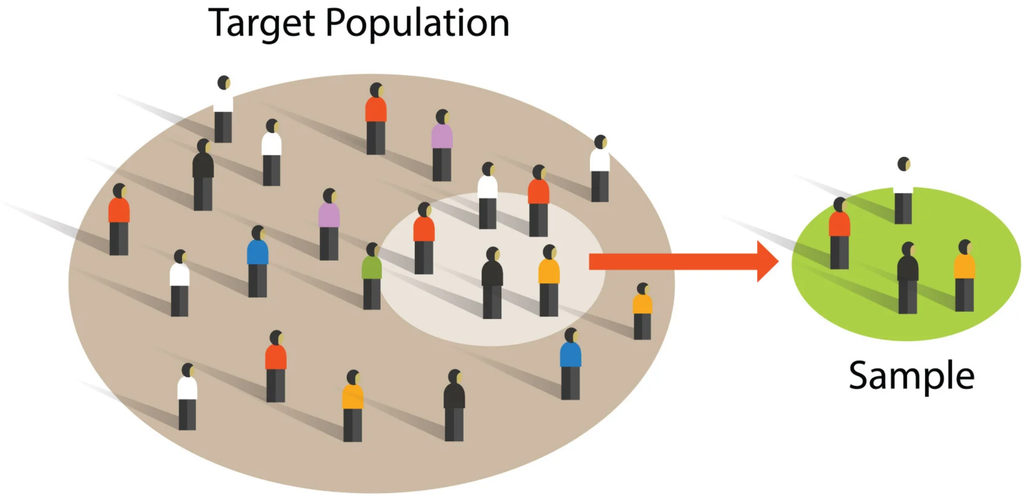
  Foundation of modern statistical practice

A central idea is the distinction between

- Population vs. Sample

  Full group of interest

  A representative (but not perfect) subset



Problem:

- Collecting data from the entire population is usually impossible

- There are costs, time constraints and even ethical barriers

→

Statistics: turning the uncertainty from incomplete data into rigorous conclusions

Broadly statistics involves three main activities

- **Estimation:**
  Figuring out an **unknown number** (like the average income in a city or the proportion of people who prefer Apple over Samsung) from data

Broadly statistics involves three main activities

- **Estimation:**
  Figuring out an **unknown number** (like the average income in a city or the proportion of people who prefer Apple over Samsung) from data

- **Inference:**
  Goes a step further → instead of just producing a number, quantify the uncertainty around it

  e.g., Confidence interval → giving a range of likely values

  Hypothesis testing → checking if an idea about the data is reasonable

Broadly statistics involves three main activities

- **Estimation:**
  Figuring out an **unknown number** (like the average income in a city or the proportion of people who prefer Apple over Samsung) from data

- **Inference:**
  Goes a step further → instead of just producing a number, quantify the uncertainty around it

  e.g., Confidence interval → giving a range of likely values
  Hypothesis testing → checking if an idea about the data is reasonable

- **Prediction:**
  Using past data to make guesses about future outcomes
  (like predicting tomorrow's weather or next year's sales)

  This is where statistics overlaps with **machine learning** and **data science**

Estimation

Estimation

- **Parameter θ**
  : a numerical characterization of the population which is **unknown**

    e.g., population mean $\mathbb{E}[X] = \mu$, population variance $\mathrm{Var}[X] = \sigma^2$
    regression coefficient β in $Y = X\beta + \epsilon$

Estimation

- **Parameter** θ
  : a numerical characterization of the population which is **unknown**

    e.g., population mean $\mathbb{E}[X] = \mu$, population variance $\mathrm{Var}[X] = \sigma^2$
    regression coefficient β in $Y = X\beta + \epsilon$

- Since θ is **unknown**, we try to learn about it using data

Estimation

- **Parameter** θ
  : a numerical characterization of the population which is **unknown**

    e.g., population mean $\mathbb{E}[X] = \mu$, population variance $\mathrm{Var}[X] = \sigma^2$
    regression coefficient β in $Y = X\beta + \epsilon$

- Since θ is **unknown**, we try to learn about it using data

- Suppose we observe a random sample $X_1, X_2, \ldots, X_n$

Construct a function $\hat{\theta} : \{X_1, \ldots, X_n\} \to \mathbb{R}$ such that $\hat{\theta}$ and θ are close ($\hat{\theta} \approx \theta$)

Estimation

- **Parameter** θ
  : a numerical characterization of the population which is **unknown**

    e.g., population mean $\mathbb{E}[X] = \mu$, population variance $\mathrm{Var}[X] = \sigma^2$
    regression coefficient β in $Y = X\beta + \epsilon$

- Since θ is **unknown**, we try to learn about it using data

- Suppose we observe a random sample $X_1, X_2, \ldots, X_n$

> Construct a function $\hat{\theta} : \{X_1, \ldots, X_n\} \to \mathbb{R}$ such that $\hat{\theta}$ and θ are close ($\hat{\theta} \approx \theta$)

Big picture: the population has some hidden truth θ, we only see a sample and we use the sample to come up with an estimate that gets us as close as possible to that truth

Estimation

Examples of Estimators

- Sample Mean

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

- Sample Variance

$$\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

- Least Squares Estimator

$$\hat{\theta} = (X^\top X)^\top X^\top Y$$

Estimation

Examples of Estimators

- Sample Mean

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

- Sample Variance

$$\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

- Least Squares Estimator

$$\hat{\theta} = (X^\top X)^\top X^\top Y$$

Key points

- **Classical estimators** (mean, variance, least squares) have explicit formulas
- **Many modern estimators** (e.g., in machine learning) are **black-box**:
  no closed-form expression, obtained by optimization algorithms

Justification of Estimators

- **Important point:** for any given parameter, there isn't just one estimator;
  in fact, there are **many possible estimators** for the same parameter
  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$ (in fact infinitely many estimators)

  For example, we want to estimate the population mean
  → sample mean, sample median, a single data point etc; each of these is technically an estimator

Justification of Estimators

- **Important point:** for any given parameter, there isn't just one estimator;
  in fact, there are **many possible estimators** for the same parameter
  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$ (in fact infinitely many estimators)

  For example, we want to estimate the population mean
  → sample mean, sample median, a single data point etc; each of these is technically an estimator

- **Key Question:**

  Which estimator should we use and why? → This is where statistics steps in!

Justification of Estimators

- **Important point:** for any given parameter, there isn't just one estimator;
  in fact, there are **many possible estimators** for the same parameter
  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$ (in fact infinitely many estimators)

  For example, we want to estimate the population mean
  → sample mean, sample median, a single data point etc; each of these is technically an estimator

- **Key Question:**

  Which estimator should we use and why? → This is where statistics steps in!

- **Statistics provides criteria** (e.g., unbiasedness, variance, consistency, efficiency)

  to justify and compare estimators

Justification of Estimators

- **Important point:** for any given parameter, there isn't just one estimator;
  in fact, there are **many possible estimators** for the same parameter
  $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$ (in fact infinitely many estimators)

  For example, we want to estimate the population mean
  → sample mean, sample median, a single data point etc; each of these is technically an estimator

- **Key Question:**
  Which estimator should we use and why? → This is where statistics steps in!

- **Statistics provides criteria** (e.g., unbiasedness, variance, consistency, efficiency)
  to justify and compare estimators

Takeaway: there are many possible estimators, but statistics gives us a framework to decide which ones are good and which ones are not

How do we compare estimators and decide which one to use?

- **Risk:** a measure of discrepancy between estimator $\widehat{\theta}$ and true parameter θ

  → how far an estimator is from the true parameter on average?

How do we compare estimators and decide which one to use?

- **Risk:** a measure of discrepancy between estimator $\widehat{\theta}$ and true parameter θ

  → how far an estimator is from the true parameter on average?

- Goal: **choose an estimator that minimizes the risk**

How do we compare estimators and decide which one to use?

- **Risk:** a measure of discrepancy between estimator $\hat{\theta}$ and true parameter θ

  → how far an estimator is from the true parameter on average?

- Goal: **choose an estimator that minimizes the risk**

- A common choice of risk: **Mean Squared Error (MSE)**

$$\text{MSE} = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \int (\hat{\theta} - \theta)^2 f_\theta(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

How do we compare estimators and decide which one to use?

- **Risk:** a measure of discrepancy between estimator $\hat{\theta}$ and true parameter θ

  → how far an estimator is from the true parameter on average?

- Goal: **choose an estimator that minimizes the risk**

- A common choice of risk: **Mean Squared Error (MSE)**

$$\text{MSE} = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \int (\hat{\theta} - \theta)^2 f_\theta(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

- **Interpretation:** imagine repeating experiments many times and averaging $(\hat{\theta} - \theta)^2$

How do we compare estimators and decide which one to use?

- **Risk:** a measure of discrepancy between estimator $\hat{\theta}$ and true parameter θ

  → how far an estimator is from the true parameter on average?

- Goal: **choose an estimator that minimizes the risk**

- A common choice of risk: **Mean Squared Error (MSE)**

$$\text{MSE} = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \int (\hat{\theta} - \theta)^2 f_\theta(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

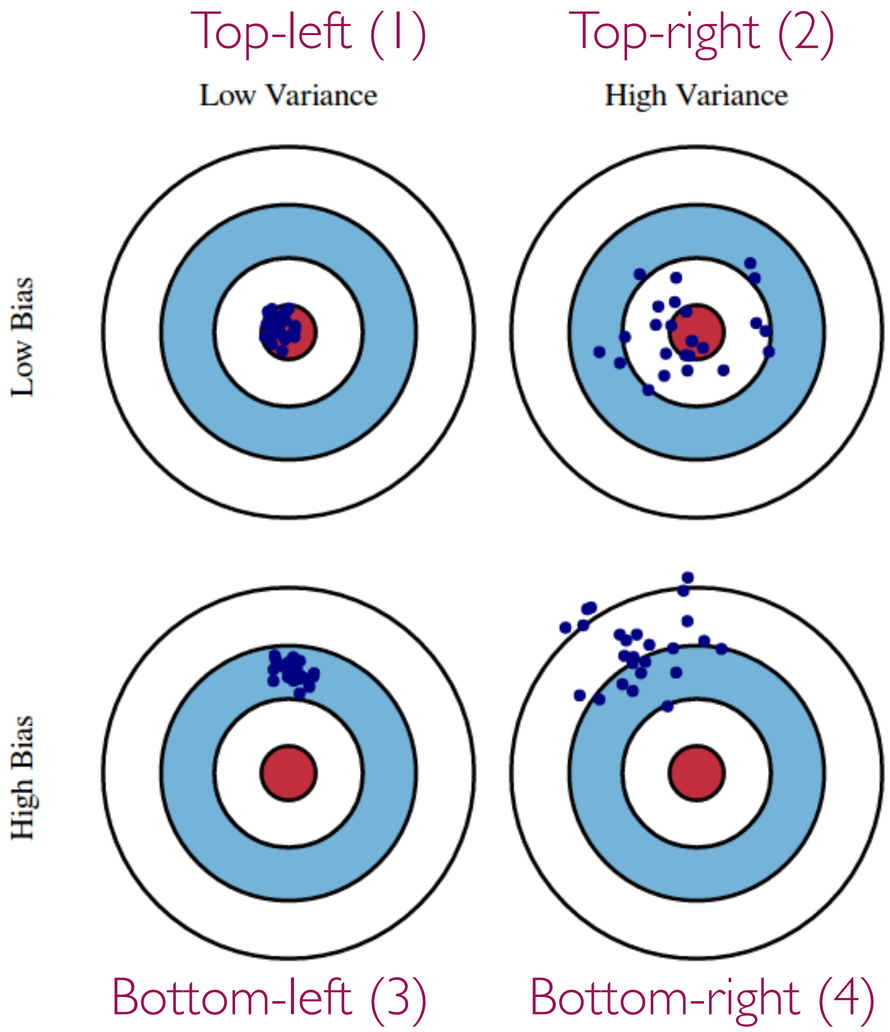
- **Interpretation:** imagine repeating experiments many times and averaging $(\hat{\theta} - \theta)^2$

- **Bias-variance decomposition:** $\text{MSE} = \underbrace{(\mathbb{E}_\theta[\hat{\theta}] - \theta)^2}_{\text{Bias}^2} + \underbrace{\text{Var}_\theta[\hat{\theta}]}_{\text{Variance}}$

Bias: how far the estimator's expected value is from the true parameter?

Variance: how much the estimator fluctuates from sample to sample

Visual illustration of bias and variance

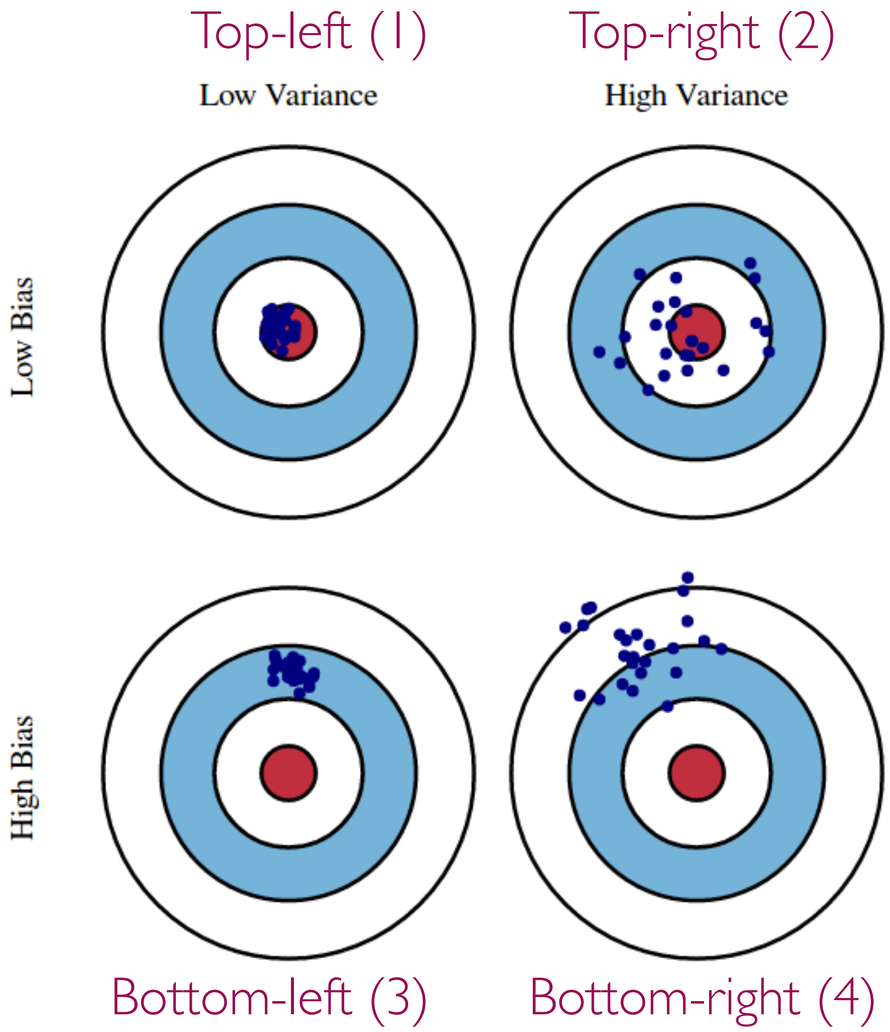


Target region (e.g., true mean)

- **Bias:** measures how far off the predicted dots from the center

  If the cloud of dots is shifted away from the center → high bias
  If they're centered right around the center → low bias

Visual illustration of bias and variance



Target region (e.g., true mean)

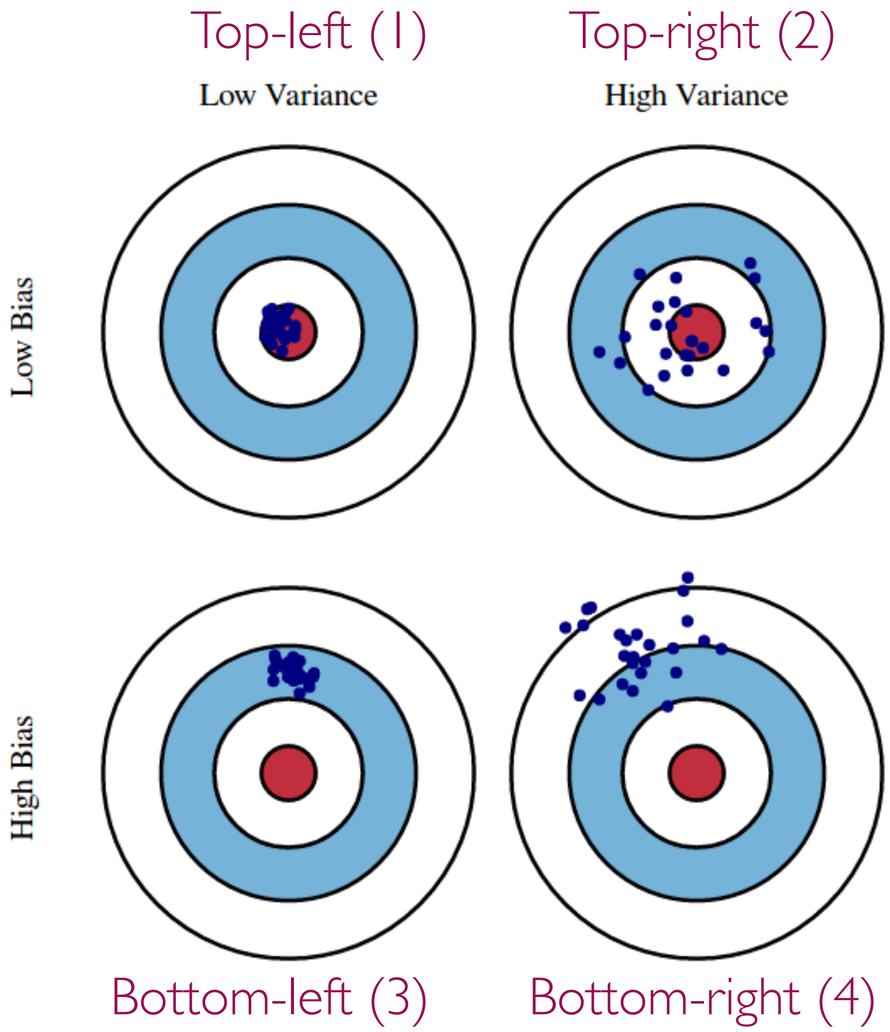
- **Bias:** measures how far off the predicted dots from the center

  If the cloud of dots is shifted away from the center → high bias
  If they're centered right around the center → low bias

- **Variance:** how scattered our predictions are

  If they're tightly clustered → low variance
  If they're scattered all over → high variance

Visual illustration of bias and variance



Target region (e.g., true mean)

- **Bias:** measures how far off the predicted dots from the center

  If the cloud of dots is shifted away from the center → high bias
  If they're centered right around the center → low bias

- **Variance:** how scattered our predictions are

  If they're tightly clustered → low variance
  If they're scattered all over → high variance

→ Choose an estimator with **low bias & low variance** (Top-left)

Optimal Estimator

Question: Is it possible to construct $\widehat{\theta}$ such that

$$\mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] \leq \mathbb{E}_\theta[(\widetilde{\theta} - \theta)^2]$$

for all possible $\theta \in \Omega$ and all possible estimators $\widetilde{\theta}$?

Optimal Estimator

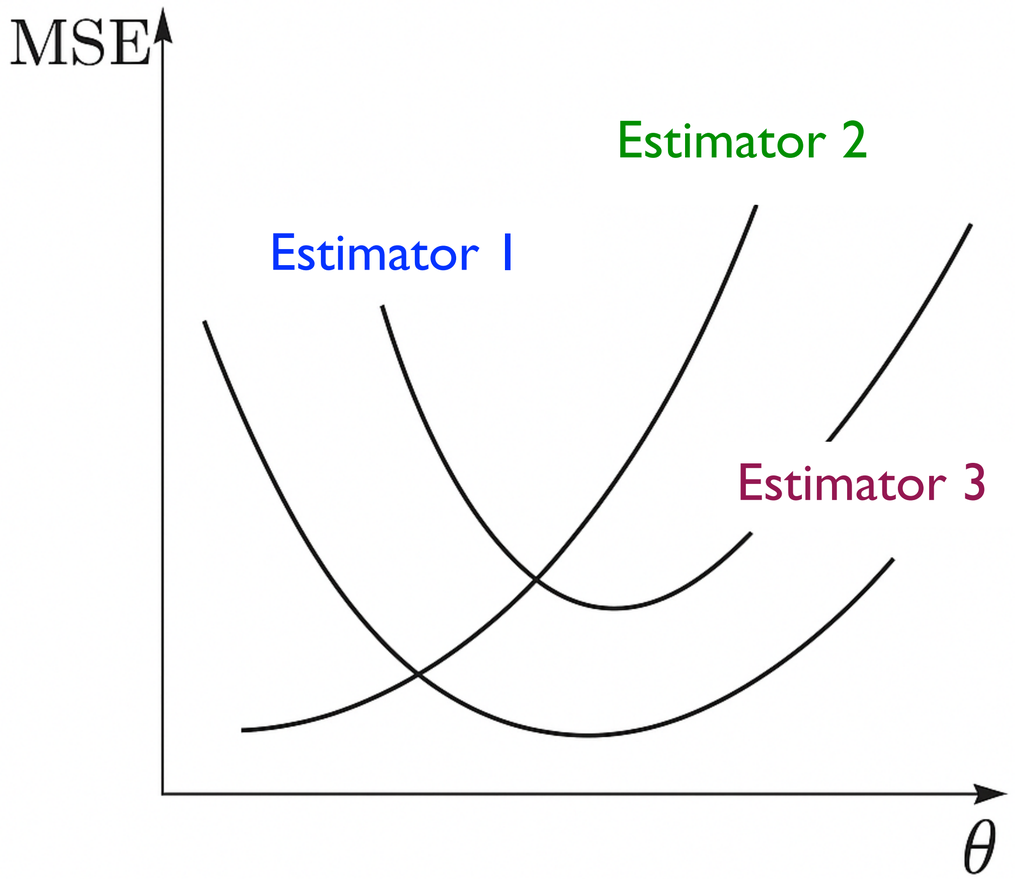
Question: Is it possible to construct $\hat{\theta}$ such that

$$\mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \leq \mathbb{E}_\theta[(\widetilde{\theta} - \theta)^2]$$

for all possible $\theta \in \Omega$ and all possible estimators $\widetilde{\theta}$?

Answer: No!
Such a universally best estimator does not exist in reasonable cases

Optimal Estimator



- Each performs **better** in some regions but **worse** in others
- No single estimator dominates across all possible parameter values

Optimal Estimator

In other words, a **global best estimator** does not exist

→ Instead we look for **optimality under constraints**

$\hat{\theta}$ chosen to be best within a restricted class

Optimal Estimator

In other words, a **global best estimator** does not exist

→ Instead we look for **optimality under constraints**

$\hat{\theta}$ **chosen to be best within a restricted class**

Examples of constrained optimality

- **MVUE** (Minimum Variance Unbiased Estimator)
- **Minimax Estimator** (minimizes worst-case risk)
- **Bayes Estimator** (minimizes posterior expected loss)

Minimum Variance Unbiased Estimator

- **Historical origin:** Developed in the early days of statistics (Lehmann, Scheffe etc)
- **Focus:** Restrict attention to unbiased estimators

  i.e., $\mathbb{E}_\theta[\widehat{\theta}] = \theta$ for all $\theta \in \Omega$

- **Goal:** Among all unbiased estimators, find one with the smallest variance

  $\mathrm{Var}_\theta(\widehat{\theta}) \leq \mathrm{Var}_\theta(\widetilde{\theta})$ for all unbiased estimators $\widetilde{\theta}$

- **Definition:** Such an estimator is called the MVUE

Minimum Variance Unbiased Estimator

Techniques

- Use of sufficient statistics
- Complete families of distributions
- Lehmann-Scheffe theorem

Minimum Variance Unbiased Estimator

Techniques

- Use of sufficient statistics
- Complete families of distributions
- Lehmann-Scheffe theorem

Limitations

- **No unbiased estimator** exists for some problems
- Typically restricted to small classes of parametric distributions
- Sometimes introducing a small bias can dramatically reduce MSE

Minimax Estimator

Dominant Paradigm:

- Chosen for **mathematical tractability**
- Works in both **parametric** and **nonparametric** settings

Minimax Estimator

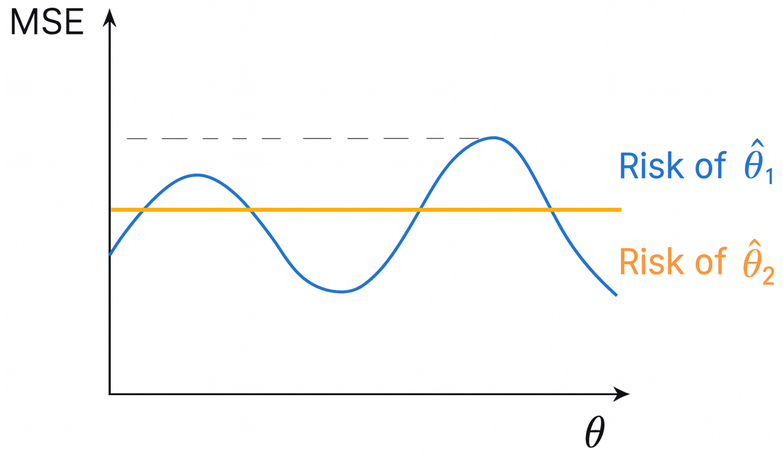
Dominant Paradigm:

- Chosen for **mathematical tractability**
- Works in both **parametric** and **nonparametric** settings

Definition

An estimator $\hat{\theta}$ is **minimax** if

$$\sup_{\theta \in \Omega} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \inf_{\tilde{\theta}} \sup_{\theta \in \Omega} \mathbb{E}_\theta[(\tilde{\theta} - \theta)^2]$$



- **Left side:** Worst-case risk of $\hat{\theta}$
- **Right side:** Minimum possible worst-case risk over all estimators $\tilde{\theta}$

Minimax Estimator

Techniques

- Le Cam's lemmas
- Information-theoretic methods (e.g., Fano's inequality)
- Concentration inequalities

Lucien Le Cam

Minimax Estimator

Lucien Le Cam

Techniques

- Le Cam's lemmas
- Information-theoretic methods (e.g., Fano's inequality)
- Concentration inequalities

Limitations

- Focuses only on **worst-case risk** → can be **overly conservative**
- **Difficult to compute** exactly for many complex models
- The minimax solution may depend on the **choice of loss function**

Hypothesis Testing

Hypothesis Testing

Motivation

- Estimation gives **numerical summaries** of parameters
- But often we want to **make decisions** based on data

This is where **hypothesis testing** comes in
→ provides a framework for making decisions based on data

Hypothesis Testing

Motivation

- Estimation gives **numerical summaries** of parameters
- But often we want to **make decisions** based on data

This is where **hypothesis testing** comes in
→ provides a framework for making decisions based on data

Examples of Questions

- Medicine: Is a **new treatment** effective?
- Probability: Is a coin **fair** (prob = 0.5)?
- Social Science: Are **two populations** equal?



Hypothesis Testing

Basic setup

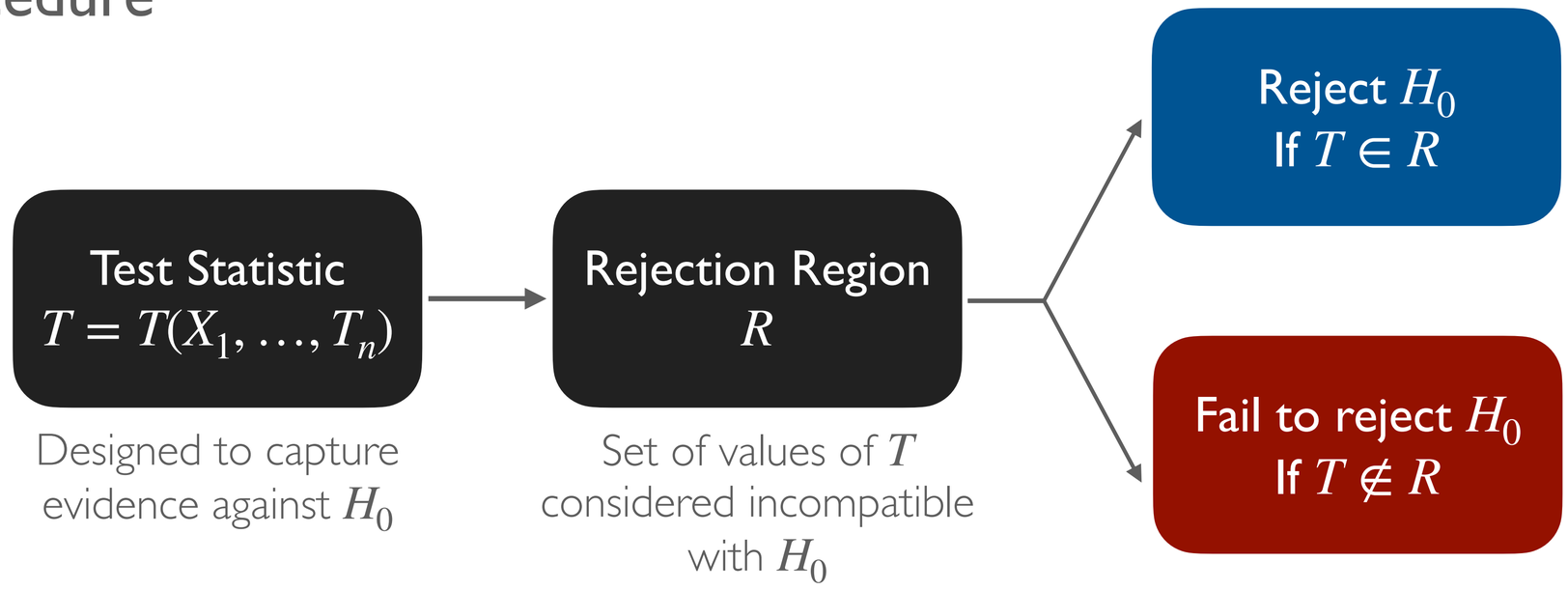
- **Null hypothesis** H_0: baseline claim (e.g., $H_0 : \theta = 0$) e.g., no effect or no difference
- **Alternative hypothesis** H_1: competing claim (e.g., $H_1 : \theta \neq 0$)

Hypothesis Testing

Basic setup

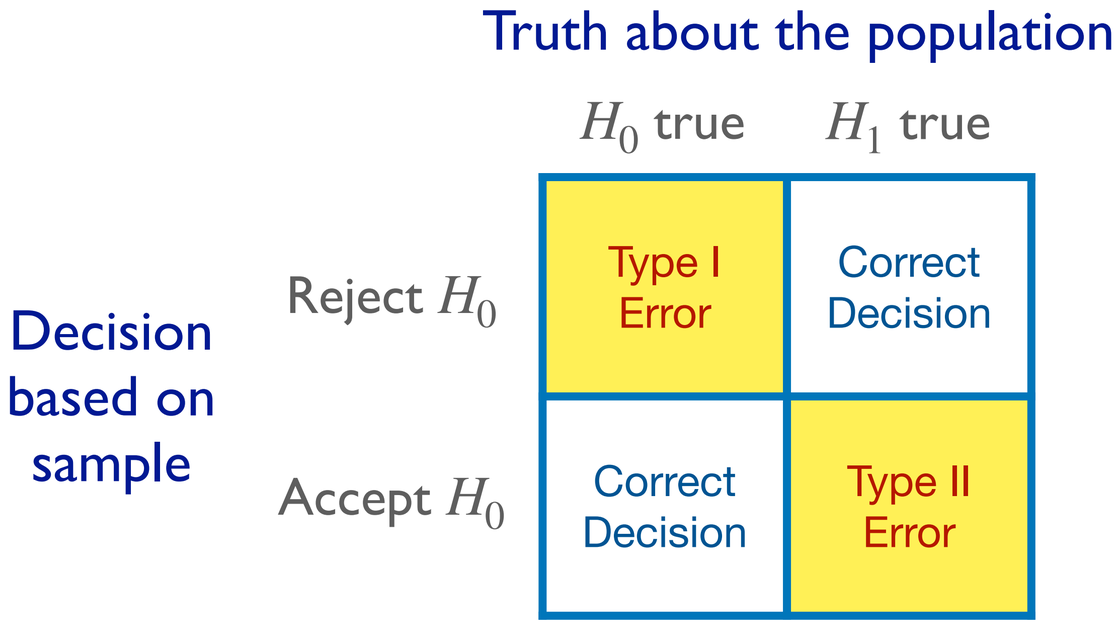
- **Null hypothesis** H_0: baseline claim (e.g., $H_0 : \theta = 0$) e.g., no effect or no difference
- **Alternative hypothesis** H_1: competing claim (e.g., $H_1 : \theta \neq 0$)

Test Procedure



Hypothesis Testing

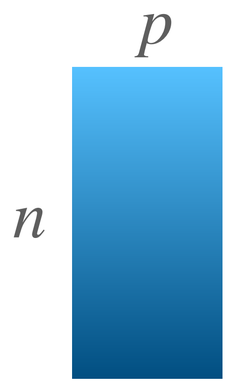
Goal: Design a test that minimizes the **type I error + the type II error**



Modern Topics

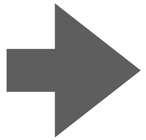
The move from classical to high-dimensional data

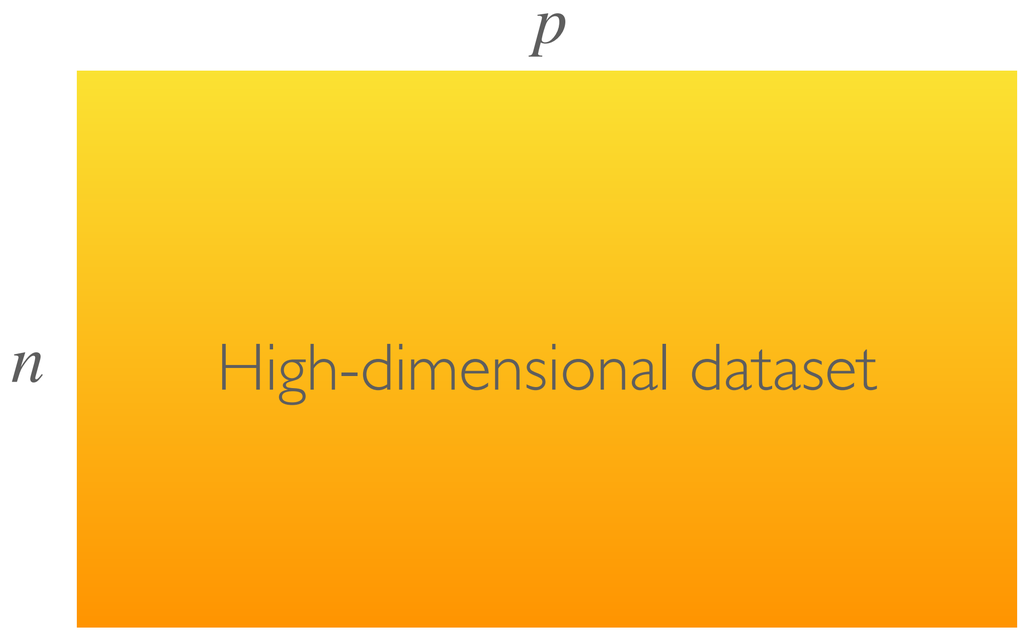
n: sample size
p: # of covariates



Classical dataset

$n \gg p$





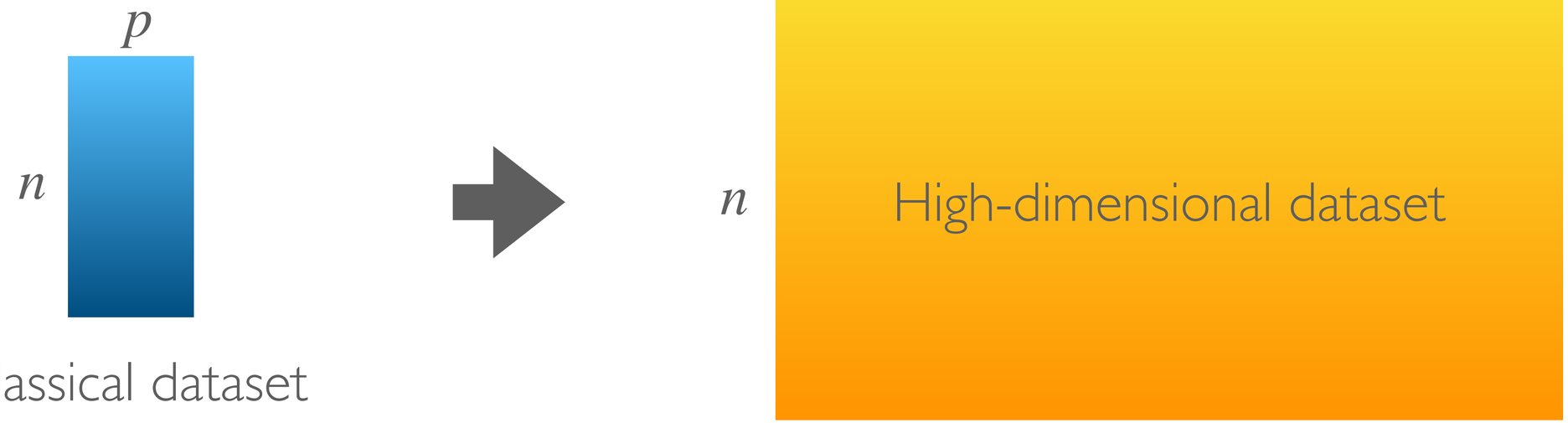
$n \ll p$ or $n \asymp p$

E.g., **medical study:** thousands of patients but only a few measured features like blood pressure and cholesterol

E.g., **genomics**: thousands of gene expression measurements but only a few hundred patients

The move from classical to high-dimensional data

n: sample size
p: # of covariates



Classical dataset

$n \gg p$

$n \ll p$ or $n \asymp p$

Classical methods break down and we need new tools

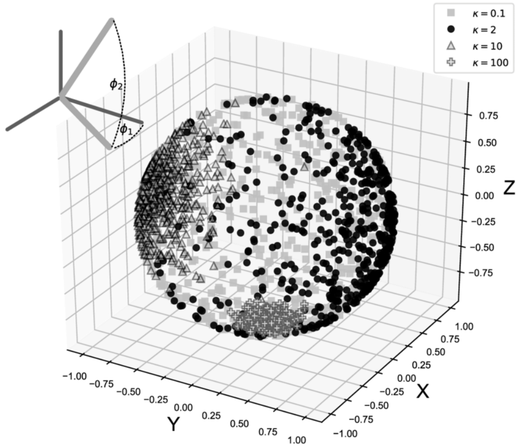
(e.g., regularization method, dimension reduction and ML approaches)

From Euclidean data to more complex data types

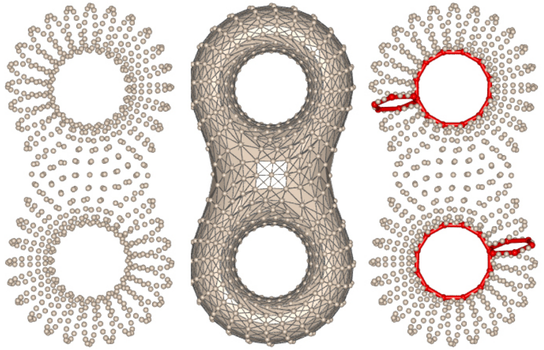
e.g., hight, weight or exam scores

→ Modern applications involve data with different structures

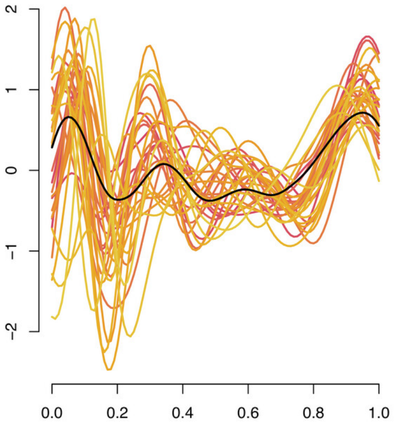
Networks



Spherical data



Topological data



Functional data

Modern topics in statistics

Privacy Issue

- Sensitive data poses privacy risks (e.g., medical records, financial data, location tracking)

Computational Issue

- High-dimensional data and complex algorithms cause computational explosion

→ need scalable, efficient methods

Memory Issue

- Massive datasets cannot fit entirely into memory

→ motivates streaming algorithms, distributed computing and compression techniques

Interpretability Issue

- Complex models often act as "black boxes"

Research interest

Research Theme: ML-Assisted Statistical Inference

Empirical Success of ML Models

- Rapid advancements and breakthroughs across **various domains**
- Innovations and applications are evolving at an **unprecedented pace**

Research Theme: ML-Assisted Statistical Inference

Empirical Success of ML Models

- Rapid advancements and breakthroughs across **various domains**
- Innovations and applications are evolving at an **unprecedented pace**

Statistical Theory and Methods

- Often applied to **critical fields** (e.g., biology) where precision is vital
- Require **rigor**, leading to **slower** but more cautious developments

Research Theme: ML-Assisted Statistical Inference

Empirical Success of ML Models

- Rapid advancements and breakthroughs across **various domains**
- Innovations and applications are evolving at an **unprecedented pace**

Statistical Theory and Methods

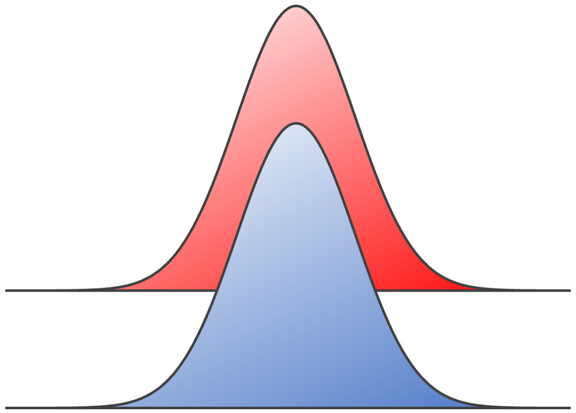
- Often applied to **critical fields** (e.g., biology) where precision is vital
- Require **rigor**, leading to **slower** but more cautious developments

Can we leverage **modern ML and AI tools** to efficiently address **traditional statistical challenges**?

Two-Sample Problem

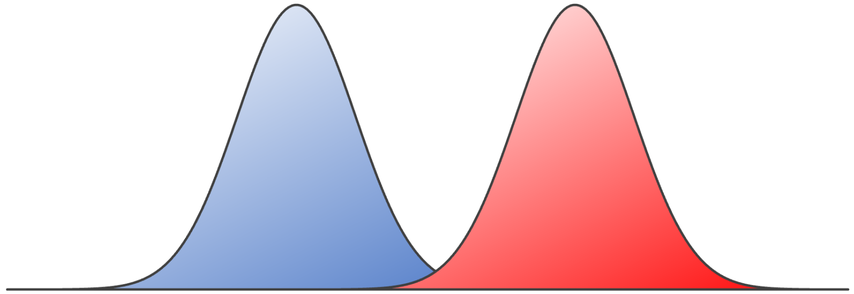
Given $\{X_1, \ldots, X_n\} \overset{\text{i.i.d.}}{\sim} P_X$ and $\{Y_1, \ldots, Y_m\} \overset{\text{i.i.d.}}{\sim} Q_Y$

we want to test whether



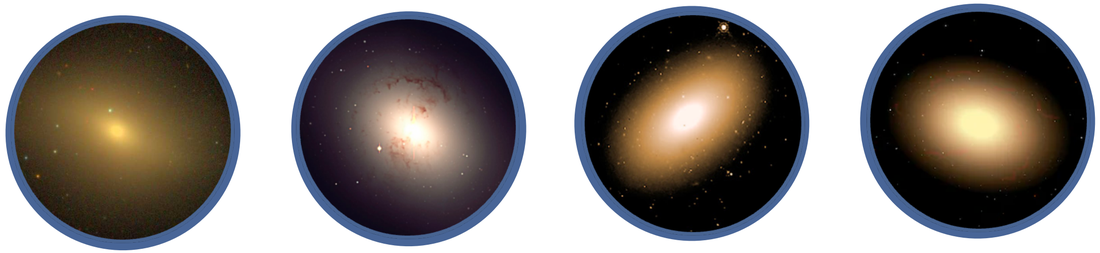
$H_0 : P_X = Q_Y$

versus



$H_1 : P_X \neq Q_Y$

Applications: Astronomy



High-mass versus Low-mass galaxies

Applications: Generative Models

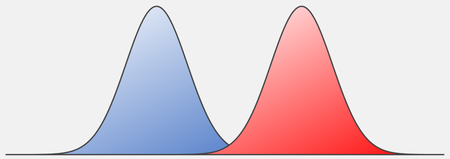
$\sim P_{\text{real}}$

$\sim Q_{\text{artificial}}$

Conventional Approach

Step 1

Compute a test statistic



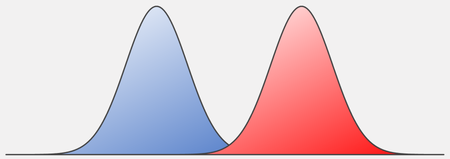
$$T_n = \int (\hat{p}_X - \hat{q}_Y)^2 d\mu$$

e.g., L_2 distance between kernel density estimates

Conventional Approach

Step I

Compute a test statistic



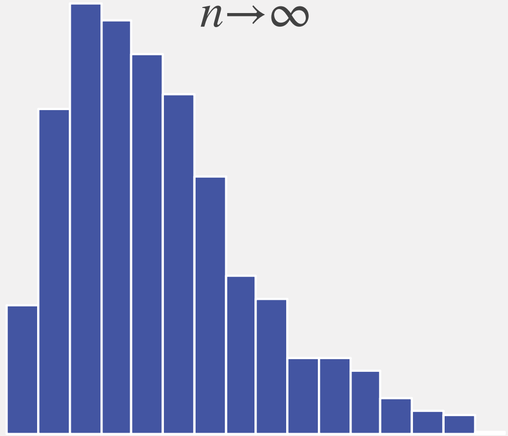
$$T_n = \int (\hat{p}_X - \hat{q}_Y)^2 d\mu$$

e.g., L_2 distance between kernel density estimates

Step II

Derive the null distribution

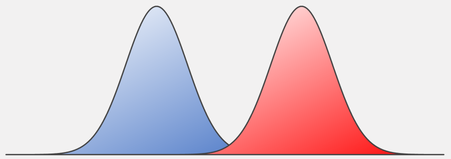
$$\lim_{n \to \infty} \mathbb{P}(T_n \leq t)$$



Conventional Approach

Step I

Compute a test statistic



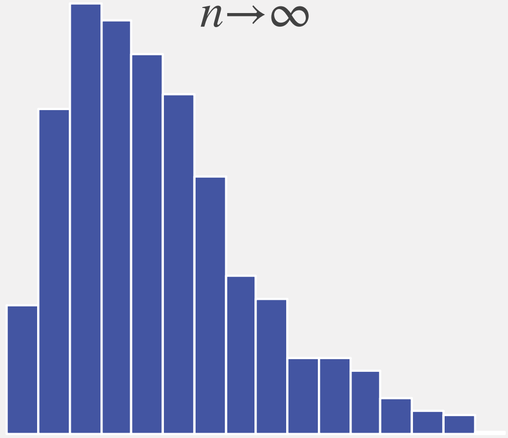
$$T_n = \int (\hat{p}_X - \hat{q}_Y)^2 d\mu$$

e.g., L_2 distance between kernel density estimates

Step II

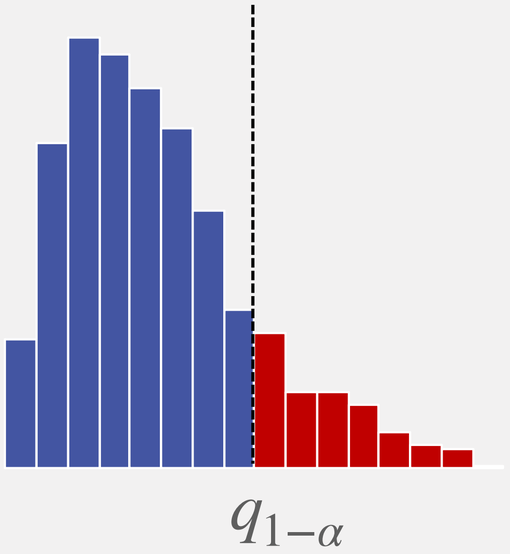
Derive the null distribution

$$\lim_{n \to \infty} \mathbb{P}(T_n \leq t)$$



Step III

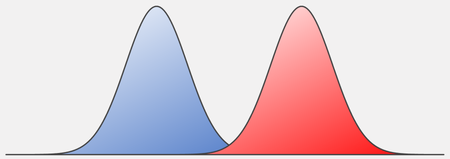
Reject H_0 if $T_n > q_{1-\alpha}$



Conventional Approach

Step 1

Compute a test statistic



$$T_n = \int (\hat{p}_X - \hat{q}_Y)^2 d\mu$$

e.g., L_2 distance between kernel density estimates

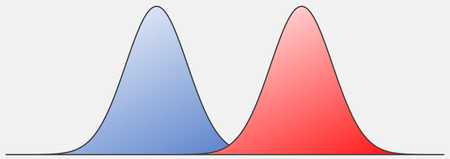
Key Challenges

- Unsupervised learning techniques
- Lack of clear evaluation metrics
- Limited tools and frameworks
- Curse of dimensionality

Conventional Approach

Step 1

Compute a test statistic



$$T_n = \int (\hat{p}_X - \hat{q}_Y)^2 d\mu$$

e.g., L_2 distance between kernel density estimates

Key Challenges

- Unsupervised learning techniques
- Lack of clear evaluation metrics
- Limited tools and frameworks
- Curse of dimensionality

Our approach

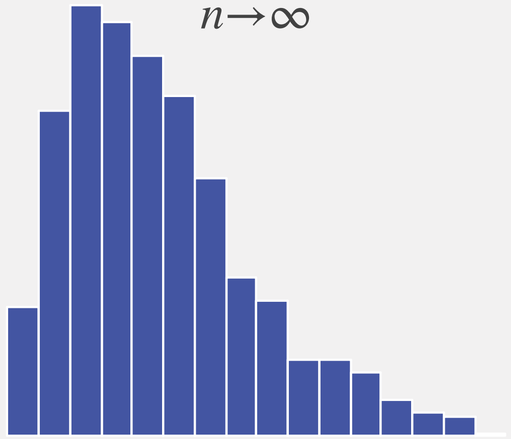
- Convert it into a **supervised learning** problem to leverage **benefits of labeled data**

Conventional Approach

Step II

Derive the null distribution

$$\lim_{n \to \infty} \mathbb{P}(T_n \leq t)$$



Key Challenges

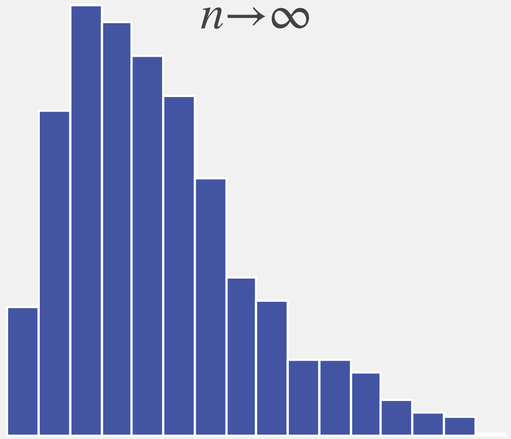
- Only asymptotically valid
- Strong assumptions to ensure validity
- Limited types of test statistics

Conventional Approach

Step II

Derive the null distribution

$$\lim_{n\to\infty} \mathbb{P}(T_n \leq t)$$



Key Challenges

- Only asymptotically valid
- Strong assumptions to ensure validity
- Limited types of test statistics

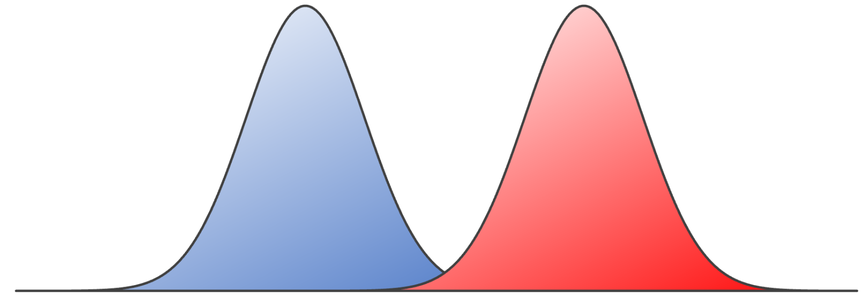
Our approach

- Use the **permutation method** that yields **finite-sample validity** for any type of test statistics

ML-Assisted Two-Sample Tests

{ **Kim**, Ramdas, Singh, Wasserman (**AoS**)
Kim, Lee, Lei (**EJS**)

Classification/Regression
(supervised learning methods)

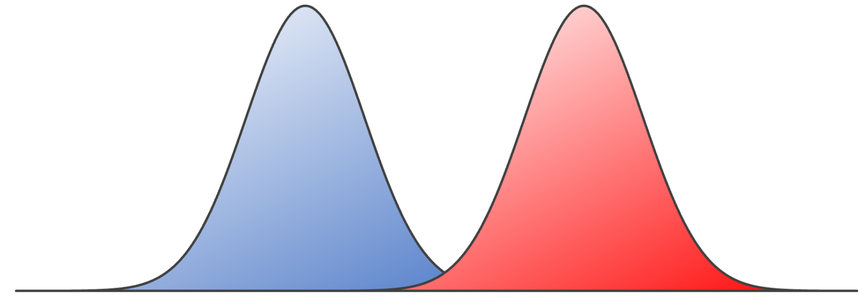


Two-sample test
(statistical problem)

ML-Assisted Two-Sample Tests

{ **Kim**, Ramdas, Singh, Wasserman (**AoS**)
Kim, Lee, Lei (**EJS**)

Classification/Regression
(supervised learning methods)

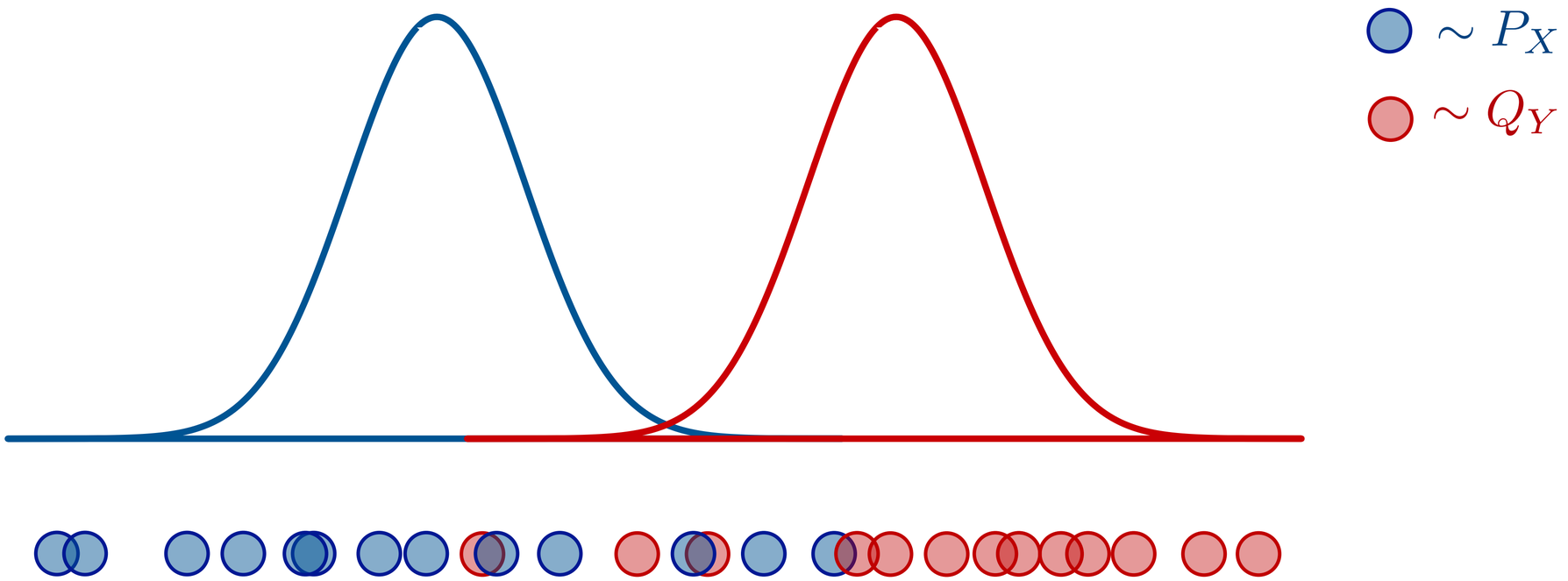


Two-sample test
(statistical problem)

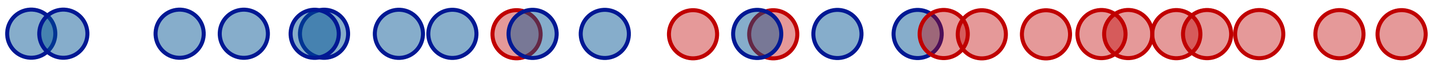
Q1. How can we **leverage ML tools** for the two-sample problem?

Q2. What are some **theoretical properties** of this approach?

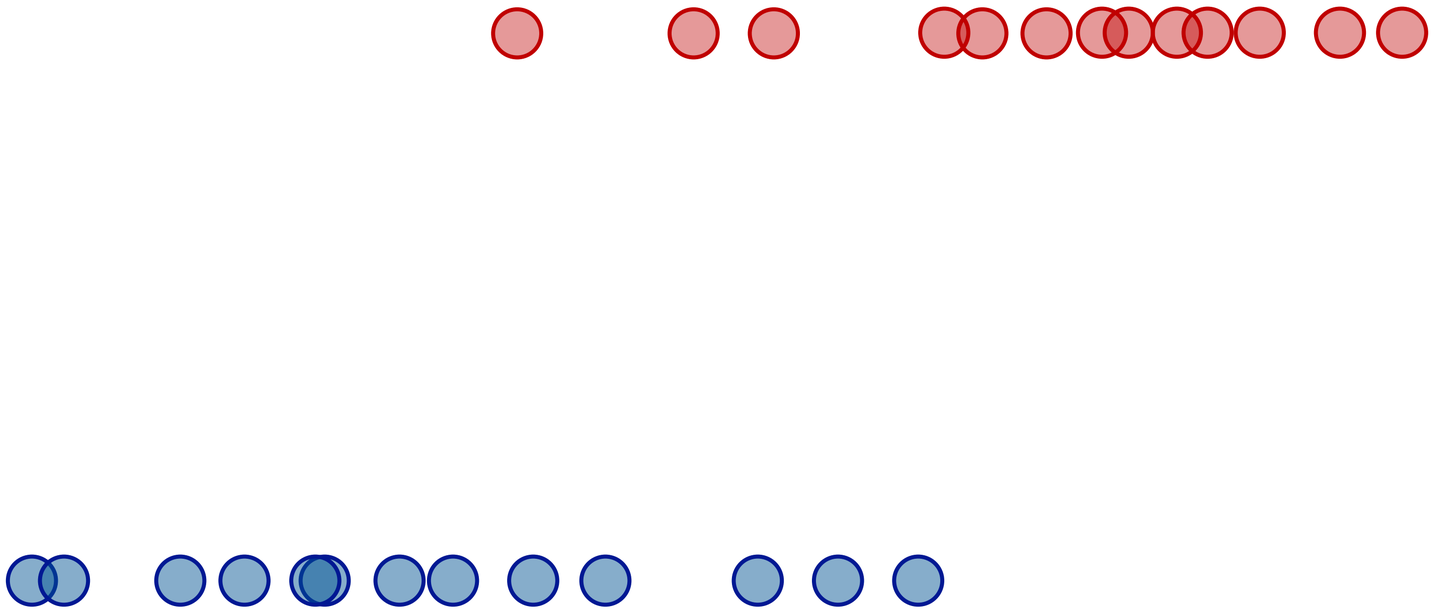
Regression-Assisted Two-Sample Test



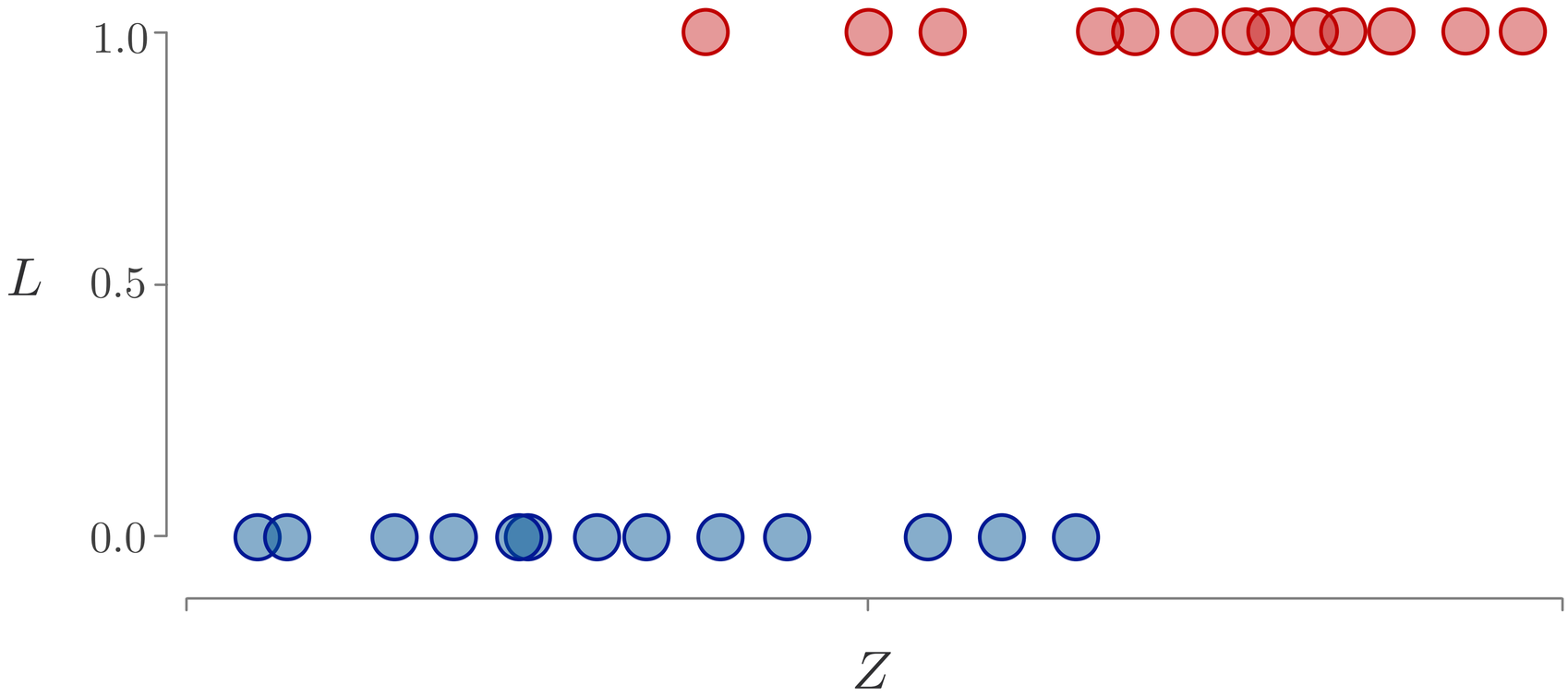
Regression-Assisted Two-Sample Test



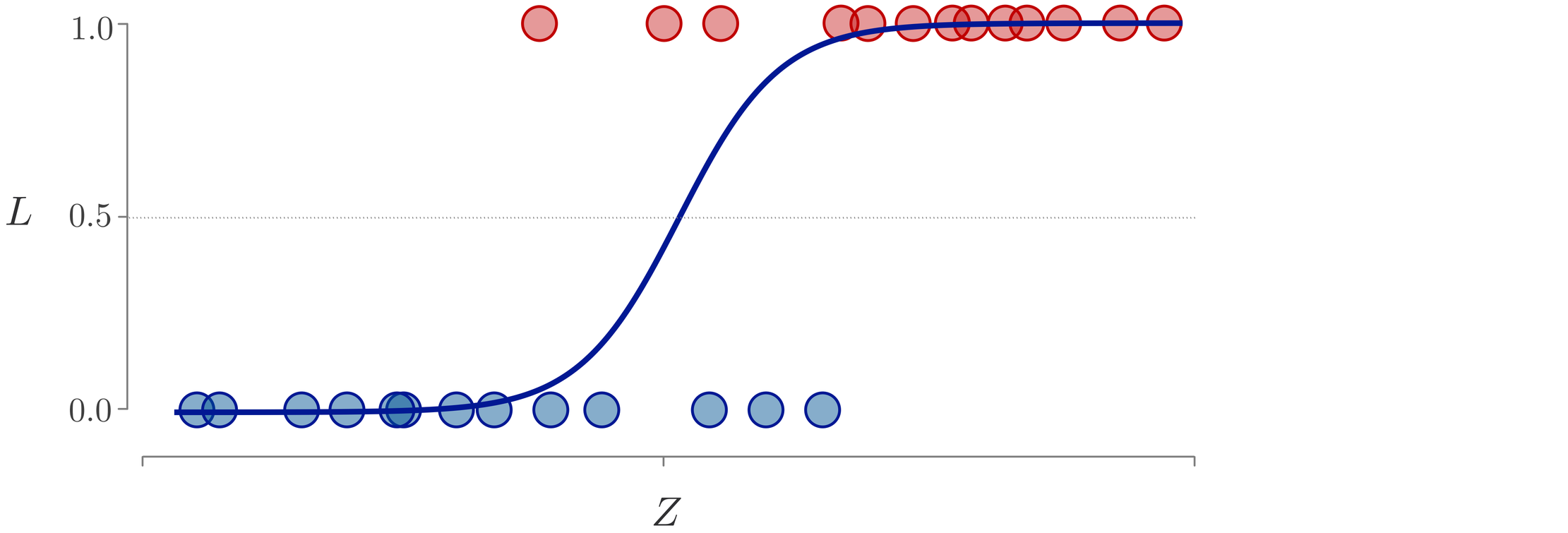
Regression-Assisted Two-Sample Test



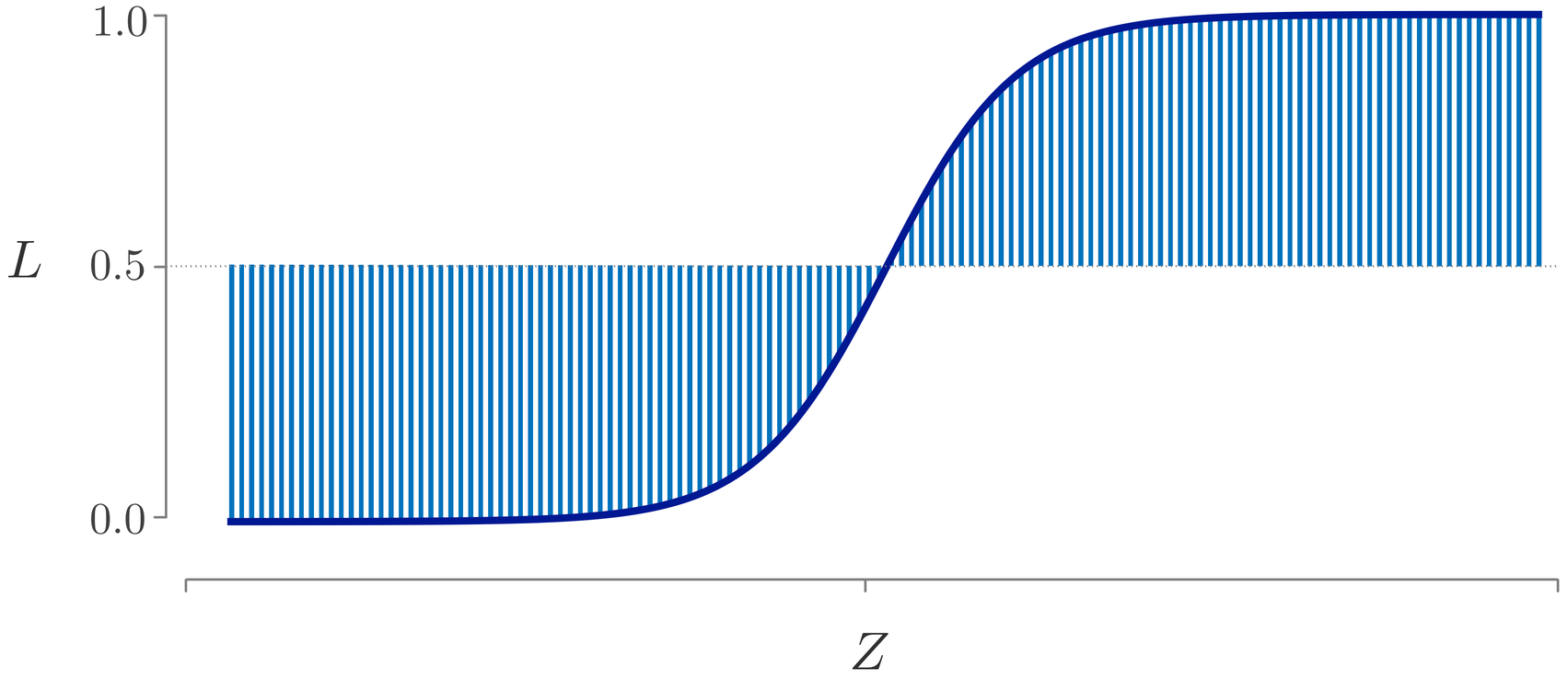
Regression-Assisted Two-Sample Test



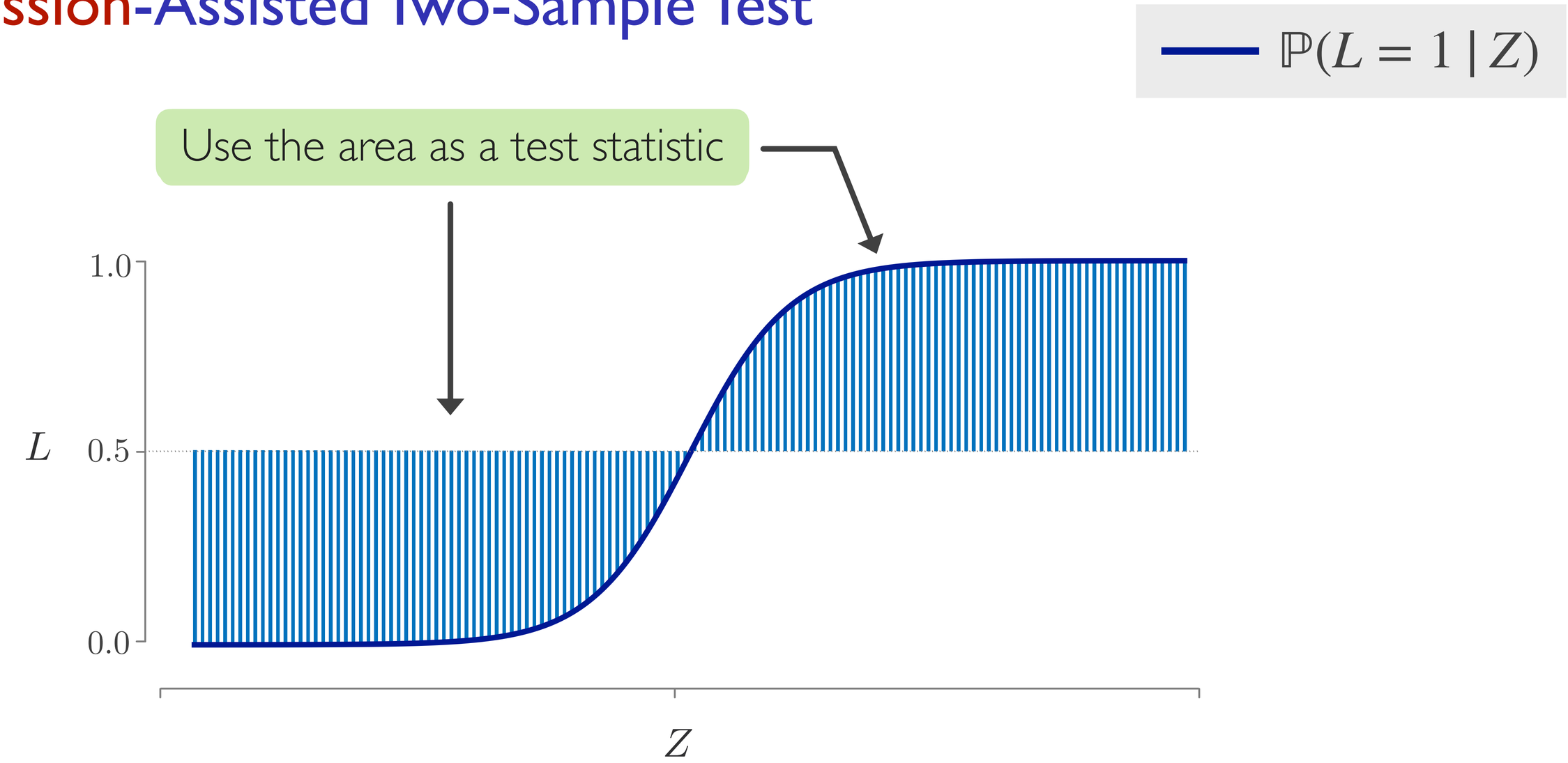
Regression-Assisted Two-Sample Test



Regression-Assisted Two-Sample Test



Regression-Assisted Two-Sample Test



Theoretical Results

Main goal: Maximize the **power** while controlling the **type I error**

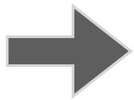
(True Positive) (False Positive)

Theoretical Results

Main goal: Maximize the **power** while controlling the **type I error**

(True Positive)

(False Positive)

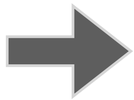


"Permutation method"

Theoretical Results

Main goal: Maximize the **power** while controlling the **type I error**

(True Positive) (False Positive)



"Permutation method"

Kim, Ramdas, Singh, Wasserman (**AoS**)

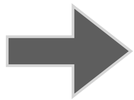
We derived the **exact asymptotic power** expression and showed that it achieves the **minimax optimal separation rate**!

$$\mathbb{E}[\phi_{\mathrm{Acc}}] = \Phi\left(-z_\alpha + \frac{n\delta^\top \Sigma^{-1}\delta}{\sqrt{32\pi d}}\right) + o(1)$$

Theoretical Results

Main goal: Maximize the **power** while controlling the **type I error**

(True Positive) (False Positive)



"Permutation method"

Kim, Ramdas, Singh, Wasserman (**AoS**)

We derived the **exact asymptotic power** expression and showed that it achieves the **minimax optimal separation rate**!

$$\mathbb{E}[\phi_{\mathrm{Acc}}] = \Phi\left(-z_\alpha + \frac{n\delta^\top \Sigma^{-1}\delta}{\sqrt{32\pi d}}\right) + o(1)$$

Kim, Lee, Lei (**EJS**)

We made a **connection** between the mean squared **error of regression** and statistical **power of the resulting test**!

$$\sup_{m\in\mathscr{M}} \mathbb{E}\int \left(\widehat{m}(x) - m(X)\right)^2 dP(x) \lesssim \delta_n$$

$$\Longrightarrow \inf_{P_X,Q_Y\in H_1(\delta_n)} \mathbb{E}_{P_X,Q_Y}[\phi_{\mathrm{reg}}] \geq 1-\beta$$

Theoretical Results

Main goal: Maximize the **power** while controlling the **type I error**

(True Positive) (False Positive)

"Permutation method"

Kim, Ramdas, Singh, Wasserman (**AoS**)

We derived the **exact asymptotic power** expression and showed that it achieves the **minimax optimal separation rate**!

$$\mathbb{E}[\phi_{\mathrm{Acc}}] = \Phi\left(-z_\alpha + \frac{n\delta^\top \Sigma^{-1}\delta}{\sqrt{32\pi d}}\right) + o(1)$$

Kim, Lee, Lei (**EJS**)

We made a **connection** between the mean squared **error of regression** and statistical **power of the resulting test**!

$$\sup_{m\in\mathscr{M}} \mathbb{E}\int \left(\widehat{m}(x) - m(X)\right)^2 dP(x) \lesssim \delta_n$$

$$\Longrightarrow \inf_{P_X,Q_Y\in H_1(\delta_n)} \mathbb{E}_{P_X,Q_Y}[\phi_{\mathrm{reg}}] \geq 1-\beta$$

"Key Takeaway"

Accurate prediction

Accurate inference

Thank you