Permutation Methods for Comparing Distributions

Ilmun Kim

Department of Mathematical Sciences
KAIST

Outline of this talk

(1) Introduction to Hypothesis Testing

(2) Permutation Tests

(3) Methods: Regression and Classification-based Approaches

(4) Theory: Power Analysis

(5) Summary

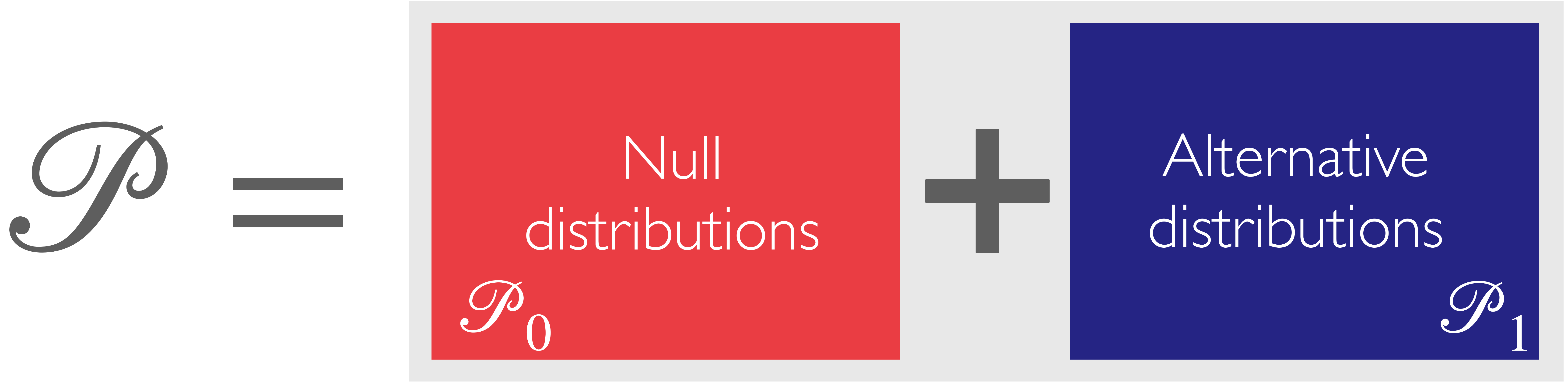
Recap: Statistical Hypothesis Testing

- Consider a class of distributions $\mathscr{P} = \mathscr{P}_0 \cup \mathscr{P}_1$

$\mathscr{P} =$ Null distributions $\mathscr{P}_0$ + Alternative distributions $\mathscr{P}_1$

Recap: Statistical Hypothesis Testing

- Consider a class of distributions $\mathscr{P} = \mathscr{P}_0 \cup \mathscr{P}_1$



- Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P \in \mathscr{P}$, **our goal** is to determine

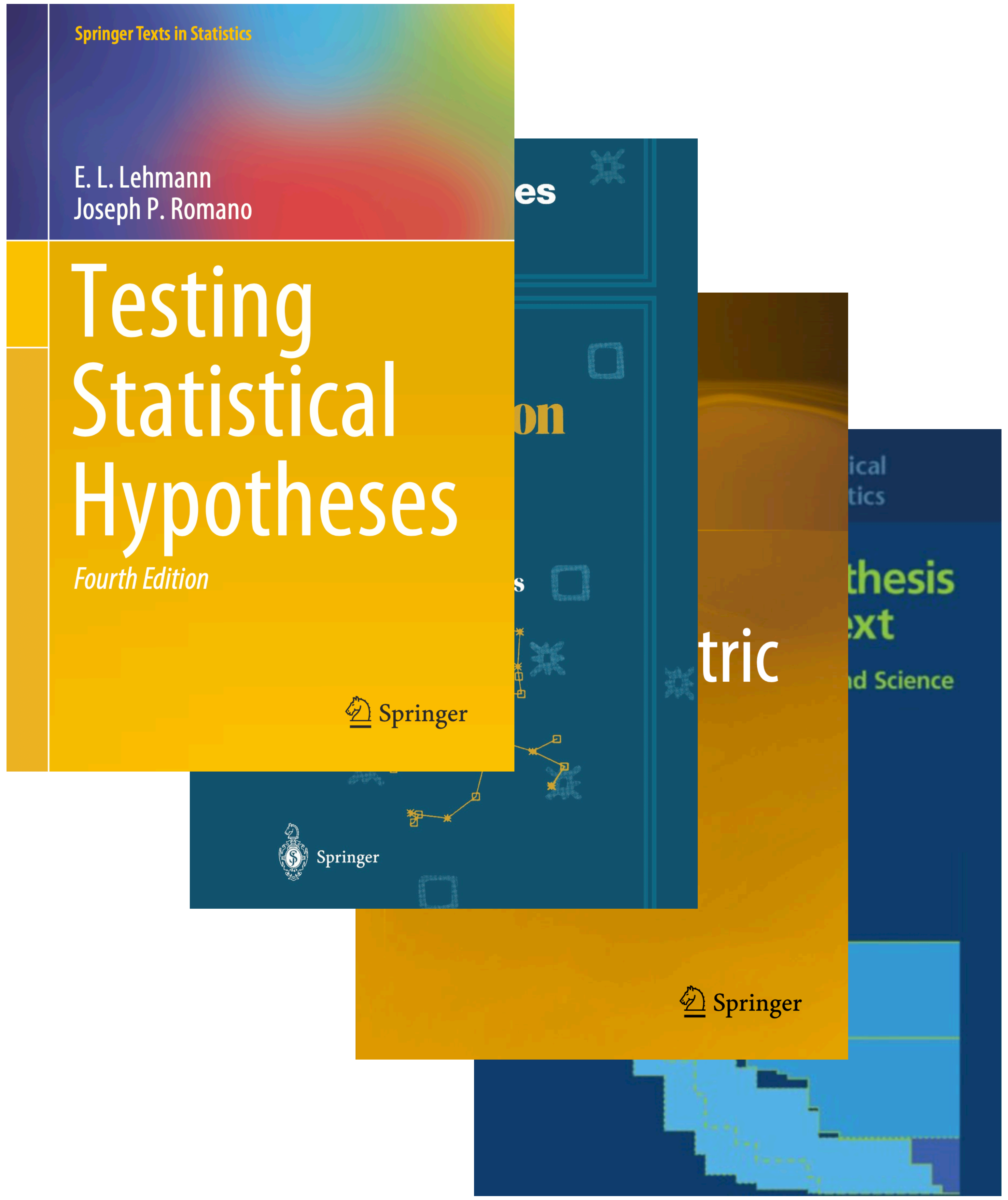
$$H_0 : P \in \mathscr{P}_0 \quad \text{versus} \quad H_1 : P \in \mathscr{P}_1$$

Recap: Statistical Hypothesis Testing

Examples include

- Mean testing
- Covariance testing
- Testing for regression models
- Two-sample testing
- Independence testing

⋮

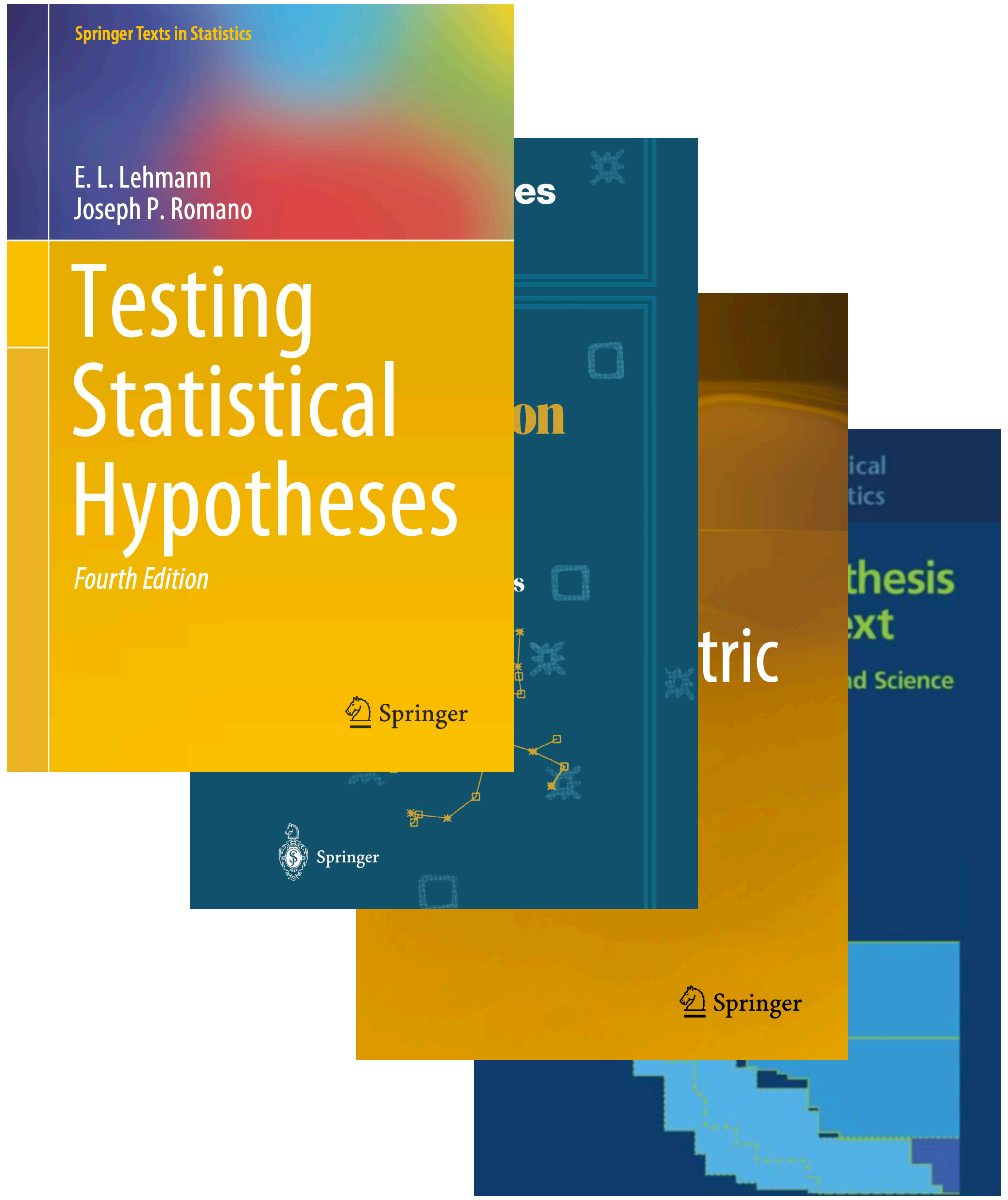


Recap: Statistical Hypothesis Testing

Examples include

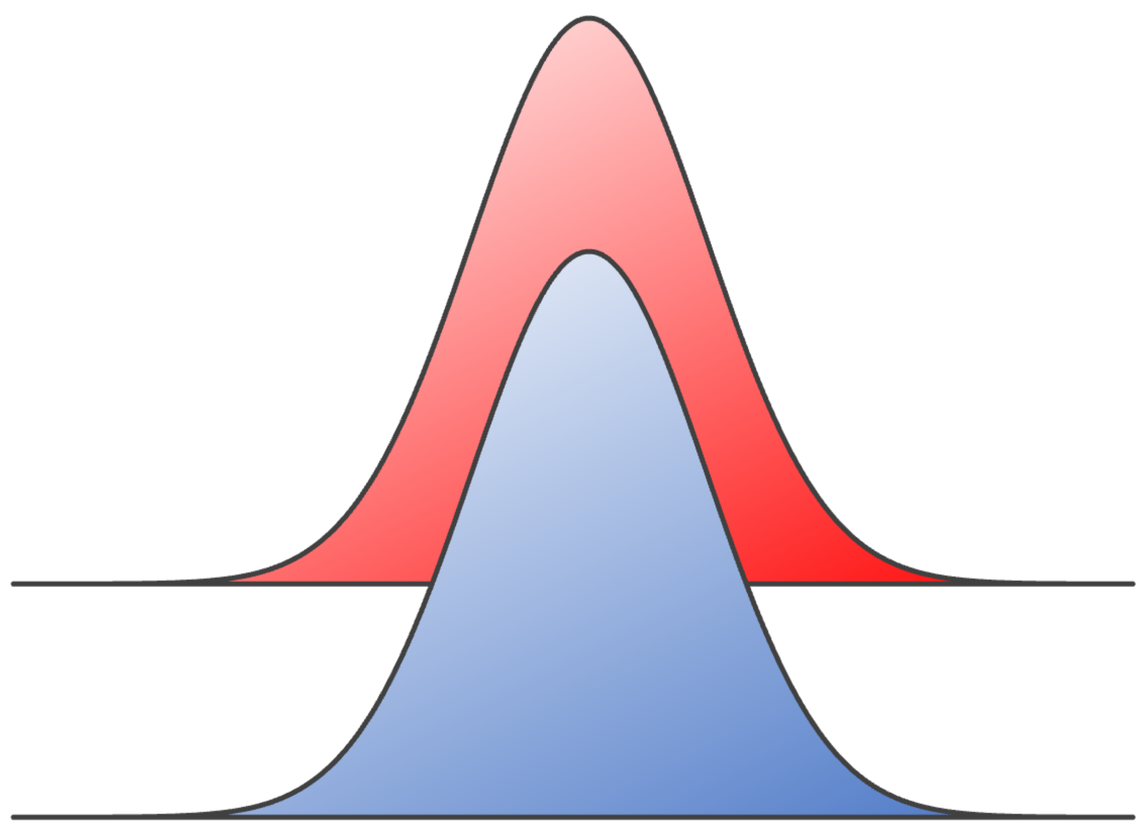
- Mean testing
- Covariance testing
- Testing for regression models
- Two-sample testing
- Independence testing

⋮



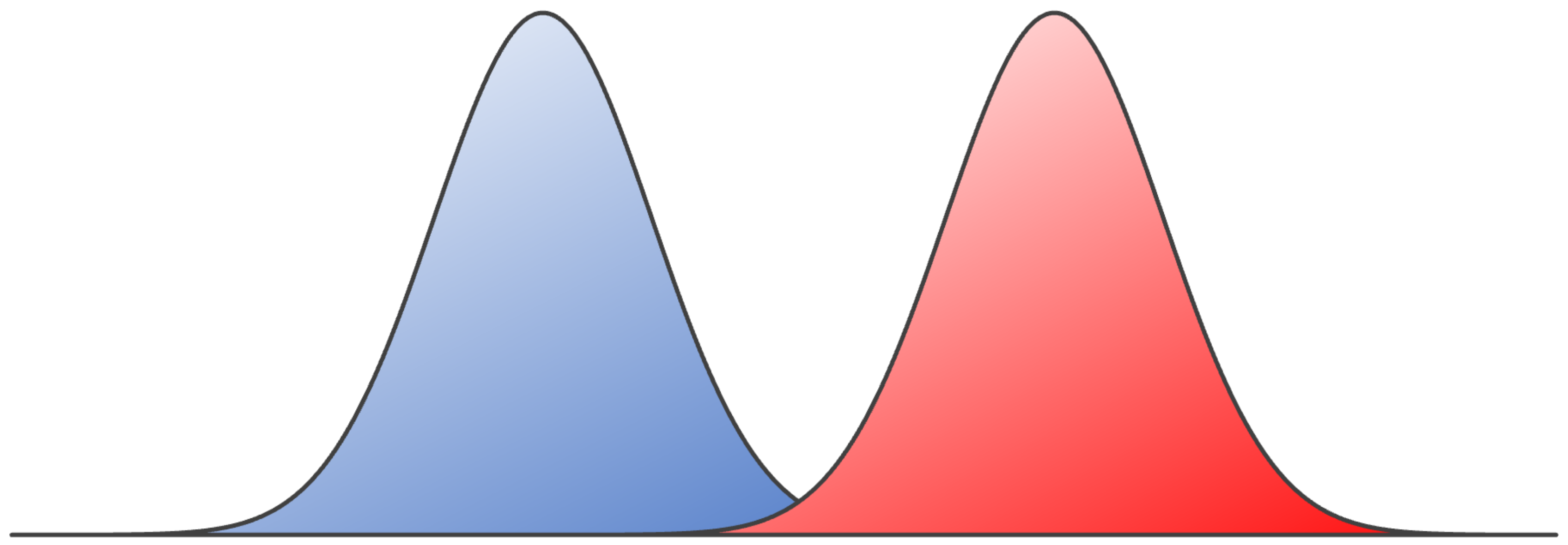
Two-Sample Problem

- 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



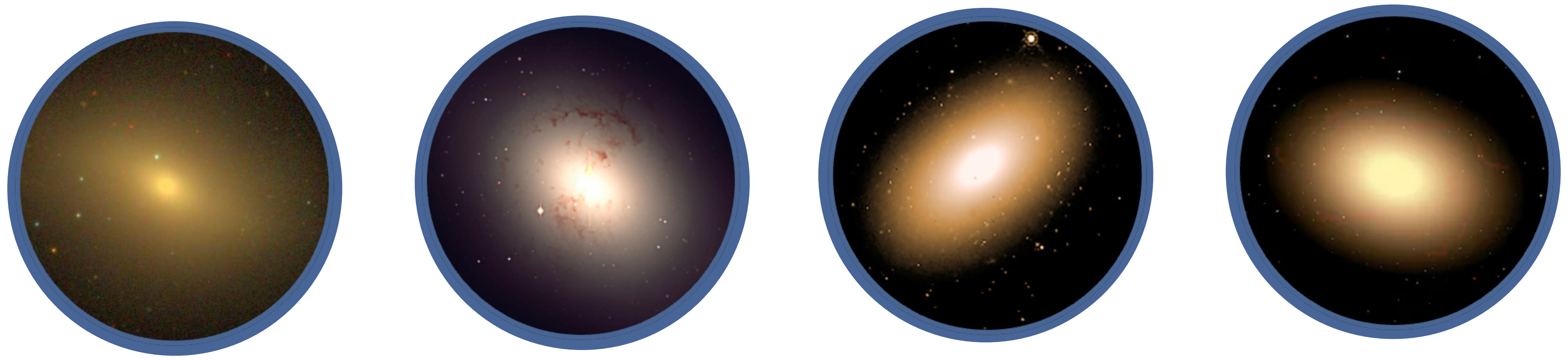
$H_0 : P_X = Q_Y$

versus



$H_1 : P_X \neq Q_Y$

Applications: Astronomy



High-mass versus Low-mass galaxies

Applications: Generative Models

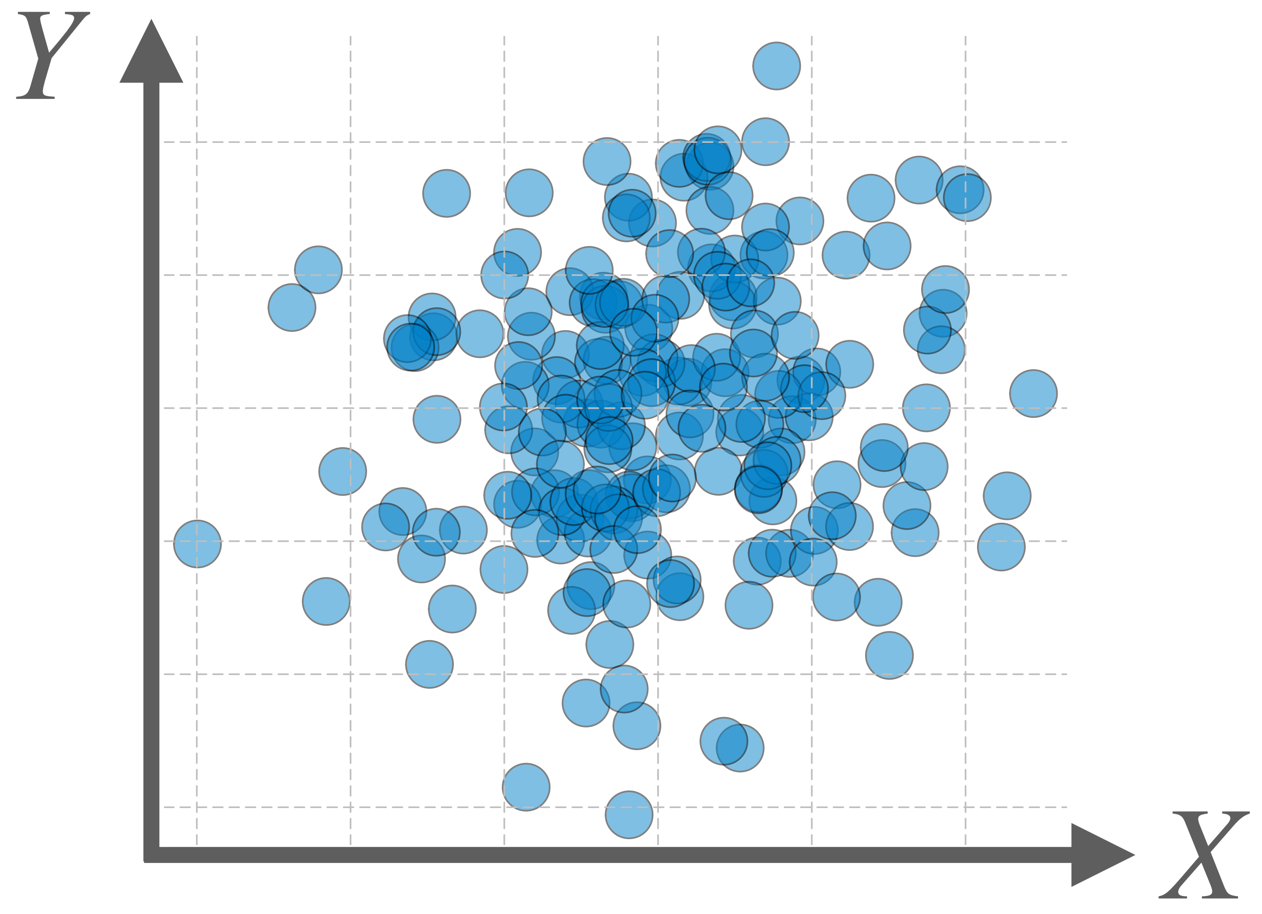
$\sim P_{\text{real}}$

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

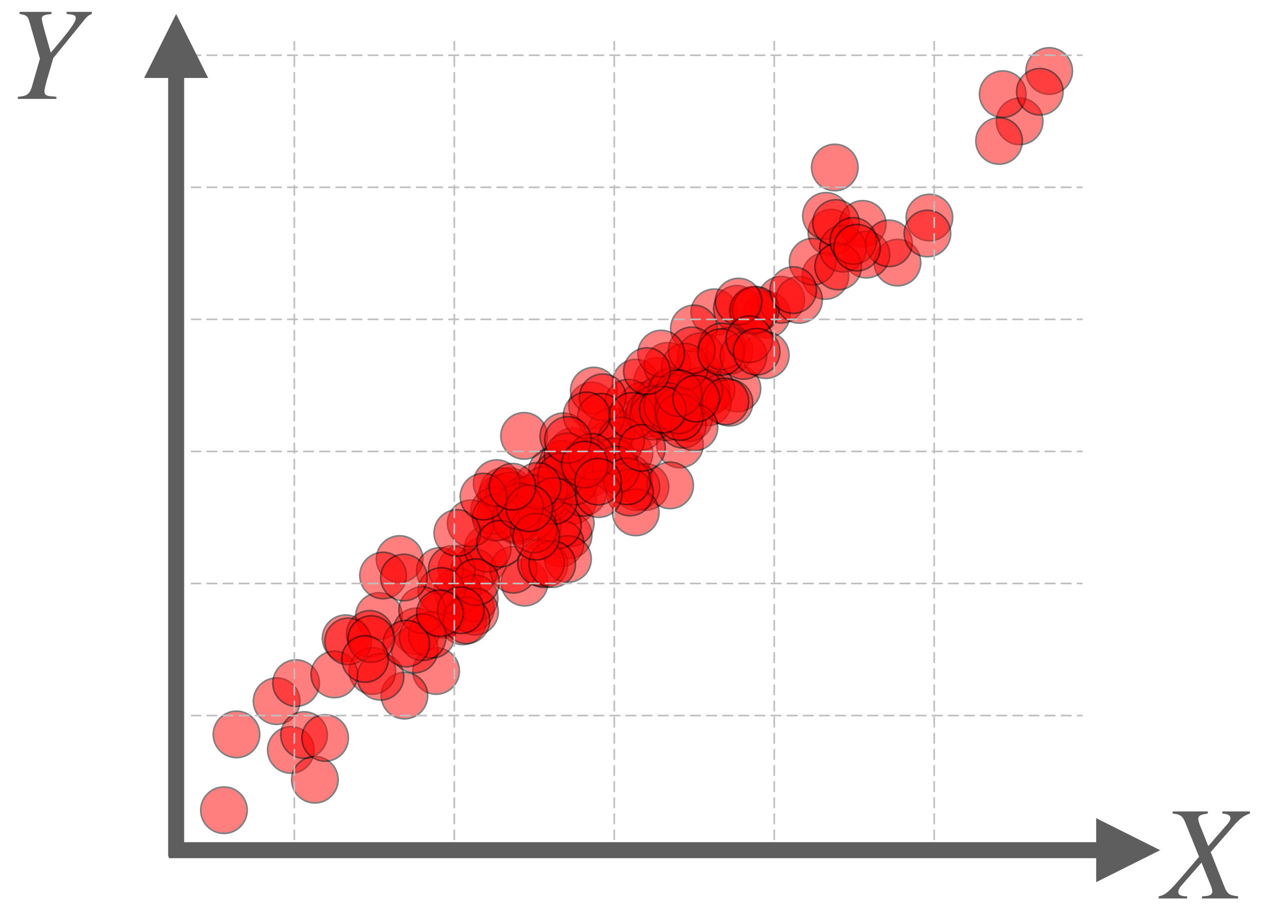
Independence Testing Problem

- Given $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{\text{i.i.d.}}{\sim} P_{X,Y}$

  we want to test whether



versus



$H_0 : P_{X,Y} = P_X P_Y$

$H_1 : P_{X,Y} \neq P_X P_Y$

Applications: Multimodal Learning

X	Y
	The image shows a brown horse with a white blaze on its face, trotting in an open grassy field.
	The image shows a majestic tiger resting on the grass, accompanied by a young tiger cub.
	The image depicts a bald eagle in mid-flight, soaring through the sky with its wings fully spread.

Recap: Statistical Hypothesis Testing

- To tackle the testing problem, we construct a test function

$$\phi : \{X_1, \ldots, X_n\} \mapsto \{0,1\}$$

Recap: Statistical Hypothesis Testing

- To tackle the testing problem, we construct a test function

$$\phi : \{X_1, \ldots, X_n\} \mapsto \{0,1\}$$

- We reject the null if $\phi = 1$ and accept the null if $\phi = 0$

Recap: Statistical Hypothesis Testing

- To tackle the testing problem, we construct a test function

$$\phi : \{X_1, \ldots, X_n\} \mapsto \{0,1\}$$

- We reject the null if $\phi = 1$ and accept the null if $\phi = 0$

- There are two types of error we care about

(Uniform) **Type I error:** $\sup_{P \in \mathscr{P}_0} \mathbb{P}_P(\phi = 1)$

(Uniform) **Type II error:** $\sup_{P \in \mathscr{P}_1} \mathbb{P}_P(\phi = 0)$

Recap: Statistical Hypothesis Testing

- Frequently, the **Type I error** is more **serious** than the **Type II error**

(Uniform) **Type I error:** $\sup_{P \in \mathscr{P}_0} \mathbb{P}_P(\phi = 1)$

(Uniform) **Type II error:** $\sup_{P \in \mathscr{P}_1} \mathbb{P}_P(\phi = 0)$

Recap: Statistical Hypothesis Testing

- Frequently, the **Type I error** is more **serious** than the **Type II error**

- Hence we **first** control the **Type I error** by level α

(Uniform) Type I error: $\sup_{P \in \mathscr{P}_0} \mathbb{P}_P(\phi = 1)$

(Uniform) Type II error: $\sup_{P \in \mathscr{P}_1} \mathbb{P}_P(\phi = 0)$

Recap: Statistical Hypothesis Testing

- Frequently, the **Type I error** is more **serious** than the **Type II error**
- Hence we **first** control the **Type I error** by level α
- And then try to **minimize** the **Type II error** (or maximize the power)

(Uniform) Type I error: $\sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\phi = 1)$

(Uniform) Type II error: $\sup_{P \in \mathcal{P}_1} \mathbb{P}_P(\phi = 0)$

Recap: Statistical Hypothesis Testing

A typical way of constructing a test function

Step 1

Compute a test statistic

e.g.,

$$T_n = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

$$\chi_n^2 = \sum_k \frac{(O_k - E_k)^2}{E_k}$$

Recap: Statistical Hypothesis Testing

A typical way of constructing a test function

Step I

Compute a test statistic

e.g.,

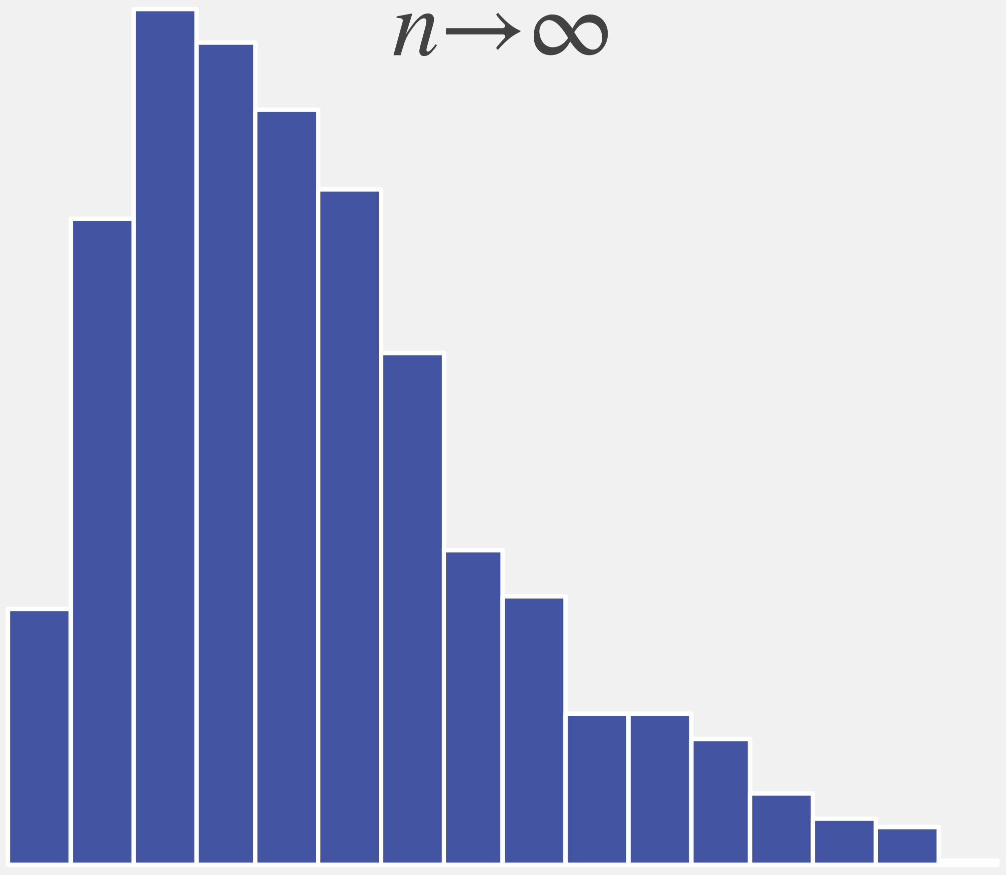
$$T_n = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

$$\chi_n^2 = \sum_k \frac{(O_k - E_k)^2}{E_k}$$

Step II

Derive the null distribution

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



Recap: Statistical Hypothesis Testing

A typical way of constructing a test function

Step I

Compute a test statistic

e.g.,

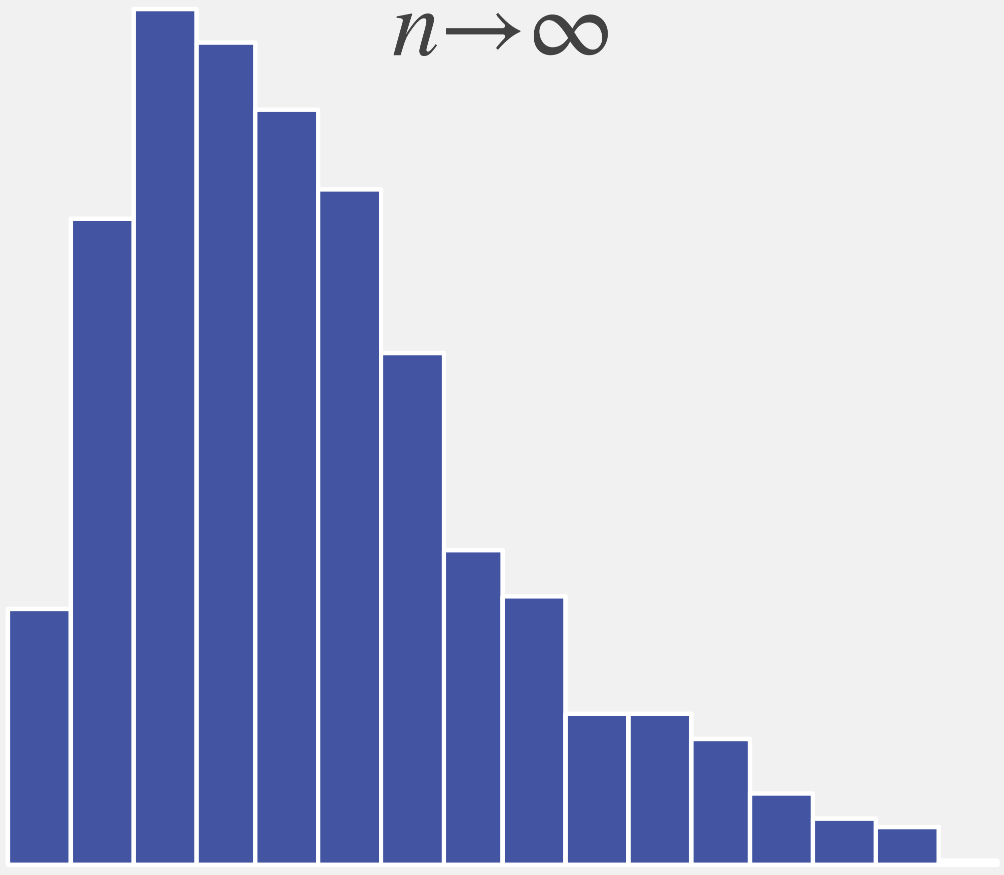
$$T_n = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

$$\chi_n^2 = \sum_k \frac{(O_k - E_k)^2}{E_k}$$

Step II

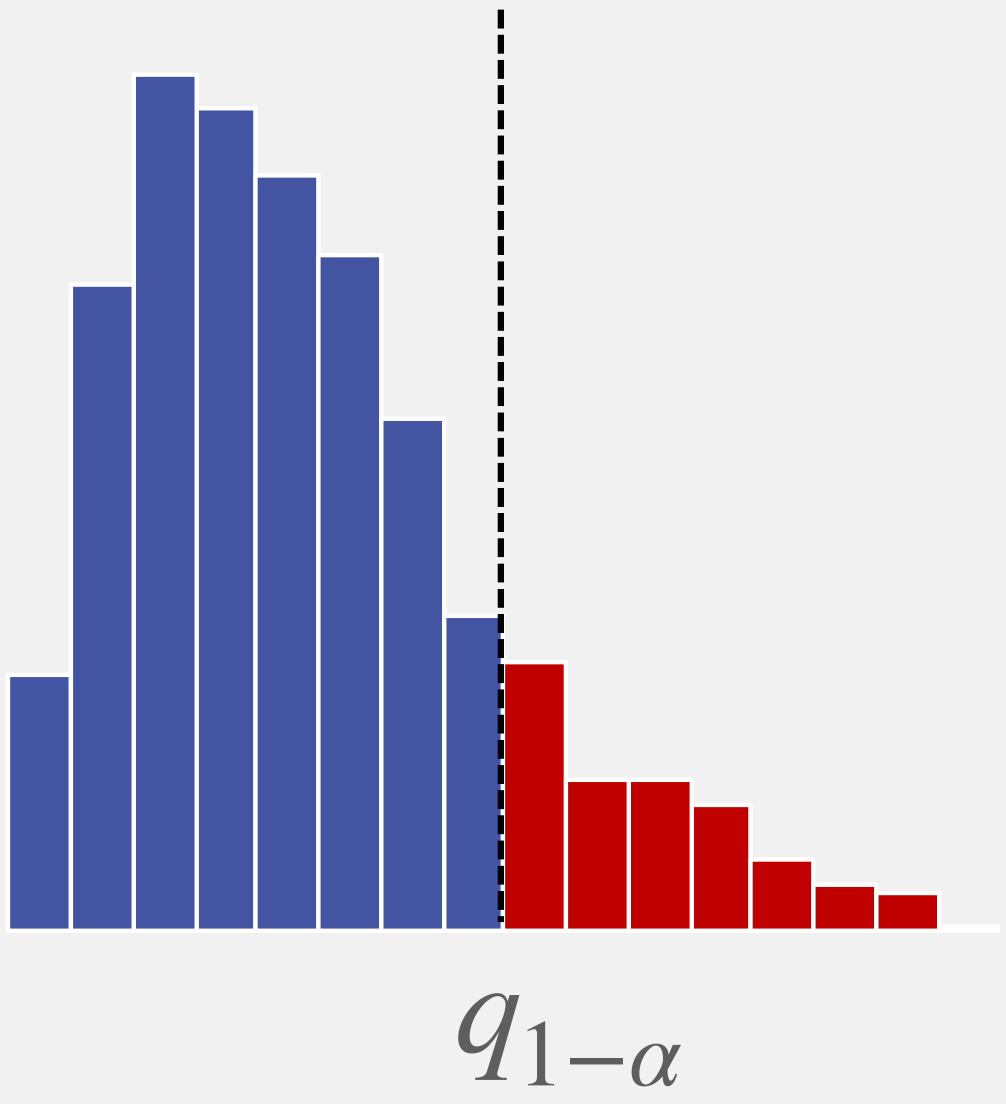
Derive the null distribution

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



Step III

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



New Challenges

- Modern data are often **large**, **high-dimensional** and **complex**

New Challenges

- Modern data are often **large**, **high-dimensional** and **complex**
- Classical **asymptotic** approaches suffer from
    - Inflated type I error
    - Suboptimal power
    - Strong assumptions

Example: χ^2-statistic in high-dimensions

Observed values

	Cat	Dog	Lion	Duck
Male	O_{11}	O_{12}	O_{13}	O_{14}
Female	O_{21}	O_{22}	O_{23}	O_{24}

Question:

Is there a significant association between gender and favorite animals?

H_0 : Gender $\perp\!\!\!\perp$ Favorite Animals

Example: χ^2-statistic in high-dimensions

Observed values

	Cat	Dog	Lion	Duck
Male	O_{11}	O_{12}	O_{13}	O_{14}
Female	O_{21}	O_{22}	O_{23}	O_{24}

Question:

Is there a significant association between gender and favorite animals?

H_0 : Gender $\perp\!\!\!\perp$ Favorite Animals

$\Rightarrow$ Pearson's chi-squared test

Example: χ^2-statistic in high-dimensions

Observed values

	Cat	Dog	Lion	Duck
Male	O_{11}	O_{12}	O_{13}	O_{14}
Female	O_{21}	O_{22}	O_{23}	O_{24}

versus

Expected values

	Cat	Dog	Lion	Duck
Male	E_{11}	E_{12}	E_{13}	E_{14}
Female	E_{21}	E_{22}	E_{23}	E_{24}

Example: χ^2-statistic in high-dimensions

Observed values

	Cat	Dog	Lion	Duck
Male	O_{11}	O_{12}	O_{13}	O_{14}
Female	O_{21}	O_{22}	O_{23}	O_{24}

versus

Expected values

	Cat	Dog	Lion	Duck
Male	E_{11}	E_{12}	E_{13}	E_{14}
Female	E_{21}	E_{22}	E_{23}	E_{24}

$$\chi^2 = \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Example: χ^2-statistic in high-dimensions

Observed values

	Cat	Dog	Lion	Duck
Male	O_{11}	O_{12}	O_{13}	O_{14}
Female	O_{21}	O_{22}	O_{23}	O_{24}

versus

Expected values

	Cat	Dog	Lion	Duck
Male	E_{11}	E_{12}	E_{13}	E_{14}
Female	E_{21}	E_{22}	E_{23}	E_{24}

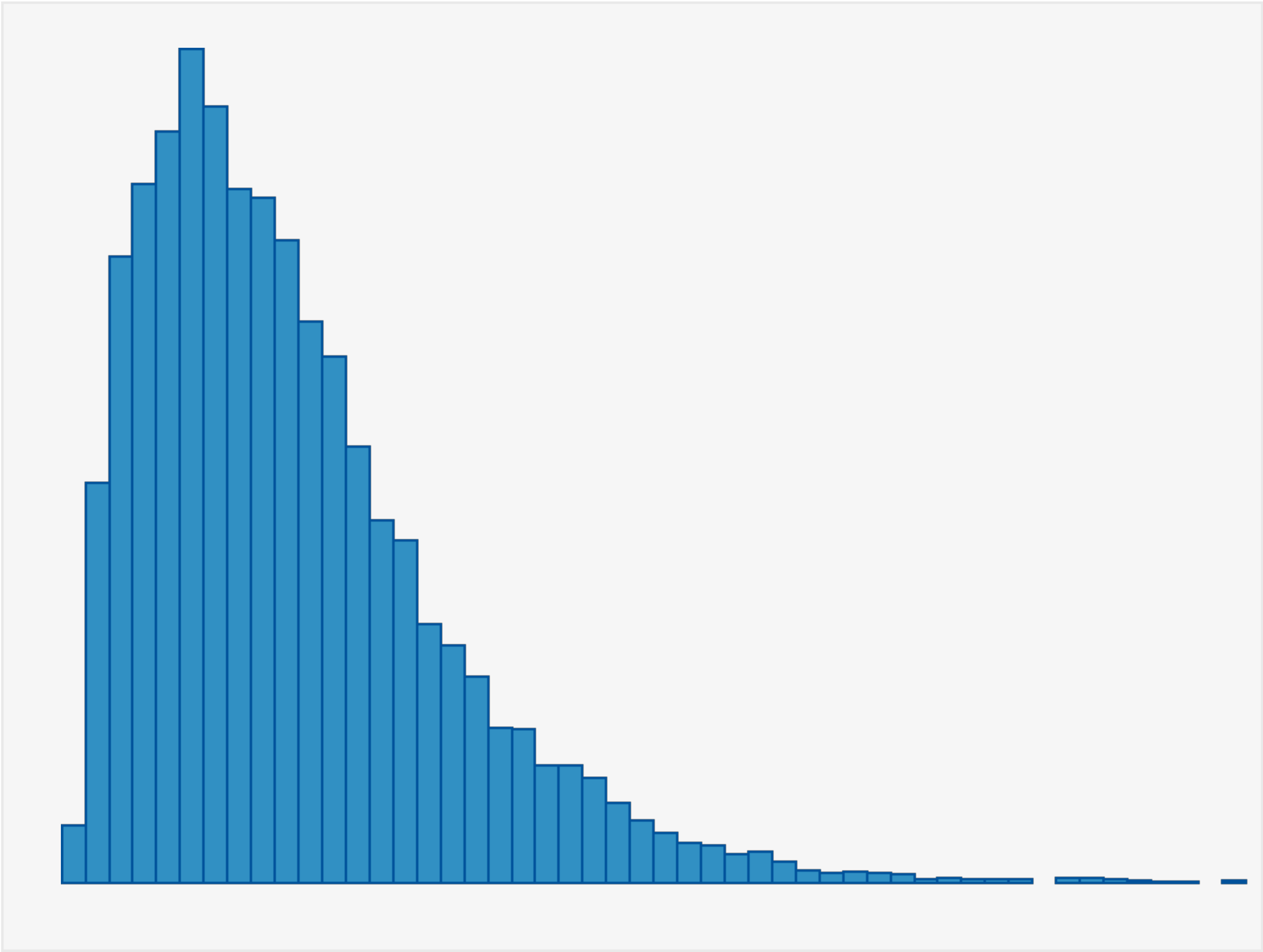
Classical asymptotic theory shows that under H_0

$$\chi^2 = \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{approx.}{\sim}$$

Chi-squared distribution with $(\ell_1 - 1)(\ell_2 - 1)$ d.f.

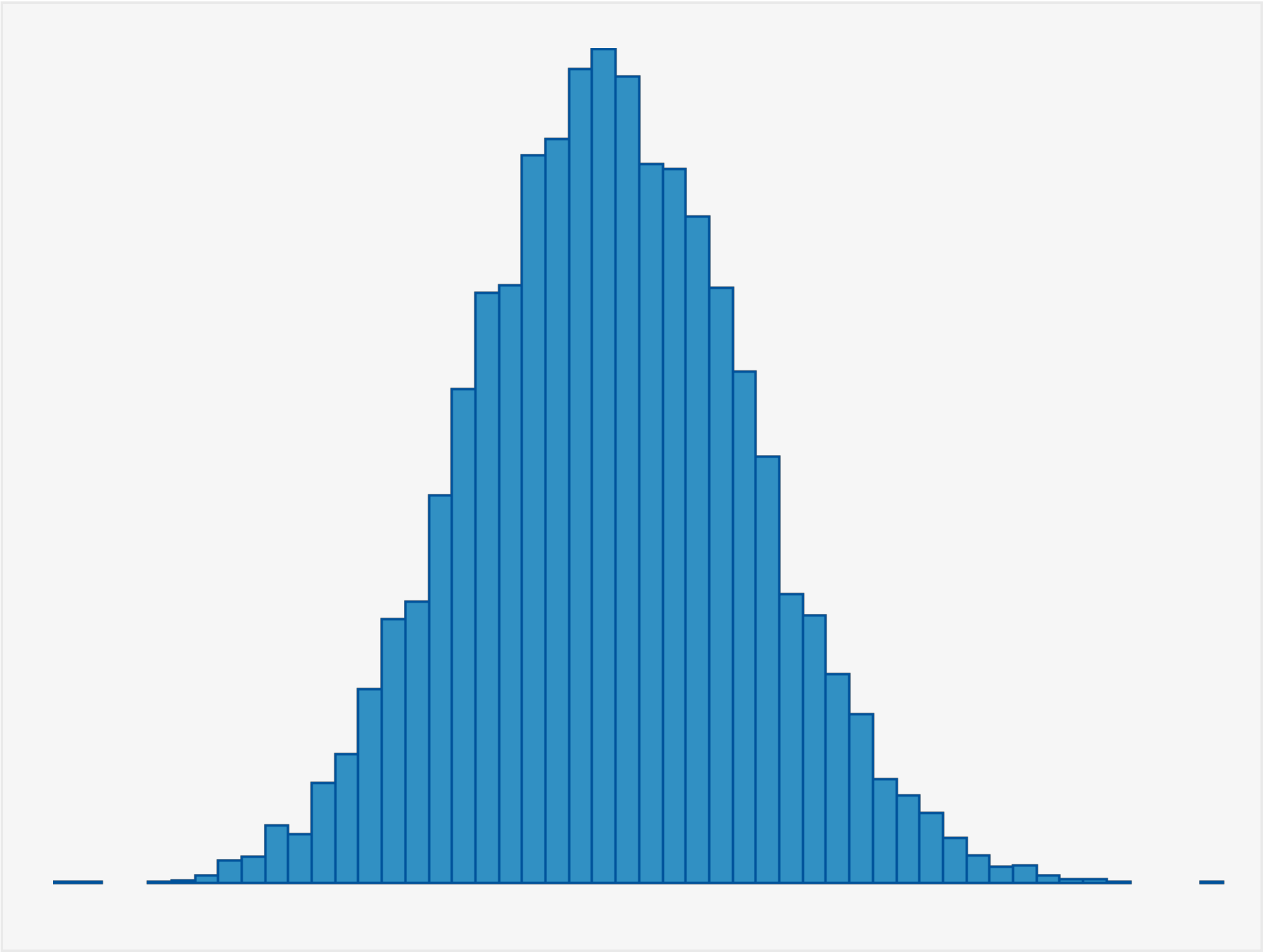
Example: χ^2-statistic in high-dimensions

Histograms of χ^2-statistic for testing independence based on $n = 100$



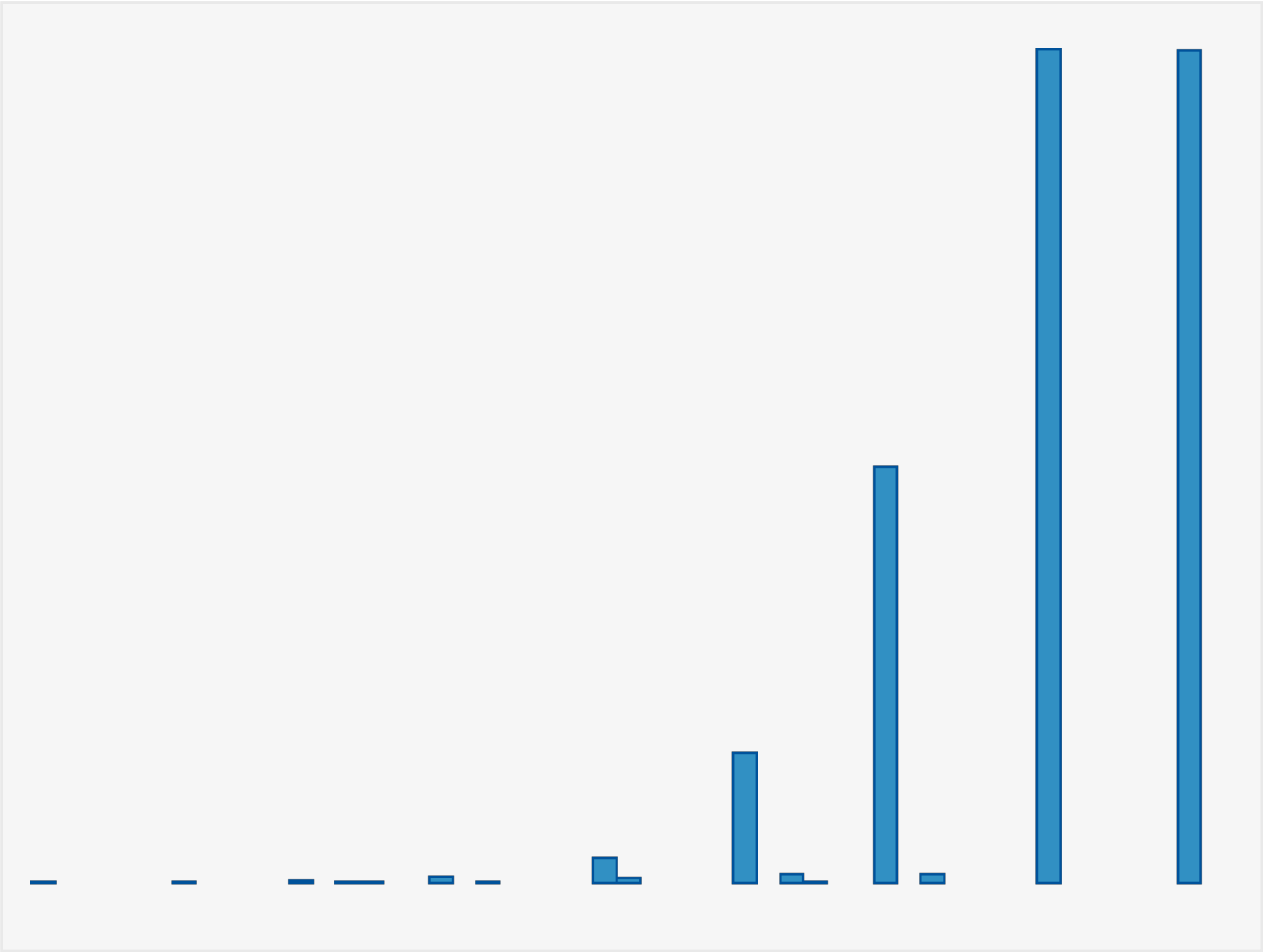
≈ Chi-square

$\ell_1 = \ell_2 = 3$



≈ Gaussian

$\ell_1 = \ell_2 = 50$



≈ Discrete

$\ell_1 = \ell_2 = 10{,}000$

We aim to develop **new** **theory** and **methods** for hypothesis testing that improve classical approaches

We aim to develop **new theory** and **methods** for hypothesis testing that improve classical approaches

Common keyword

“Permutation tests”

- Freeman, **Kim**, Lee (2019, MNRAS)
- **Kim**, Lee, Lei (2019, EJS)
- **Kim**, Balakrishnan, Wasserman (2021, AoS)
- **Kim**, Balakrishnan, Wasserman (2022, AoS)
- **Kim**, Ramdas, Singh, Wasserman (2021, AoS)
- **Kim** (2021, Bernoulli)
- Schrab, **Kim**, Guedj, Gretton (2022, NeurIPS)
- **Kim**, Neykov, Balakrishnan, Wasserman (2022, AoS)
- Schrab, **Kim**, Albert, Laurent, Guedj, Gretton (2023, JMRL)
- **Kim**, Neykov, Balakrishnan, Wasserman (2023, *submitted*)
- **Kim**, Schrab (2024, *submitted*)
- Schrab, **Kim** (2024, *submitted*)
- Choi, **Kim** (2024, *submitted*)

Permutation tests

- The use of permutation methods dates back to Fisher in 1935

Permutation tests

- The use of permutation methods dates back to Fisher in 1935

For i = 1 to B

Sample $\{X_1, \ldots, X_n\}$ → a. Permute: $X_{\pi_1}, \ldots, X_{\pi_n}$ b. Compute: T_i^π → Compare T_n with $\{T_1^\pi, \ldots, T_B^\pi\}$

Permutation tests

- The use of permutation methods dates back to Fisher in 1935

For i = 1 to B

Sample $\{X_1, \ldots, X_n\}$ → a. Permute: $X_{\pi_1}, \ldots, X_{\pi_n}$ b. Compute: T_i^{π} → Compare T_n with $\{T_1^{\pi}, \ldots, T_B^{\pi}\}$

- Permutation p-value: $$p_{\text{perm}} = \frac{1}{B+1}\left\{1 + \sum_{i=1}^{B} \mathbb{1}(T_i^{\pi} \geq T_n)\right\}$$

(Reject the null when $p_{\text{perm}} \leq \alpha$)

Key features of permutation tests

$$\sup_{P \in \mathscr{P}_0} \mathbb{P}_P(p_{\text{perm}} \le \alpha) \le \alpha \quad \begin{cases} \text{for any } \alpha \in (0,1) \\ \text{for any } n \ge 1 \end{cases}$$

Class of all null distributions

- **Uniform**, **non-asymptotic** type I error control

Key features of permutation tests

$$\sup_{P \in \mathscr{P}_0} \mathbb{P}_P(p_{\mathrm{perm}} \leq \alpha) \leq \alpha \quad \begin{cases} \text{for any } \alpha \in (0,1) \\ \text{for any } n \geq 1 \end{cases}$$

Class of all null distributions

- **Uniform**, **non-asymptotic** type I error control
- **Distribution-free** for any type of test statistics

Key features of permutation tests

$$\sup_{P \in \mathscr{P}_0} \mathbb{P}_P(p_{\mathrm{perm}} \le \alpha) \le \alpha \quad \begin{cases} \text{for any } \alpha \in (0,1) \\ \text{for any } n \ge 1 \end{cases}$$

Class of all null distributions

- **Uniform**, **non-asymptotic** type I error control
- **Distribution-free** for any type of test statistics
- It does not depend on **unspecified constants**

Key features of permutation tests

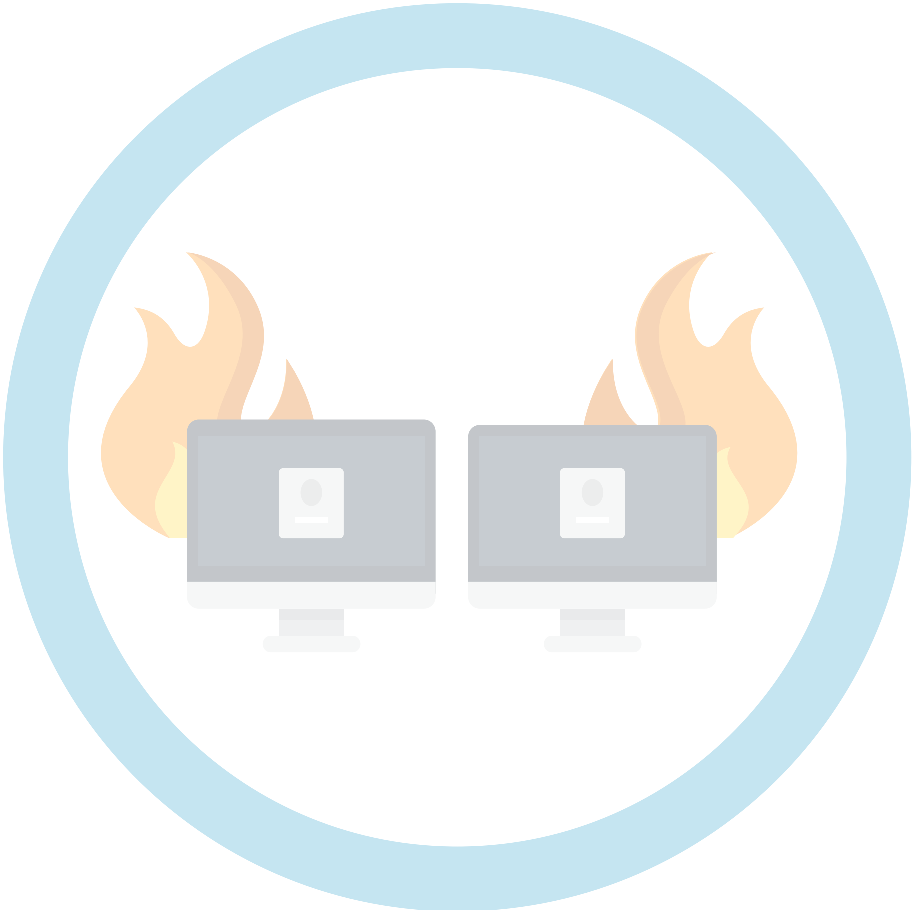
$$\sup_{P \in \mathscr{P}_0} \mathbb{P}_P(p_{\text{perm}} \leq \alpha) \leq \alpha \quad \begin{cases} \text{for any } \alpha \in (0,1) \\ \text{for any } n \geq 1 \end{cases}$$

Class of all null distributions

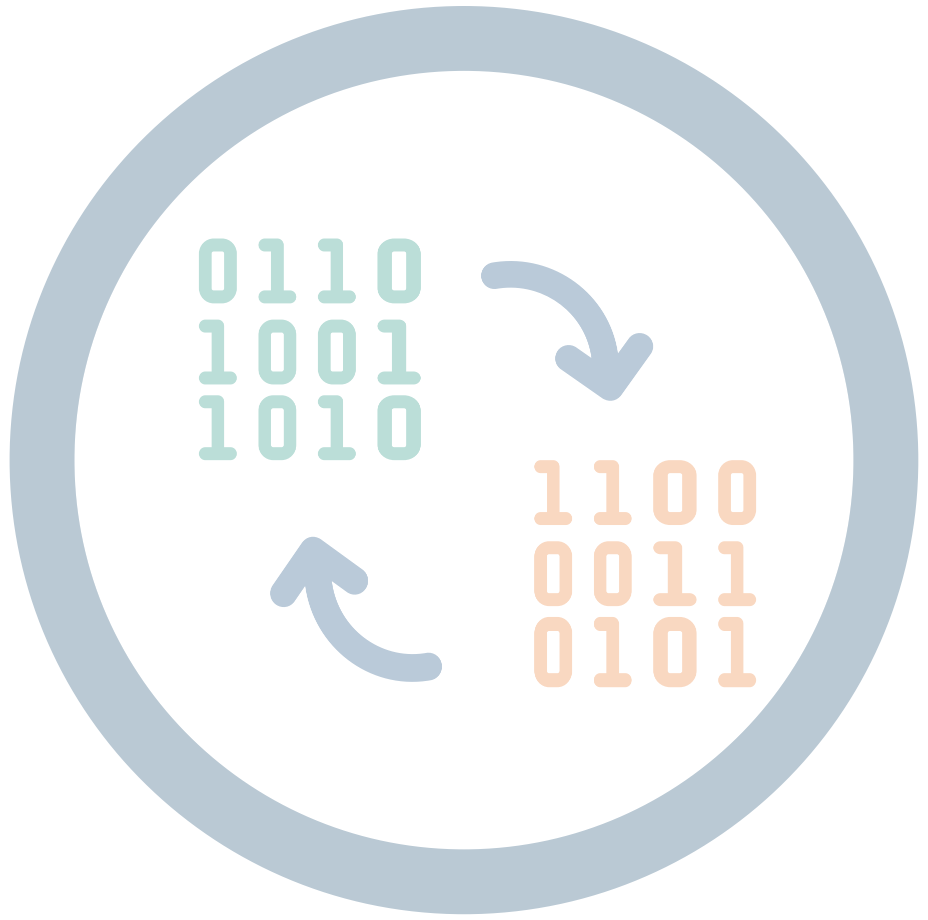
- **Uniform**, **non-asymptotic** type I error control
- **Distribution-free** for any type of test statistics
- It does not depend on **unspecified constants**
- All we need is $\{X_1, \ldots, X_n\}$ are **exchangeable** under the null

Three main challenges of the permutation approach

Power Analysis



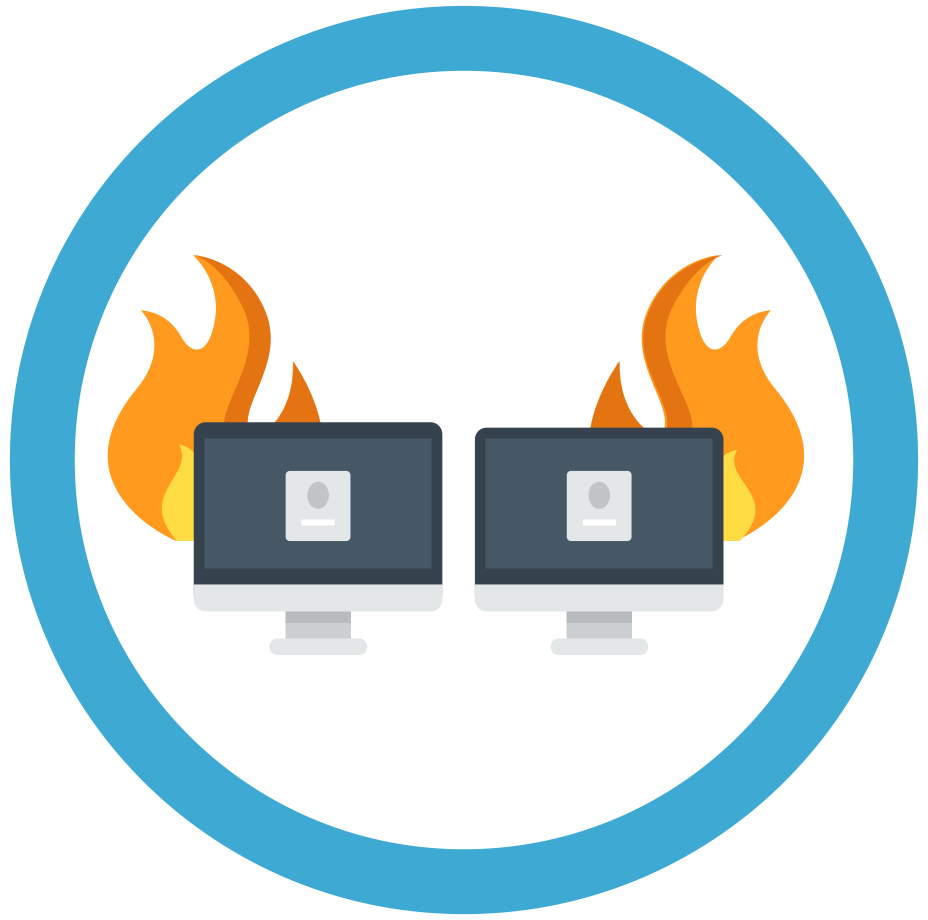
Computational Complexity



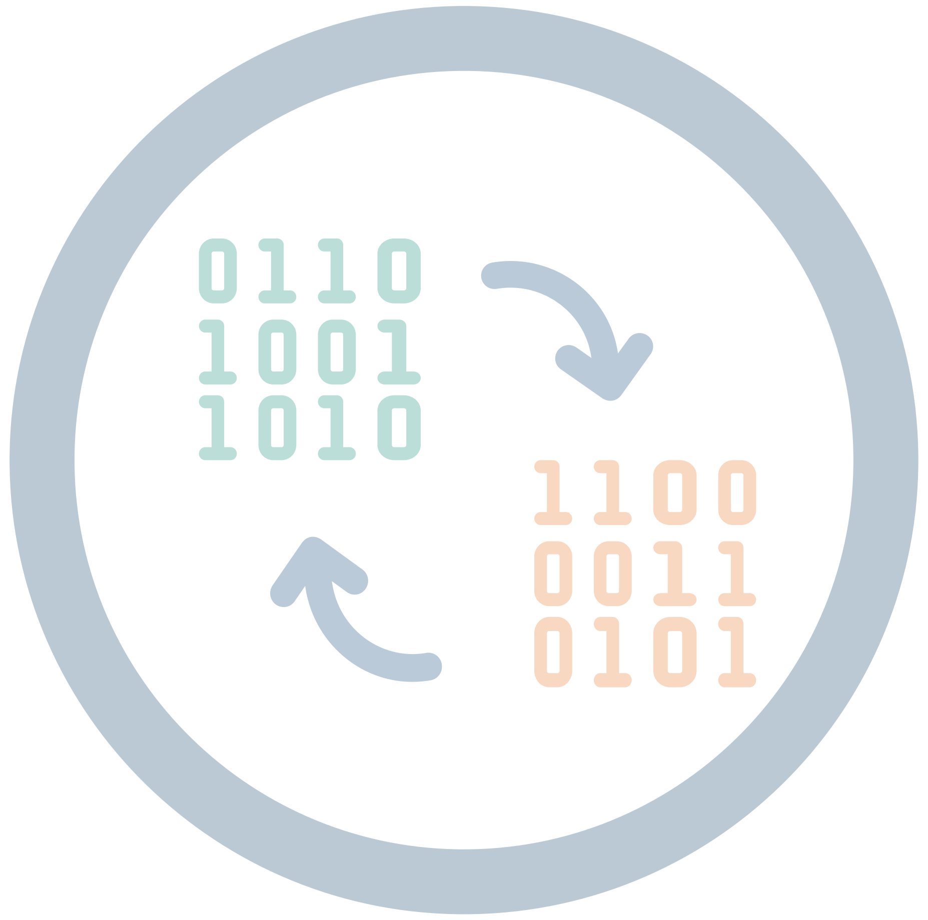
Non-exchangeable Data

Three main challenges of the permutation approach

Power Analysis



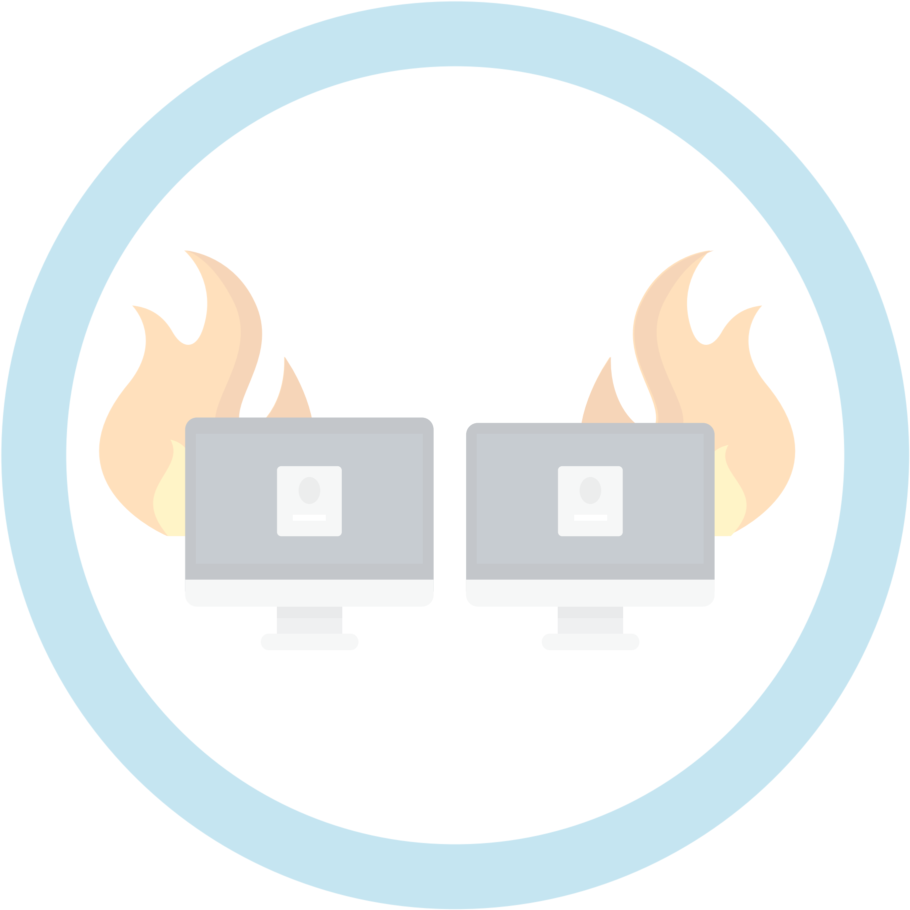
Computational Complexity



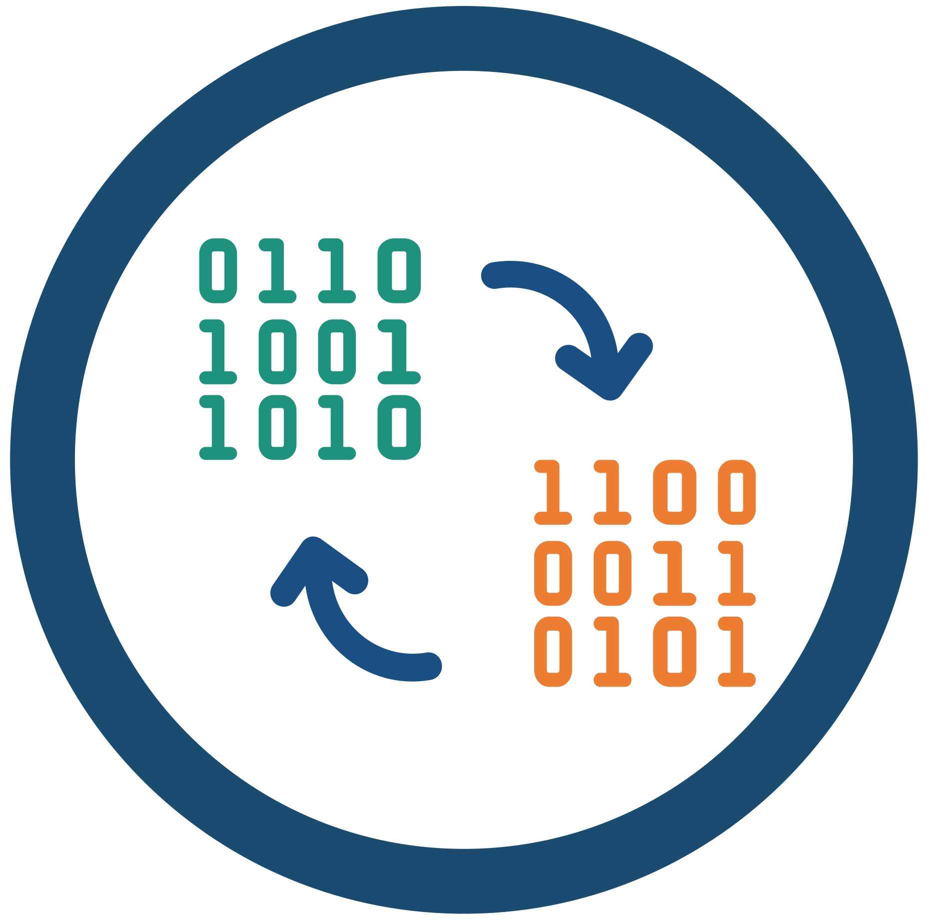
Non-exchangeable Data

Three main challenges of the permutation approach

Power Analysis



Computational
Complexity



Non-exchangeable
Data

Part I. Methodological Contributions

Part II. Theoretical Contributions

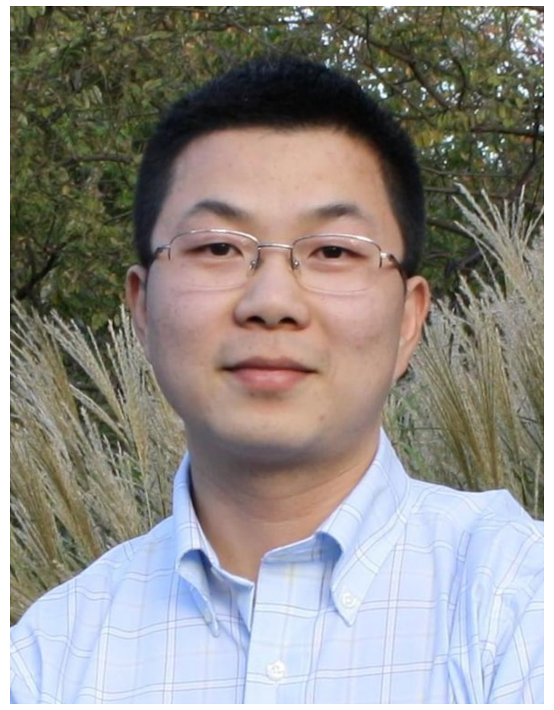
Electronic Journal of Statistics
Vol. 13 (2019) 5253–5305
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1648

Global and local two-sample tests via regression

Ilmun Kim, Ann B. Lee, and Jing Lei

Carnegie Mellon University
Department of Statistics and Data Science
5000 Forbes Avenue
Pittsburgh, PA 15213

Ann Lee
(CMU)



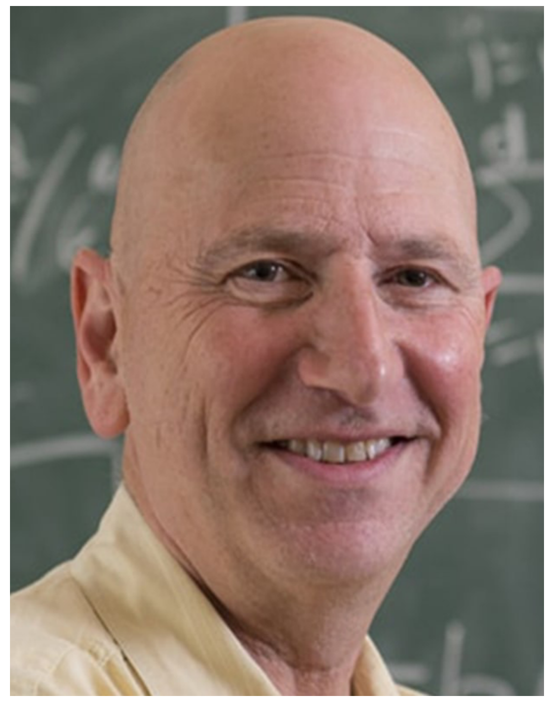
Jing Lei
(CMU)

The Annals of Statistics
2021, Vol. 49, No. 1, 411–434
https://doi.org/10.1214/20-AOS1962

CLASSIFICATION ACCURACY AS A PROXY FOR TWO-SAMPLE TESTING

BY ILMUN KIM[1,*], AADITYA RAMDAS[1,†], AARTI SINGH[1,‡] AND LARRY WASSERMAN[1,§]

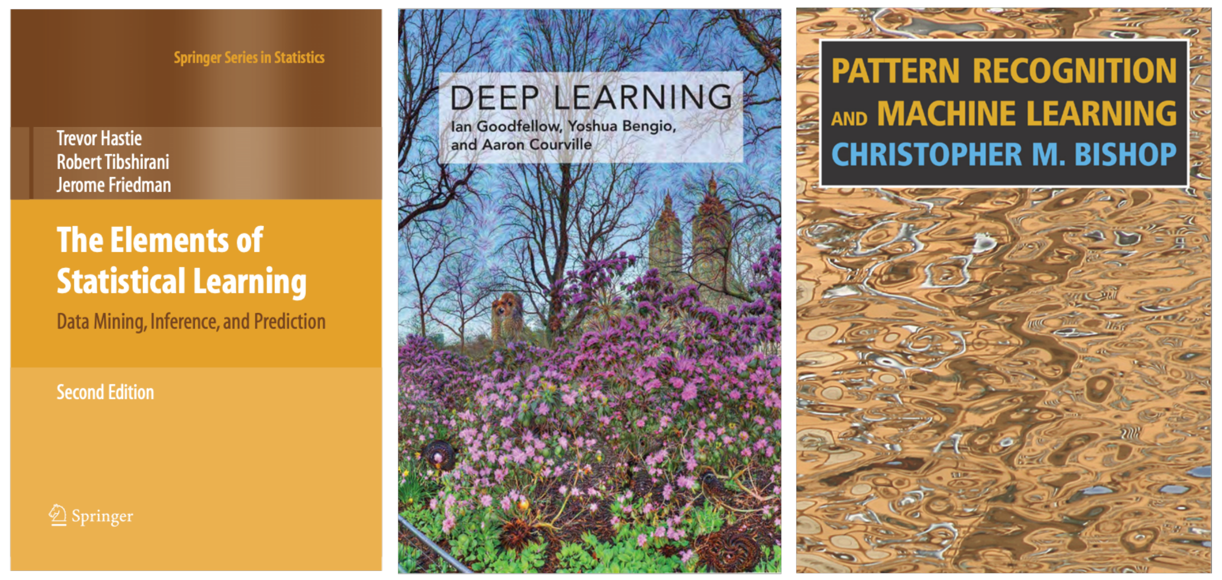
Aaditya Ramdas
(CMU)



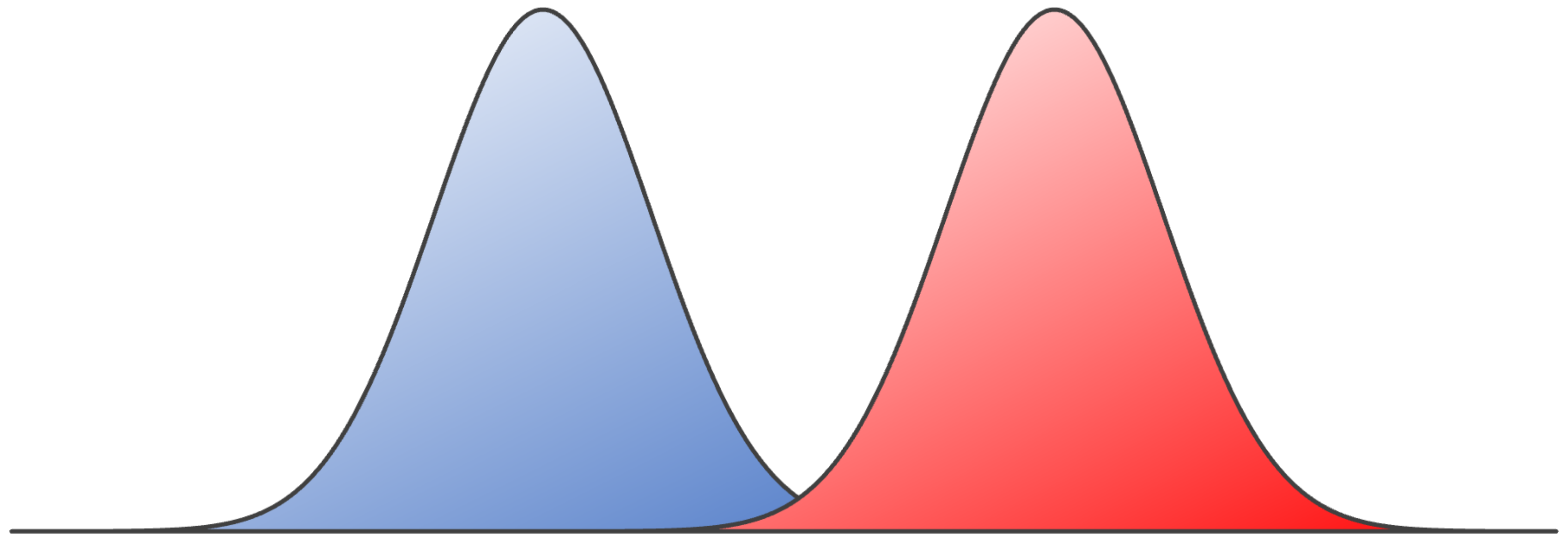
Larry Wasserman
(CMU)

Aarti Singh
(CMU)

We propose a flexible framework for two-sample testing

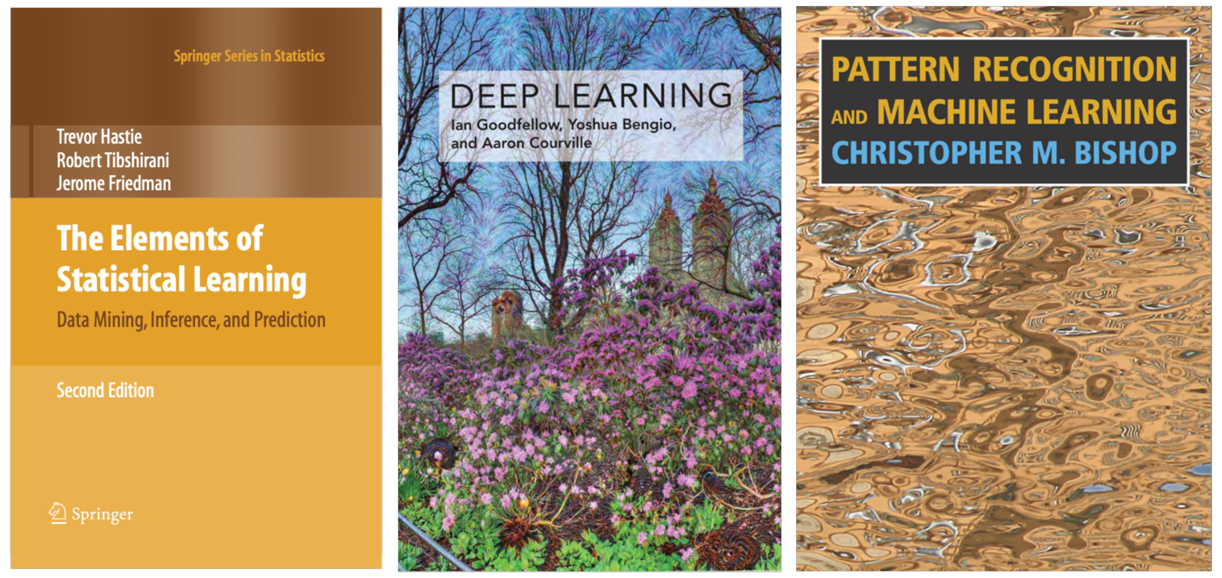


Classification/Regression

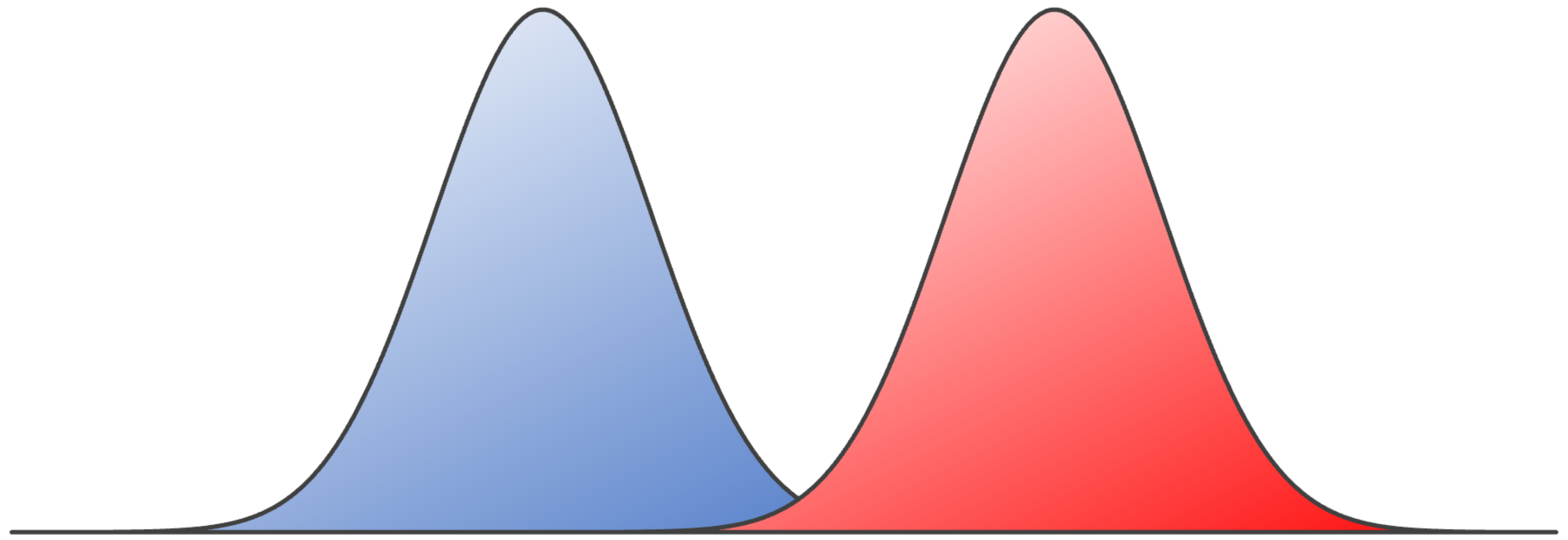


Two-sample test

We propose a flexible framework for two-sample testing



Classification/Regression

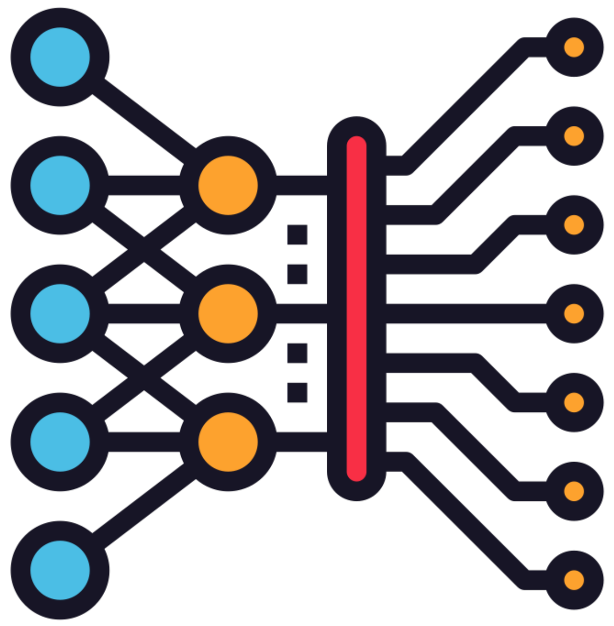


Two-sample test

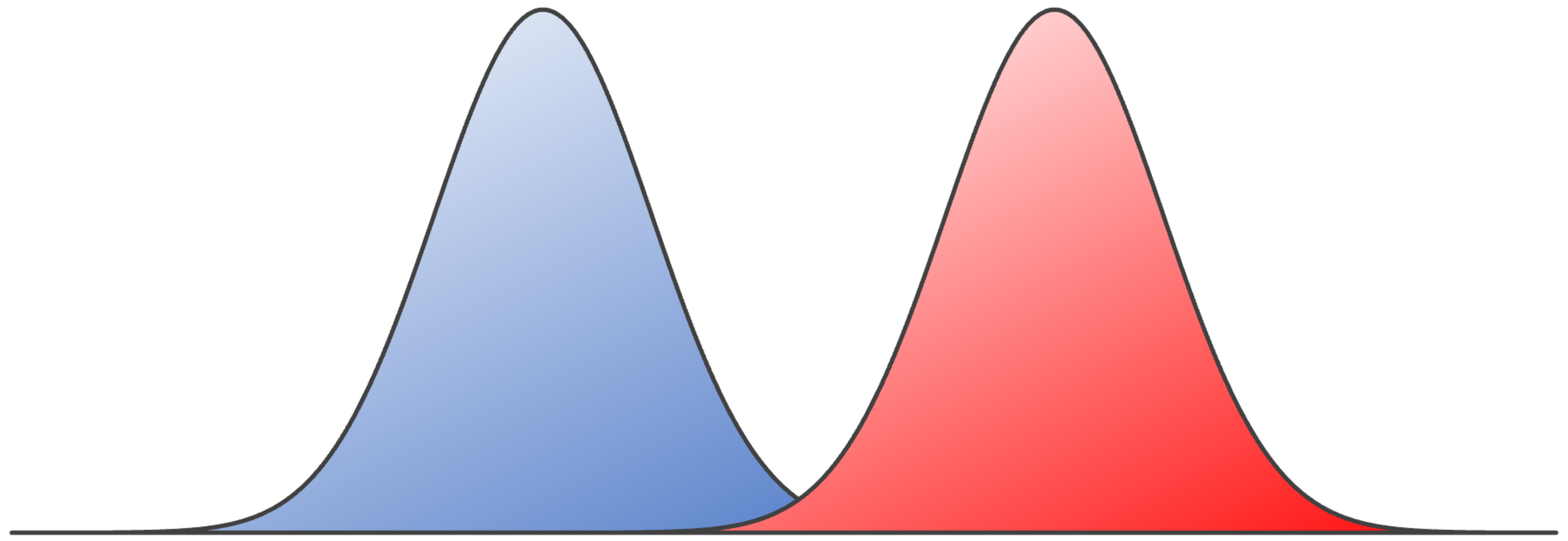
Idea

- Permutation tests present a **valid** p-value for **any** test statistic
- We can take advantage of **modern** algorithms in machine learning

We propose a flexible framework for two-sample testing



Neural networks

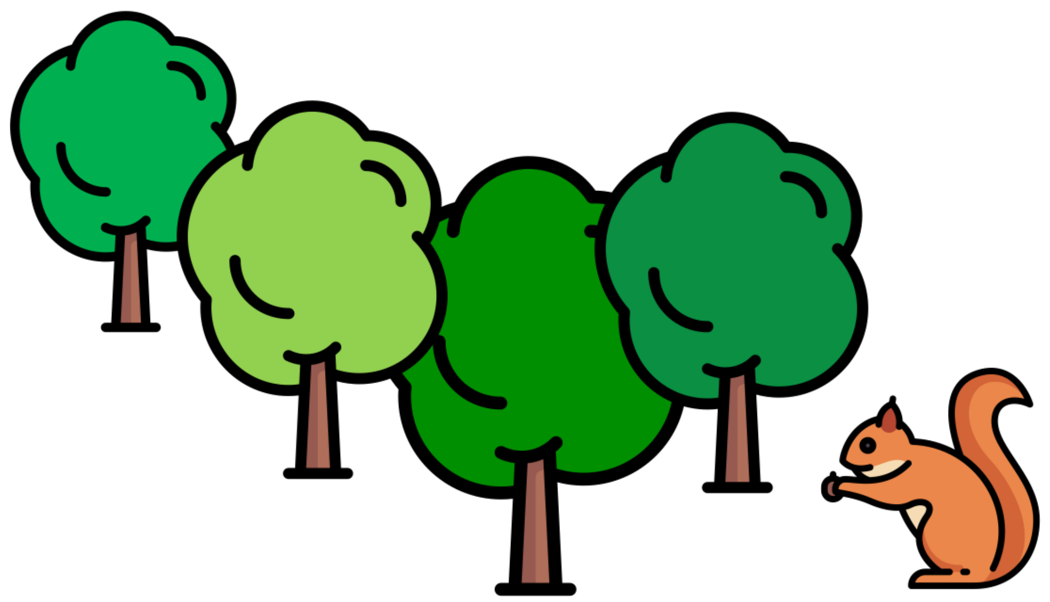


Two-sample test

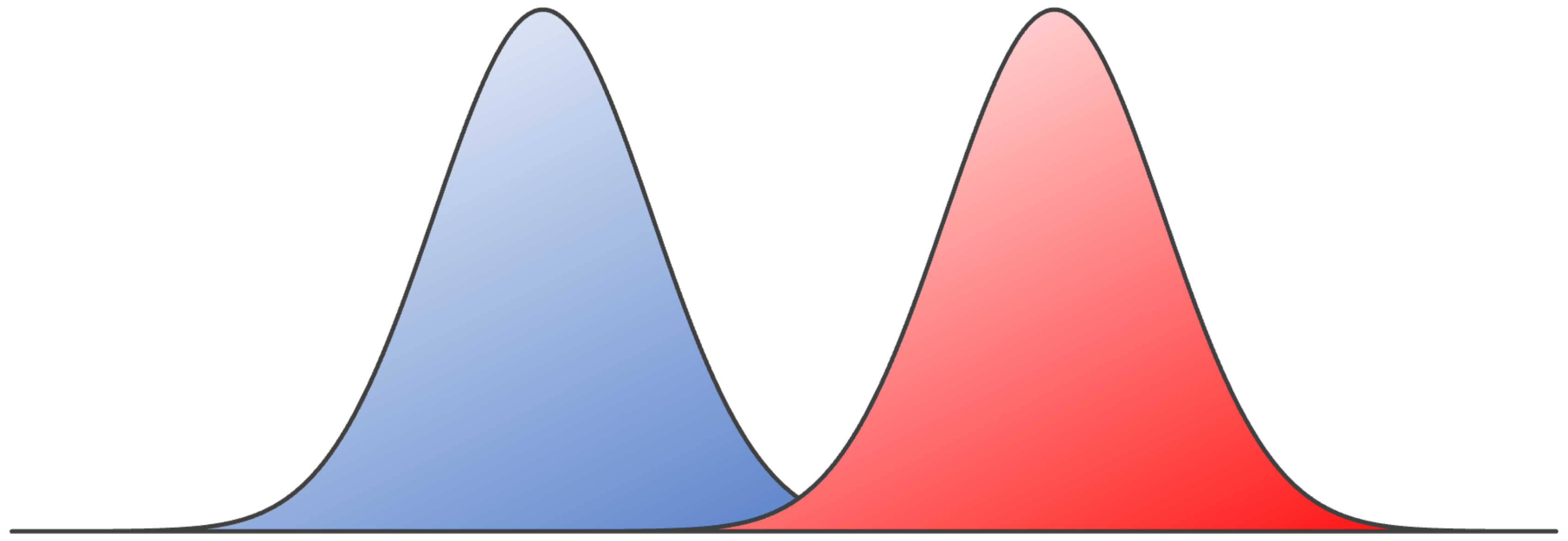
Idea

- Permutation tests present a **valid** p-value for **any** test statistic
- We can take advantage of **modern** algorithms in machine learning

We propose a flexible framework for two-sample testing



Random forests

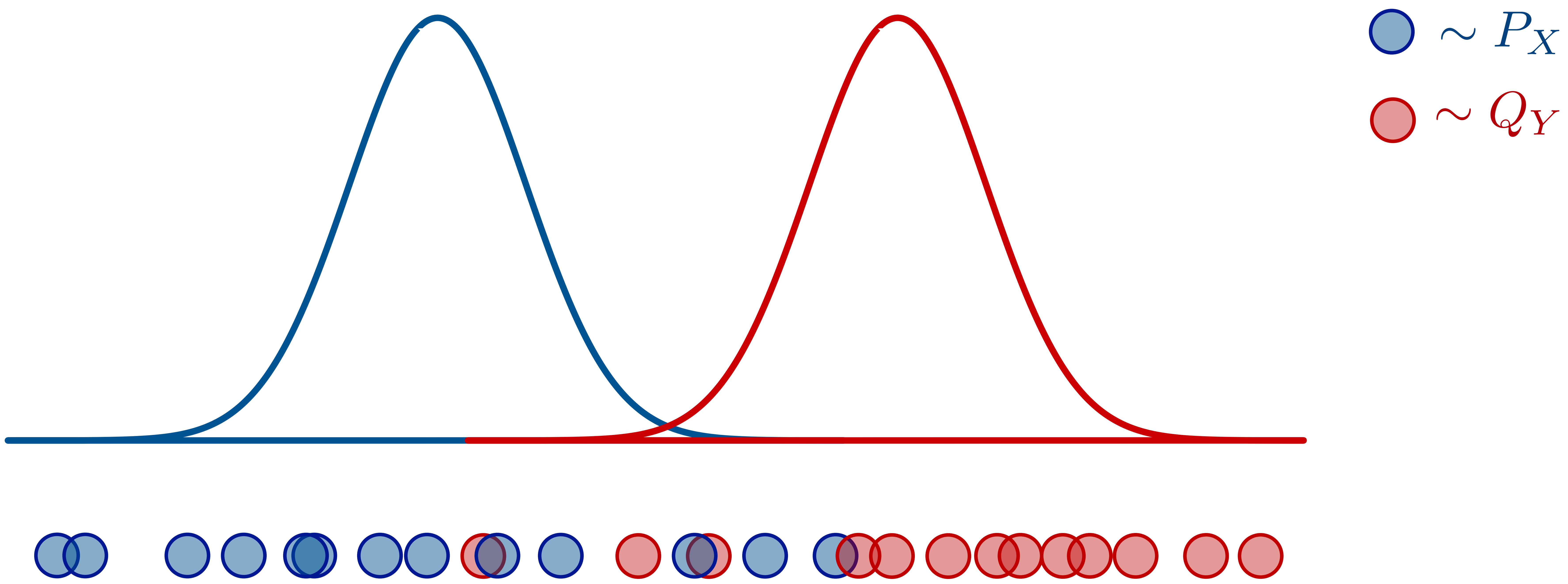


Two-sample test

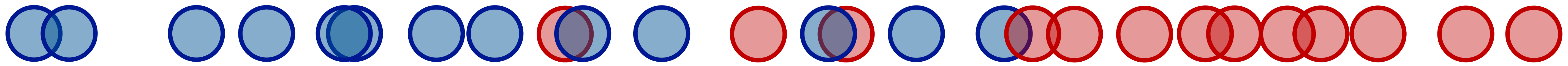
Idea

- Permutation tests present a **valid** p-value for **any** test statistic
- We can take advantage of **modern** algorithms in machine learning

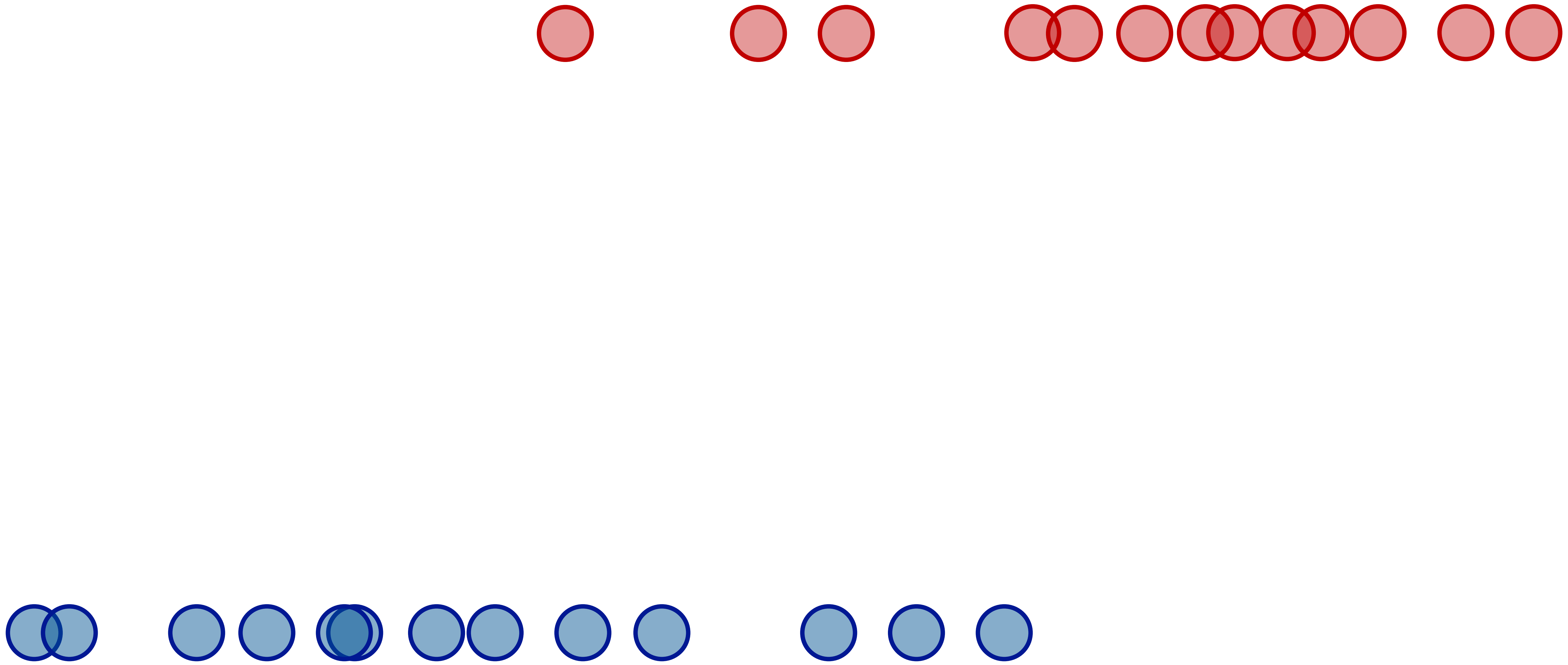
Regression-based two-sample test



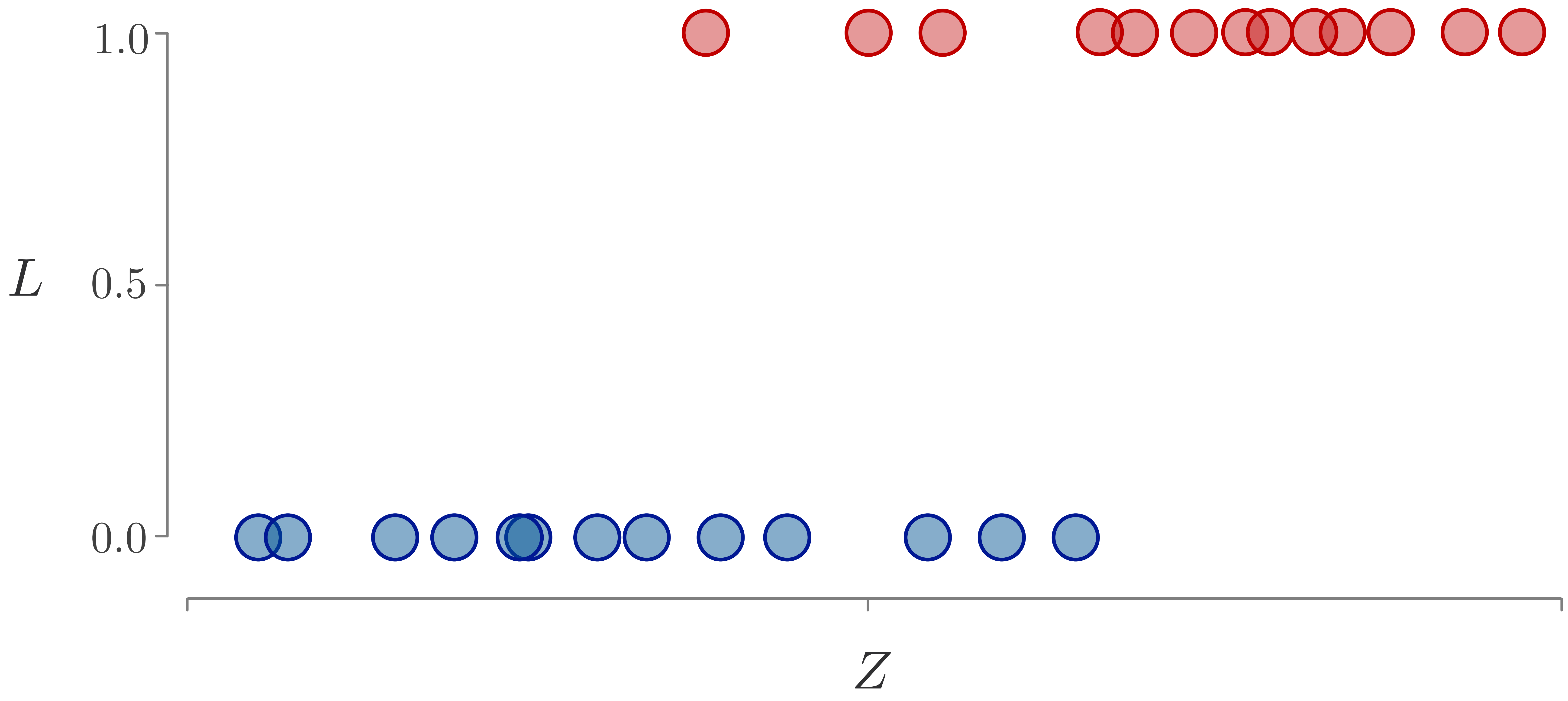
Regression-based two-sample test



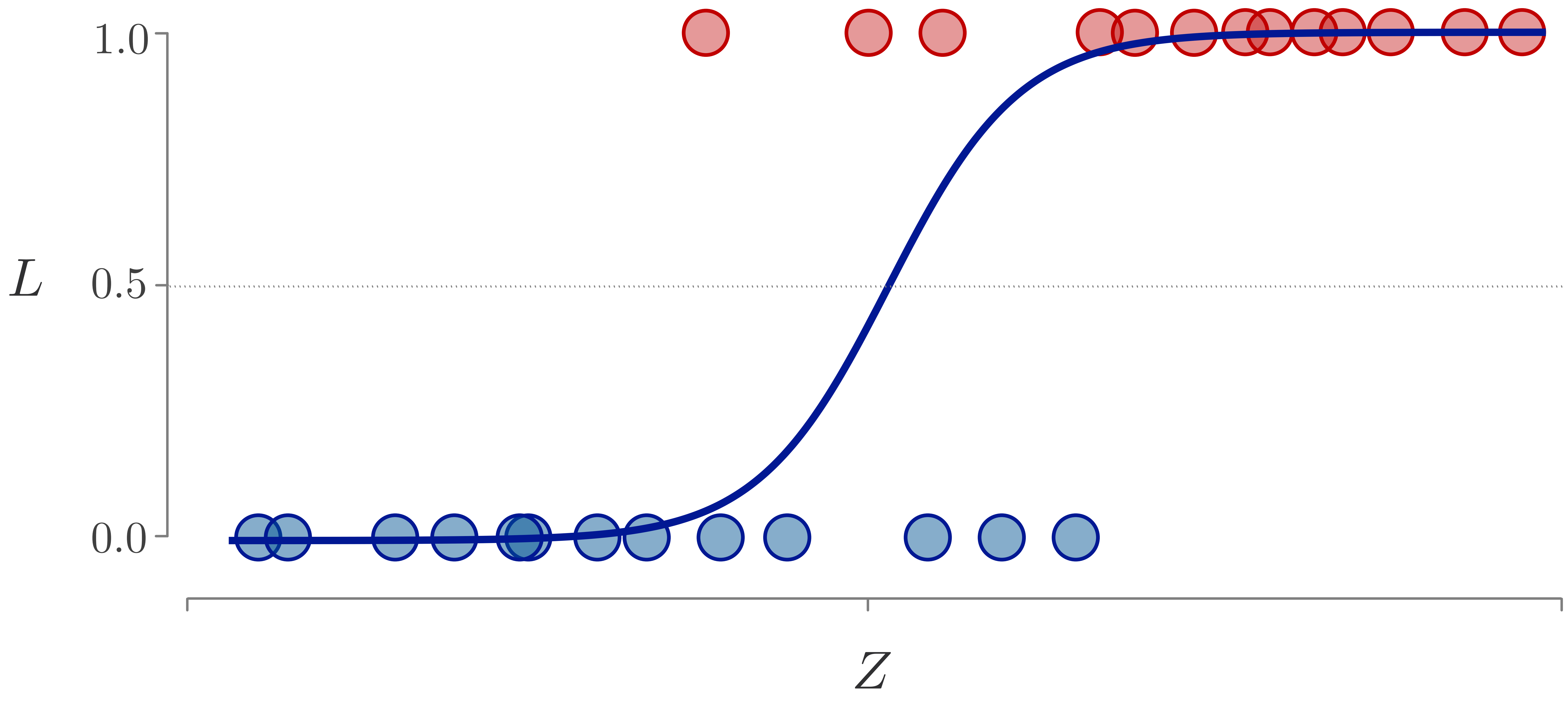
Regression-based two-sample test



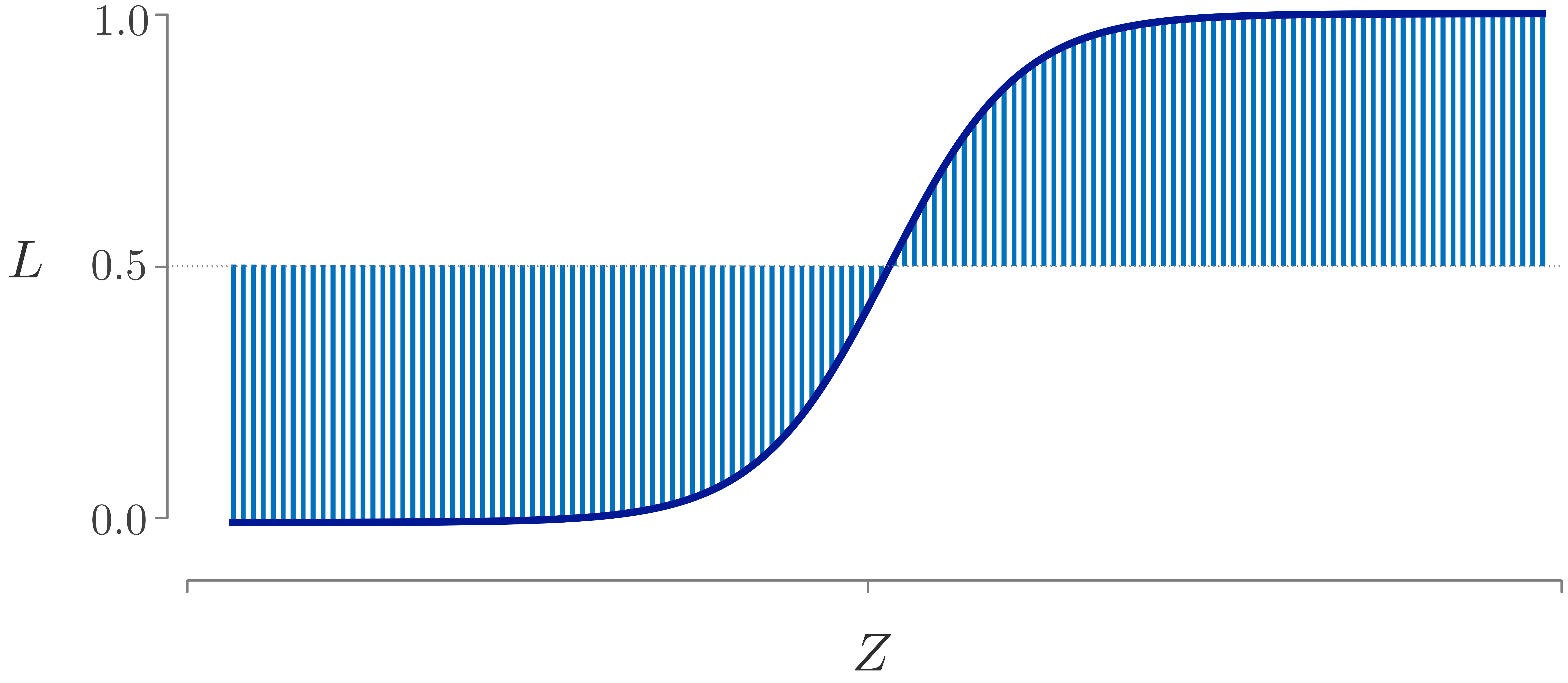
Regression-based two-sample test



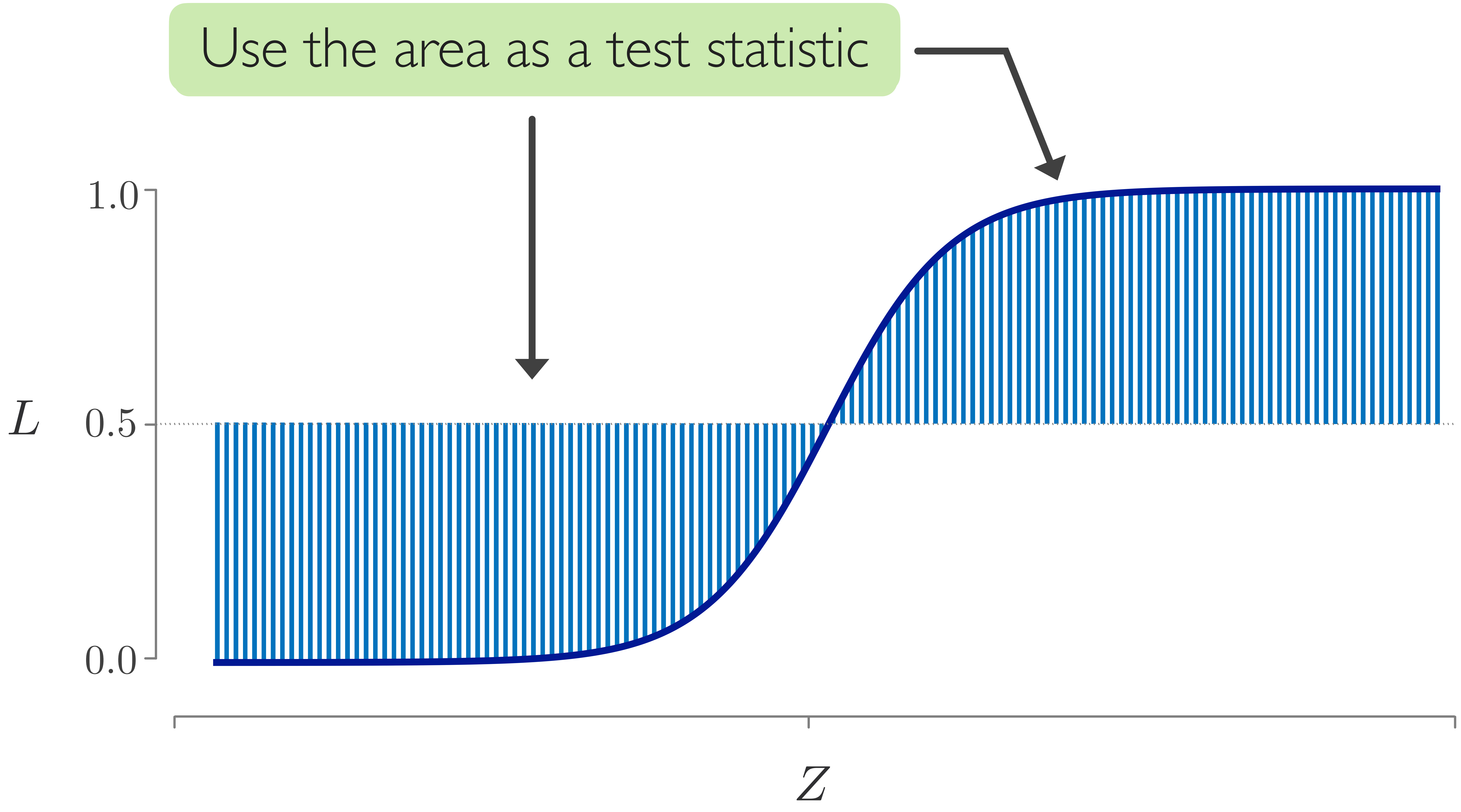
Regression-based two-sample test



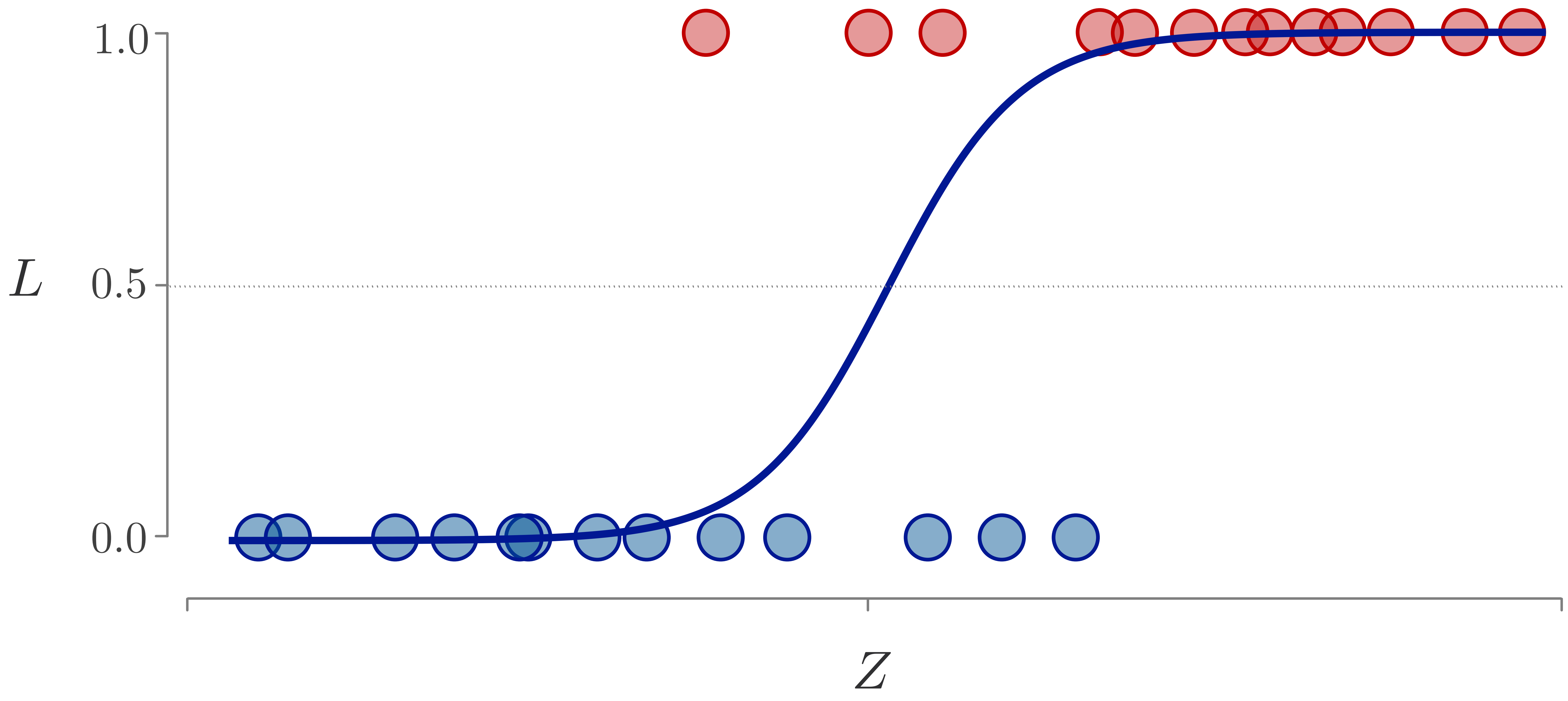
Regression-based two-sample test



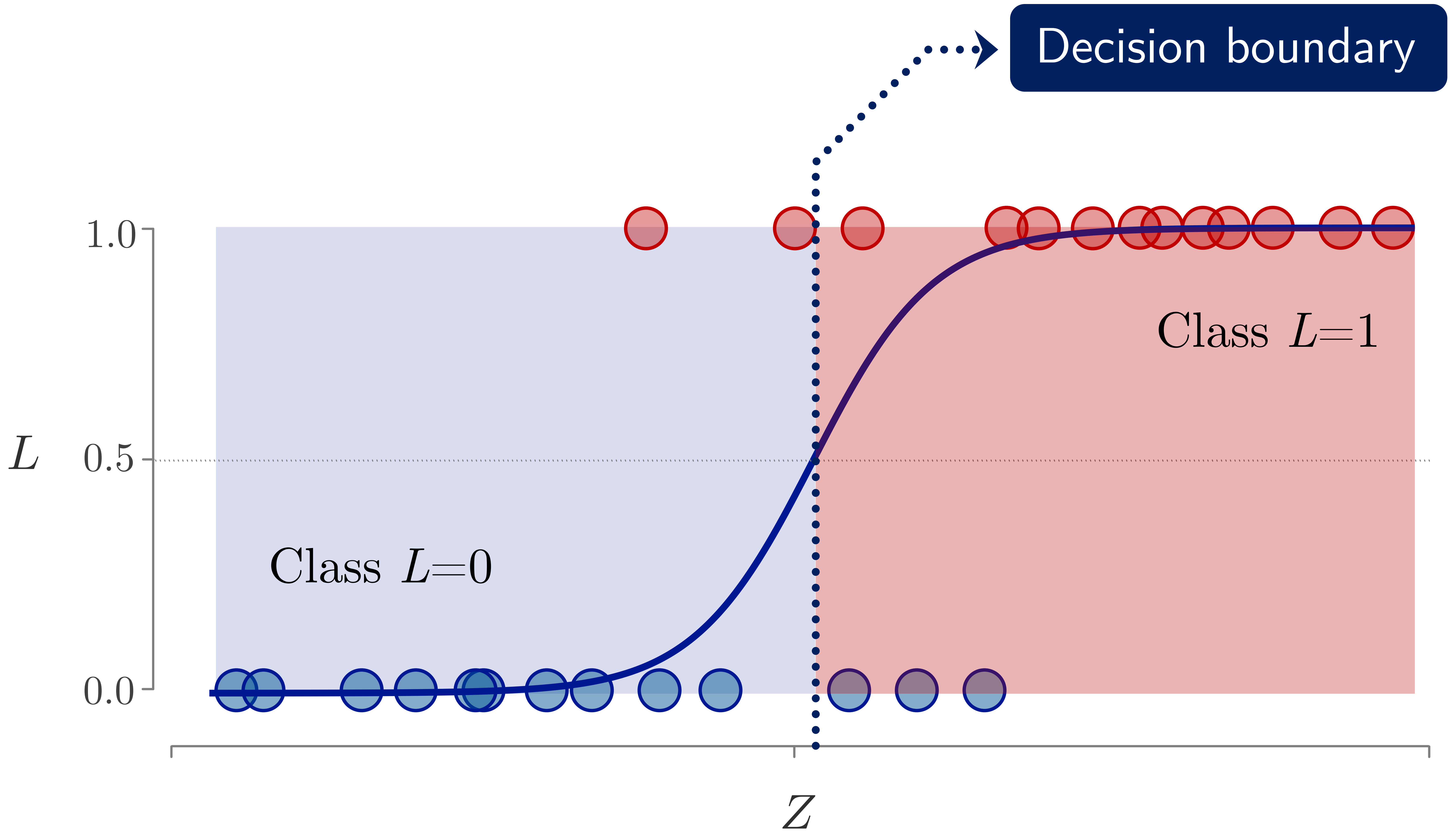
Regression-based two-sample test



Classification accuracy-based two-sample test



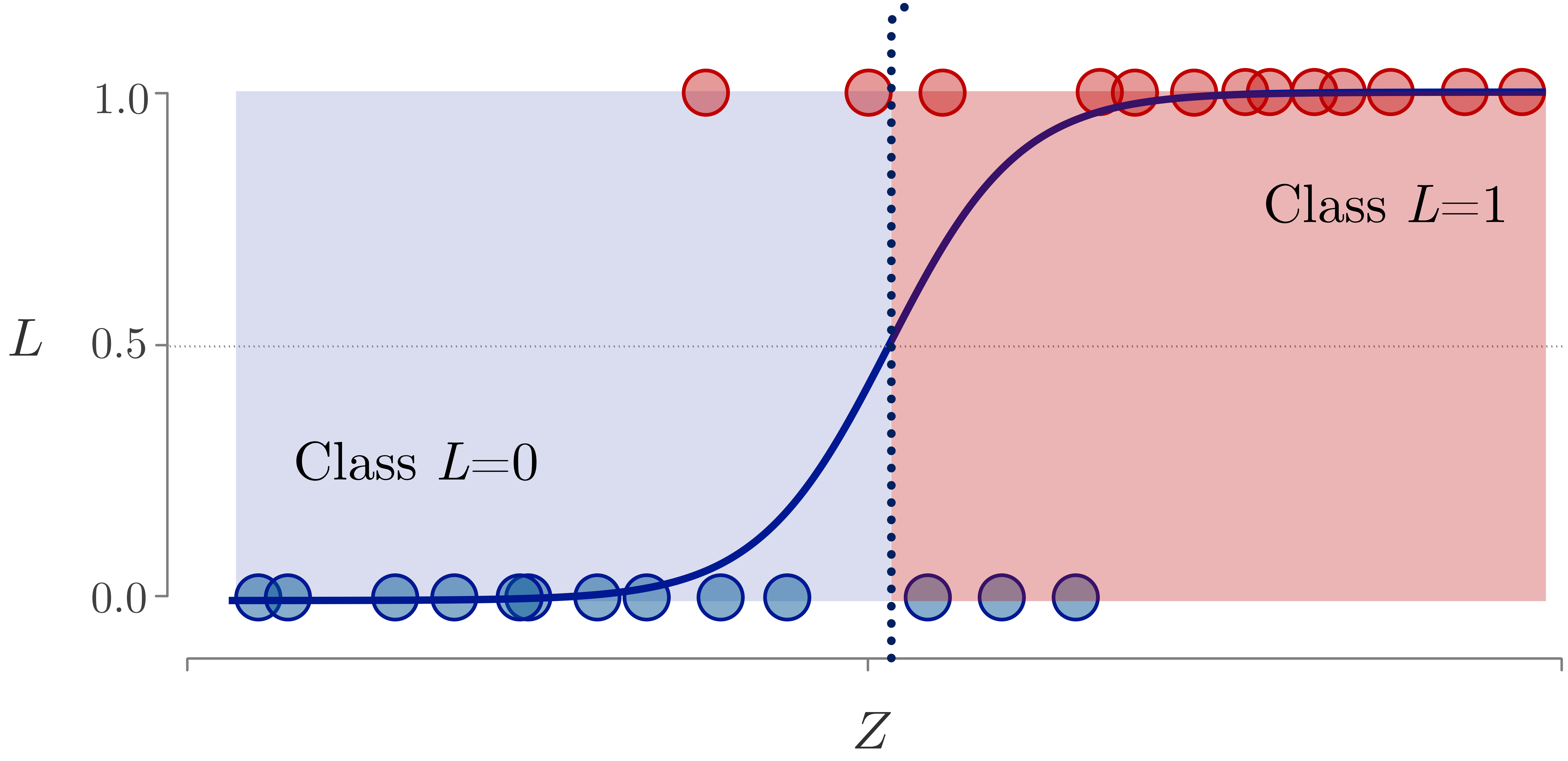
Classification accuracy-based two-sample test



Classification accuracy-based two-sample test

Use the accuracy as a test statistic

Decision boundary



Test statistics

$$
\mathcal{D}_N := \begin{array}{c} \begin{matrix} Z & L \end{matrix} \\ \begin{bmatrix} X_1 & 0 \\ \vdots & \vdots \\ X_n & 0 \\ Y_1 & 1 \\ \vdots & \vdots \\ Y_m & 1 \end{bmatrix} \end{array}
$$

$m = n$
for simplicity

Test statistics

$$\mathcal{D}_N := \begin{array}{cc} Z & L \\ \left[\begin{array}{cc} X_1 & 0 \\ \vdots & \vdots \\ X_n & 0 \\ Y_1 & 1 \\ \vdots & \vdots \\ Y_m & 1 \end{array}\right] \end{array}$$

$m = n$
for simplicity

Train a regressor

$$\widehat{R} : \mathcal{D}_N \mapsto [0,1]$$

Test statistics

$$\mathcal{D}_N := \begin{array}{c} \begin{array}{cc} Z & L \end{array} \\ \begin{bmatrix} X_1 & 0 \\ \vdots & \vdots \\ X_n & 0 \\ Y_1 & 1 \\ \vdots & \vdots \\ Y_m & 1 \end{bmatrix} \end{array}$$

$m = n$
for simplicity

Train a regressor

$$\widehat{R} : \mathcal{D}_N \mapsto [0, 1]$$

Mean square error

$$N^{-1} \sum_{i \in [N]} \left(\widehat{R}(Z_i) - 1/2\right)^2$$

Test statistics

$$\mathcal{D}_N := \begin{array}{cc} Z & L \\ \left[\begin{array}{cc} X_1 & 0 \\ \vdots & \vdots \\ X_n & 0 \\ Y_1 & 1 \\ \vdots & \vdots \\ Y_m & 1 \end{array}\right] & \end{array}$$

$m = n$
for simplicity

Train a regressor

$$\widehat{R} : \mathcal{D}_N \mapsto [0, 1]$$

Mean square error

$$N^{-1} \sum_{i \in [N]} \left(\widehat{R}(Z_i) - 1/2\right)^2$$

Train a classifier

$$\widehat{C} : \mathcal{D}_N \mapsto \{0, 1\}$$

Test statistics

$$\mathcal{D}_N := \begin{array}{c} \begin{array}{cc} Z & L \end{array} \\ \begin{bmatrix} X_1 & 0 \\ \vdots & \vdots \\ X_n & 0 \\ Y_1 & 1 \\ \vdots & \vdots \\ Y_m & 1 \end{bmatrix} \end{array}$$

$m = n$
for simplicity

Train a regressor → **Mean square error**

$$\widehat{R} : \mathcal{D}_N \mapsto [0, 1]$$

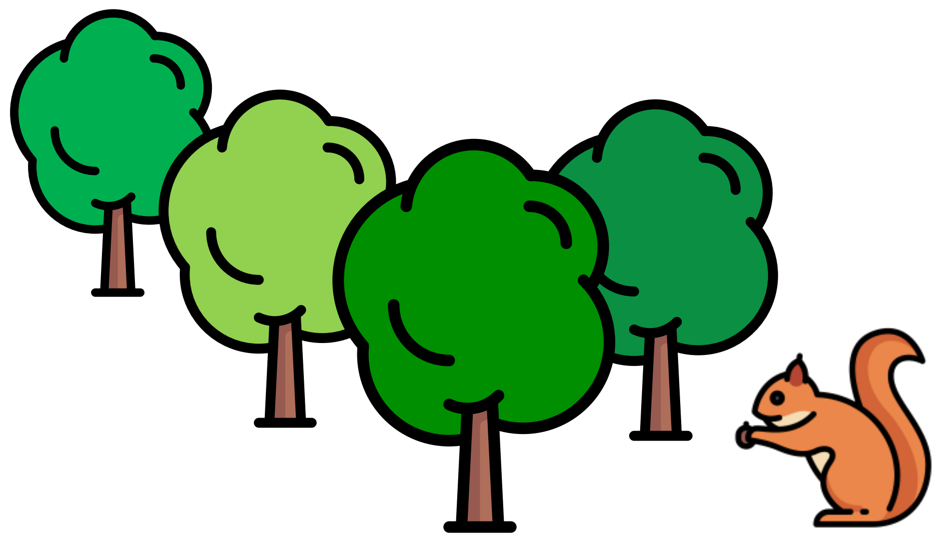
$$N^{-1} \sum_{i \in [N]} \left(\widehat{R}(Z_i) - 1/2\right)^2$$

Train a classifier → **Accuracy**

$$\widehat{C} : \mathcal{D}_N \mapsto \{0, 1\}$$

$$N^{-1} \sum_{i \in [N]} I\left(\widehat{C}(Z) = 1\right)$$

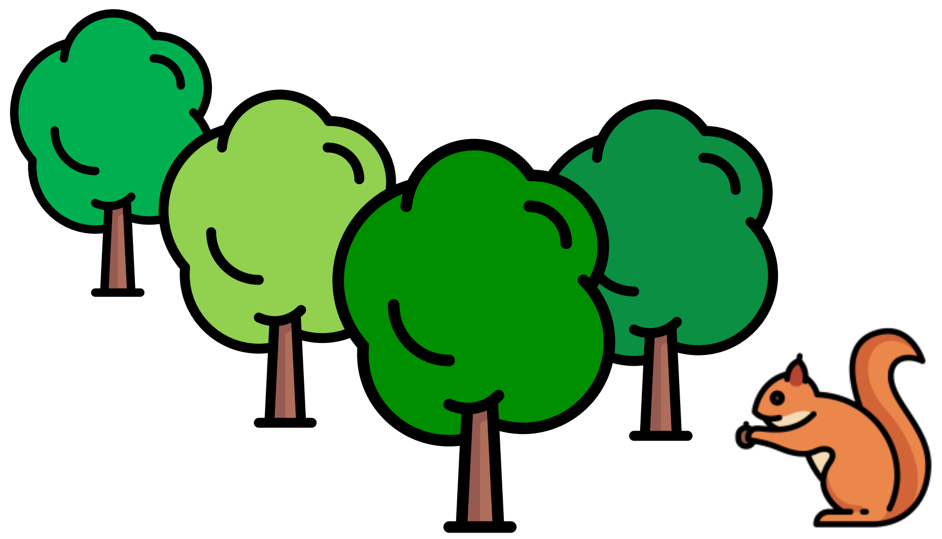
Empirical results



Random forests

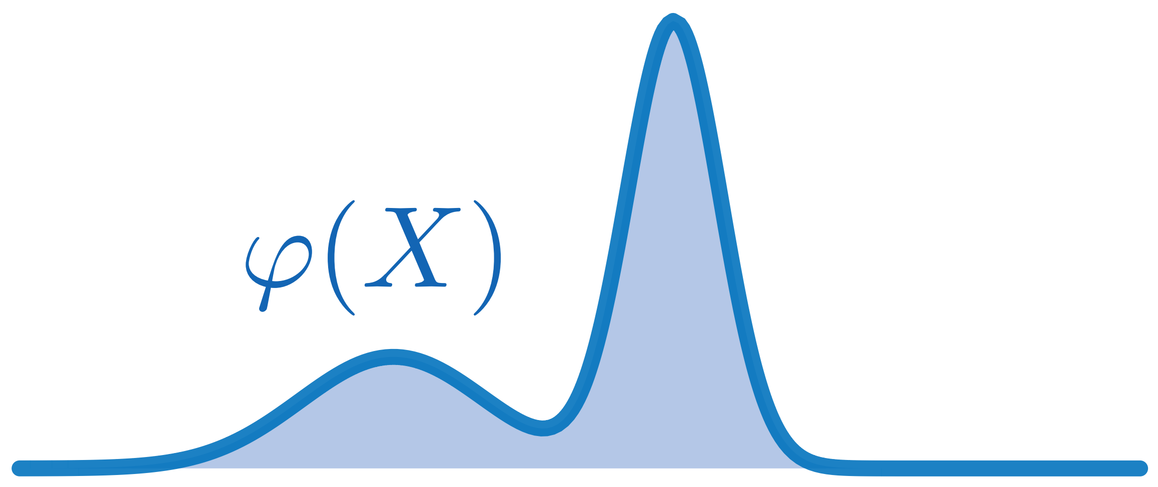
- Screen irrelevant variables
- Robust to outliers
- Handle various data types
- Empirical success

Empirical results



Random forests

versus

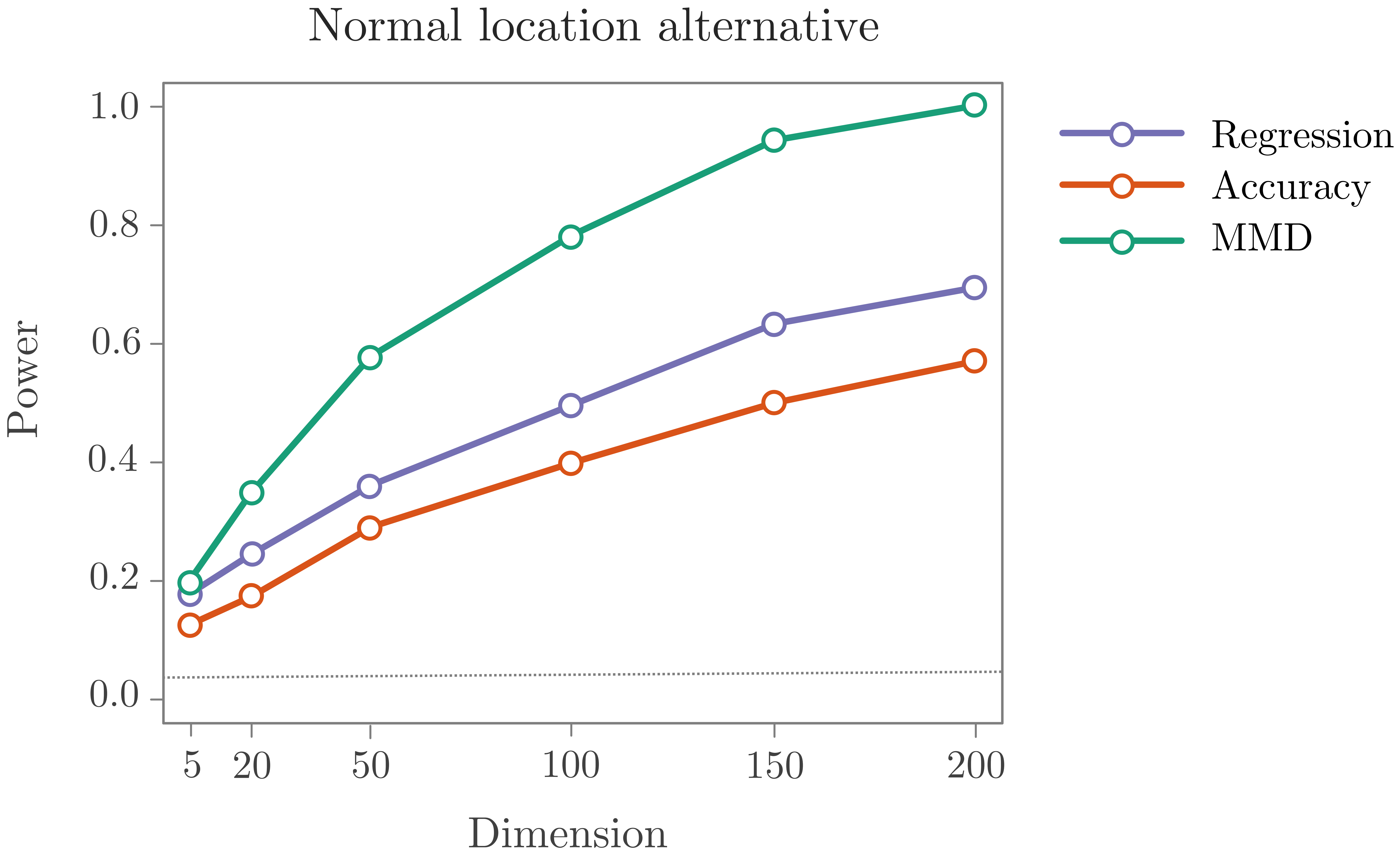


Maximum Mean Discrepancy

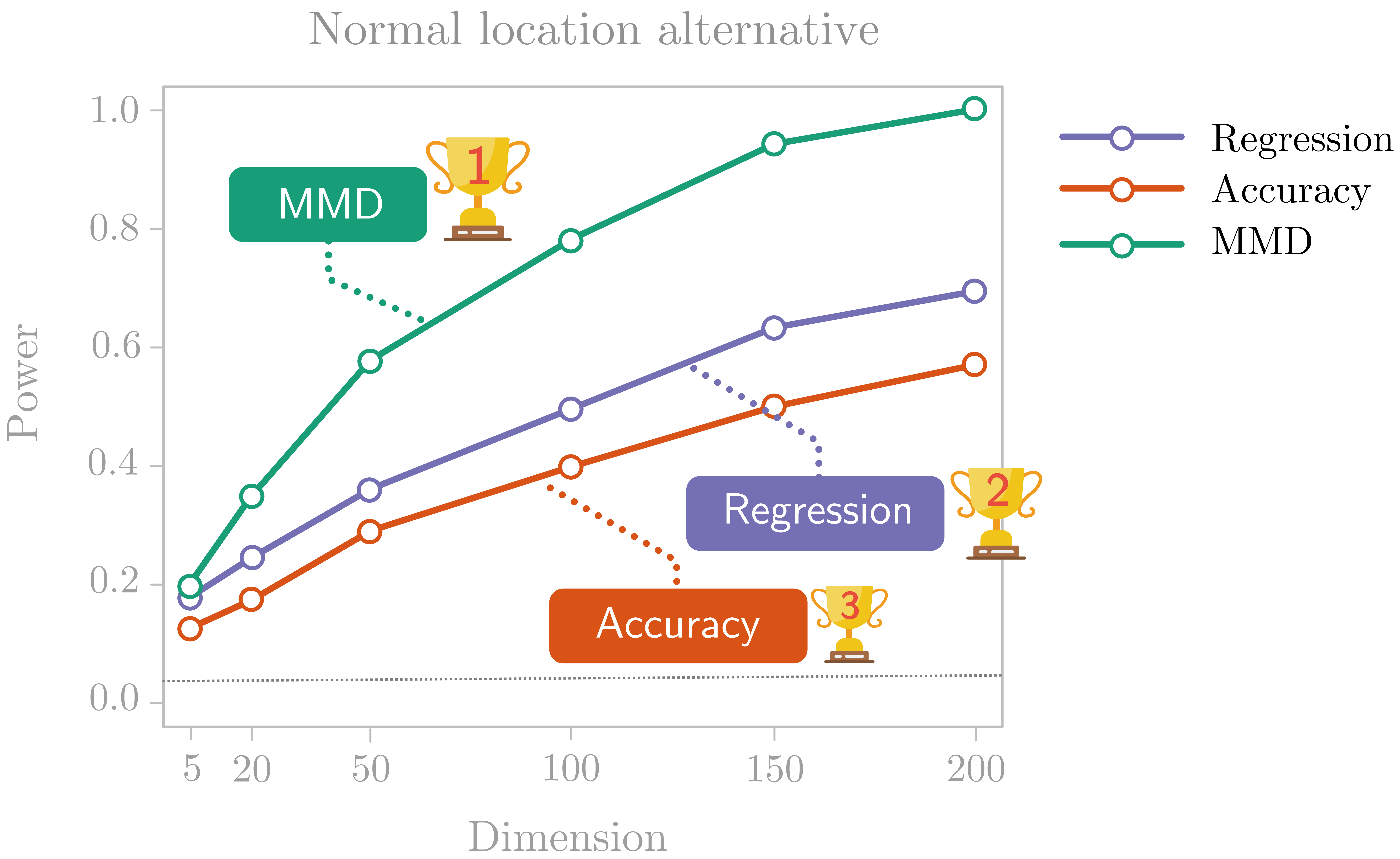
- Screen irrelevant variables
- Robust to outliers
- Handle various data types
- Empirical success

- Gretton et al. (2012)
- Popular method in ML
- Simple calculation formula
- Theoretical support

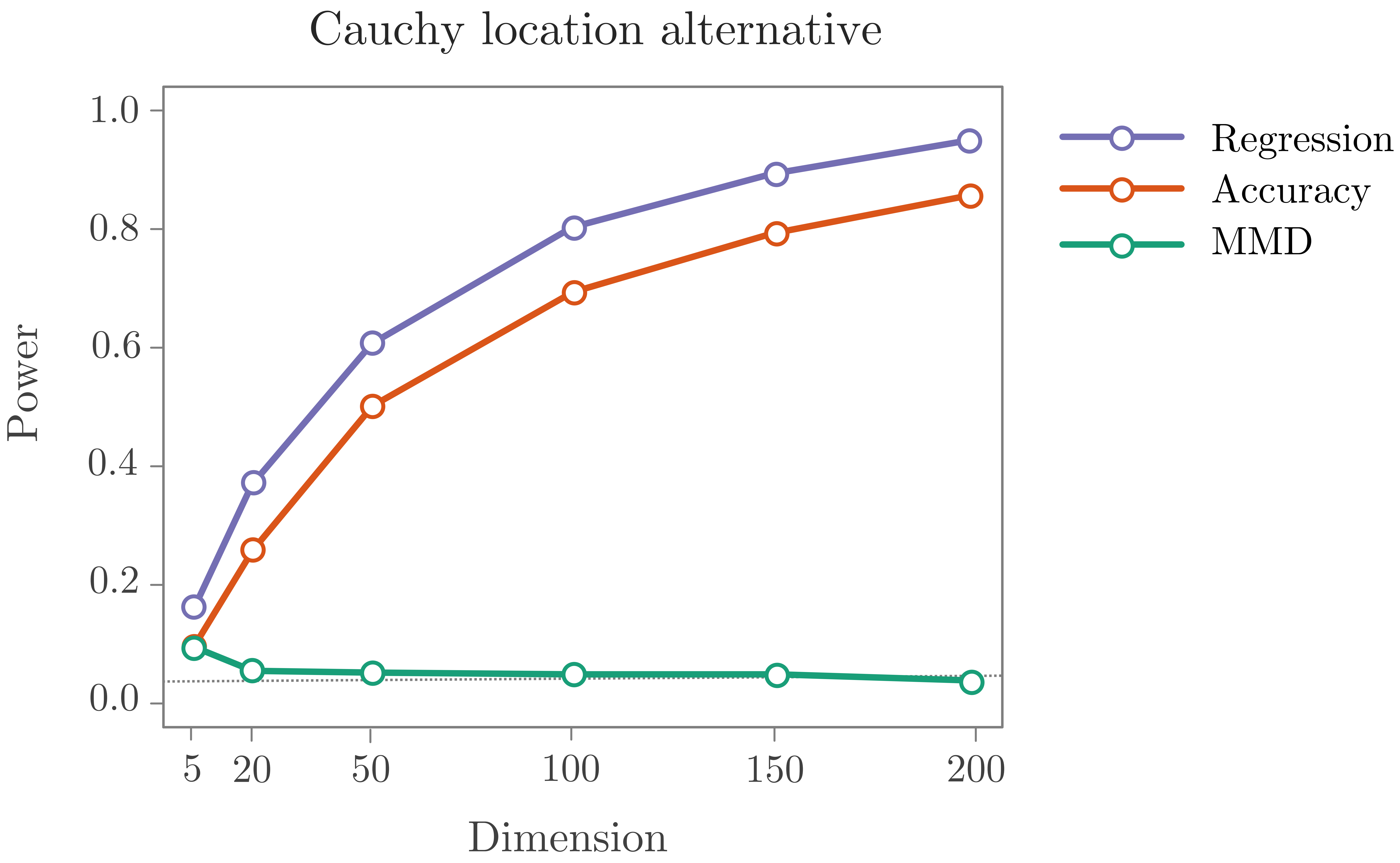
Power comparison



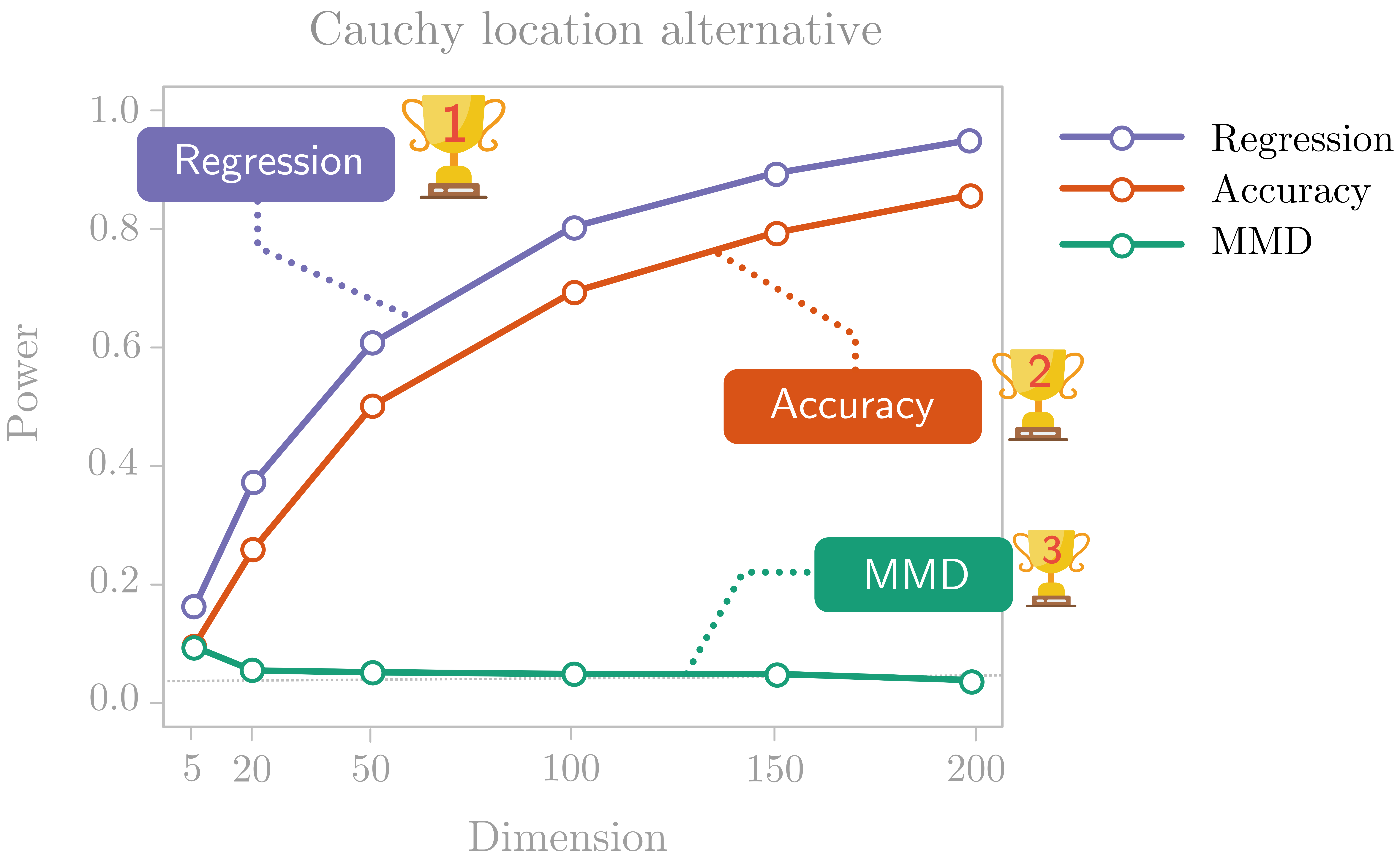
Power comparison



Power comparison



Power comparison



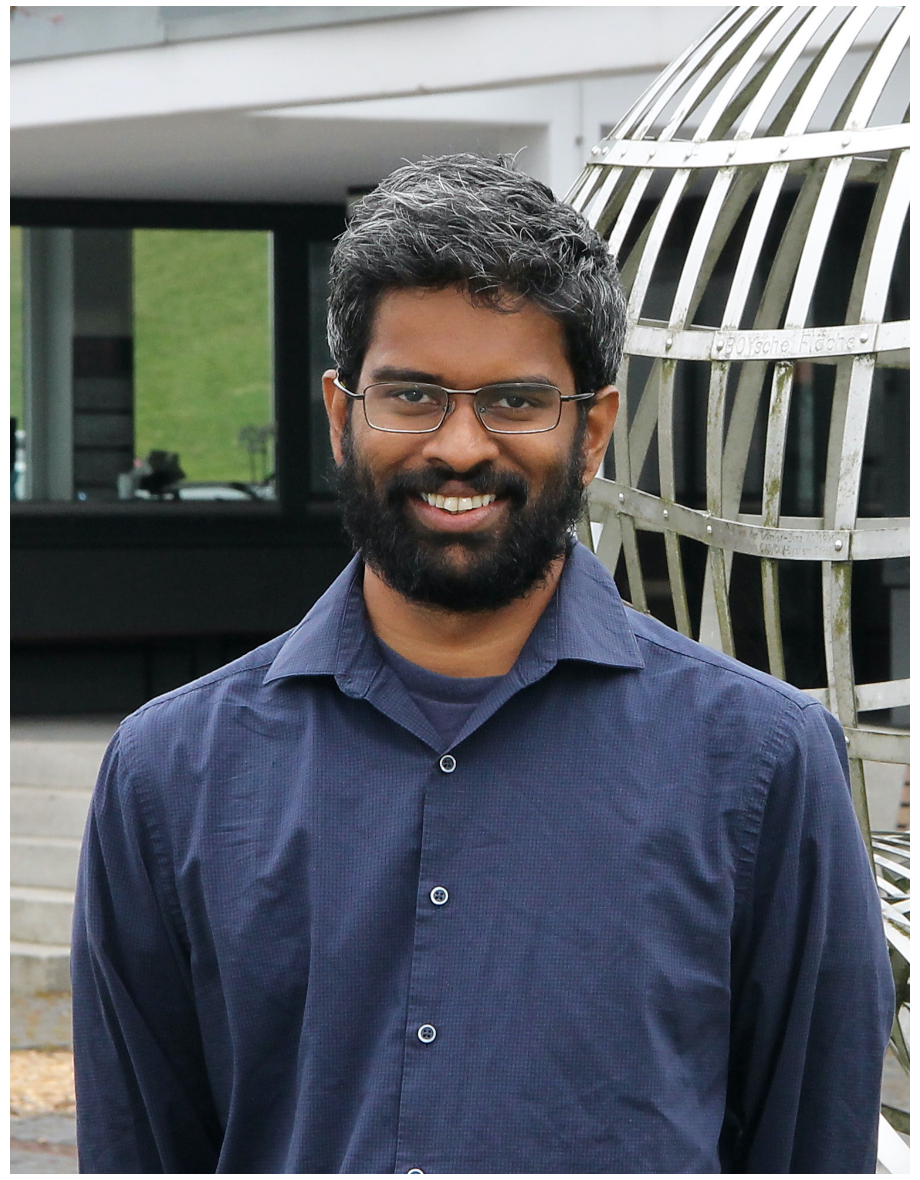
Part 1. Methodological Contributions

Part II. Theoretical Contributions

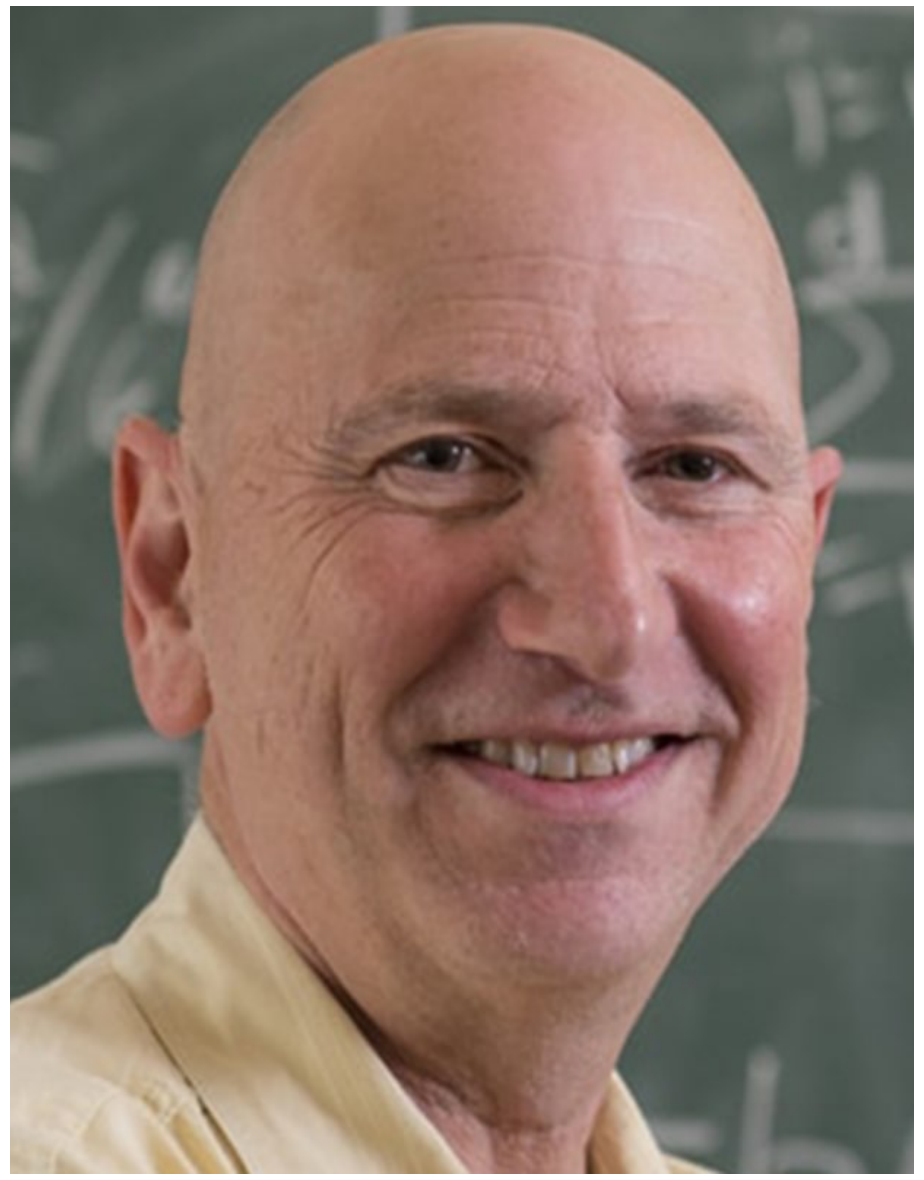
The Annals of Statistics
2022, Vol. 50, No. 1, 225–251
https://doi.org/10.1214/21-AOS2103

MINIMAX OPTIMALITY OF PERMUTATION TESTS

BY ILMUN KIM[1], SIVARAMAN BALAKRISHNAN[2,*] AND LARRY WASSERMAN[2,†]



Siva Balakrishnan
(CMU)



Larry Wasserman
(CMU)

Challenge: Random critical value

Remark the permutation test **rejects** H_0 when

$$\underbrace{T}_{\text{Test statistic}} > \underbrace{q_{1-\alpha}}_{\text{Critical value}} := \text{Quantile}_{1-\alpha}(T, T_1^{\pi}, \ldots, T_B^{\pi})$$

Challenge: Random critical value

Remark the permutation test **rejects** H_0 when

$$\underbrace{T}_{\text{Test statistic}} > \underbrace{q_{1-\alpha}}_{\text{Critical value}} := \mathrm{Quantile}_{1-\alpha}(T, T_1^\pi, \ldots, T_B^\pi)$$

Our **goal** is to identify **non-asymptotic conditions** under which

$$\sup_{P \in \mathscr{P}_1} \mathbb{P}_P(T \leq q_{1-\alpha}) \leq \beta \quad (\text{Say, } \beta = 0.05)$$

Uniform Type II error over $\mathscr{P}_1$

Challenge: Random critical value

Random quantity +
Lacking independence

Remark the permutation test **rejects** H_0 when

$$\boxed{T} > \boxed{q_{1-\alpha}} := \mathrm{Quantile}_{1-\alpha}(T, T_1^{\pi}, \ldots, T_B^{\pi})$$

Test statistic Critical value

Our **goal** is to identify **non-asymptotic conditions** under which

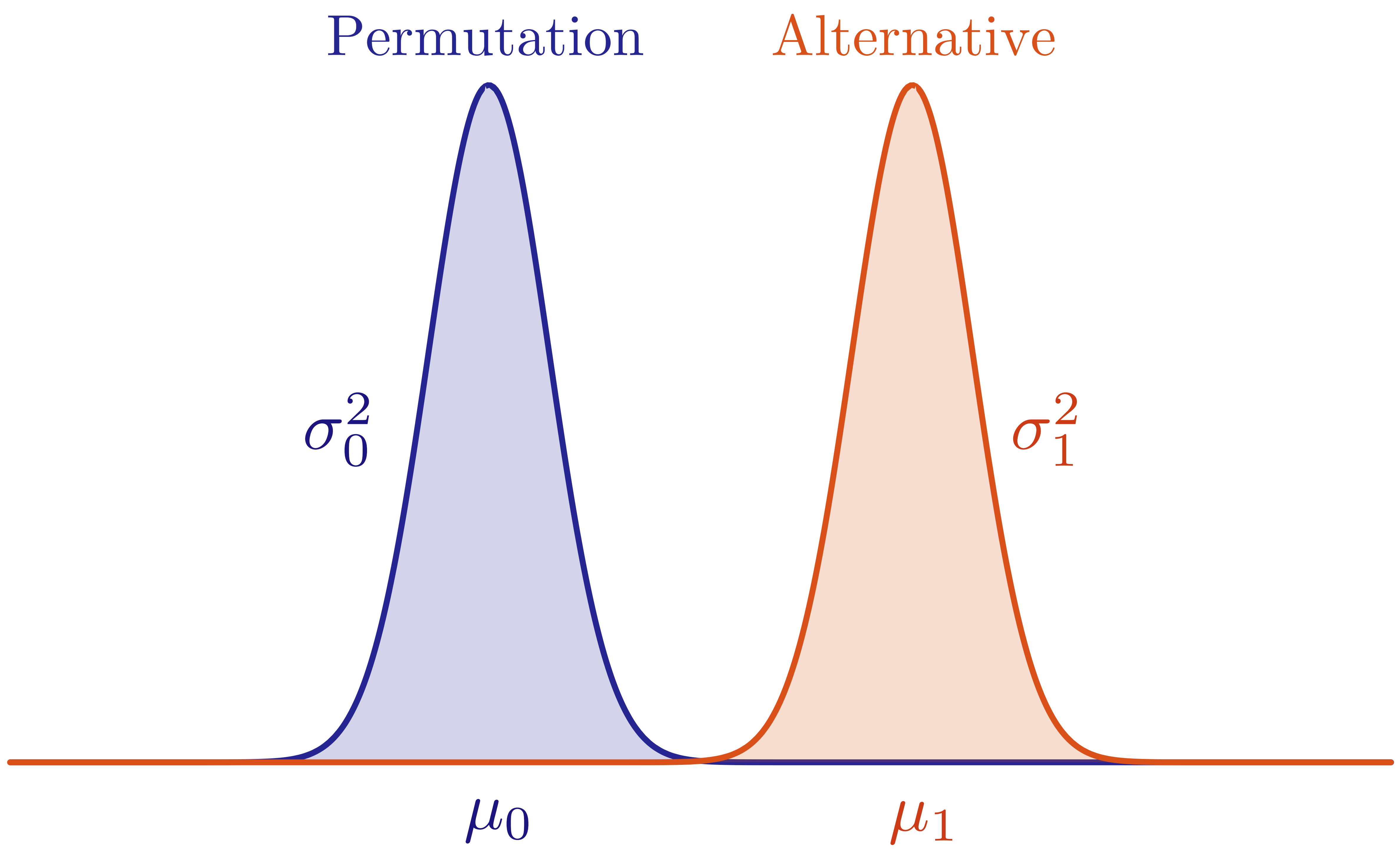
$$\sup_{P \in \mathscr{P}_1} \mathbb{P}_P(T \le q_{1-\alpha}) \le \beta \quad \text{(Say, } \beta = 0.05\text{)}$$

Uniform Type II error over $\mathscr{P}_1$

Our proposal: Two moments method

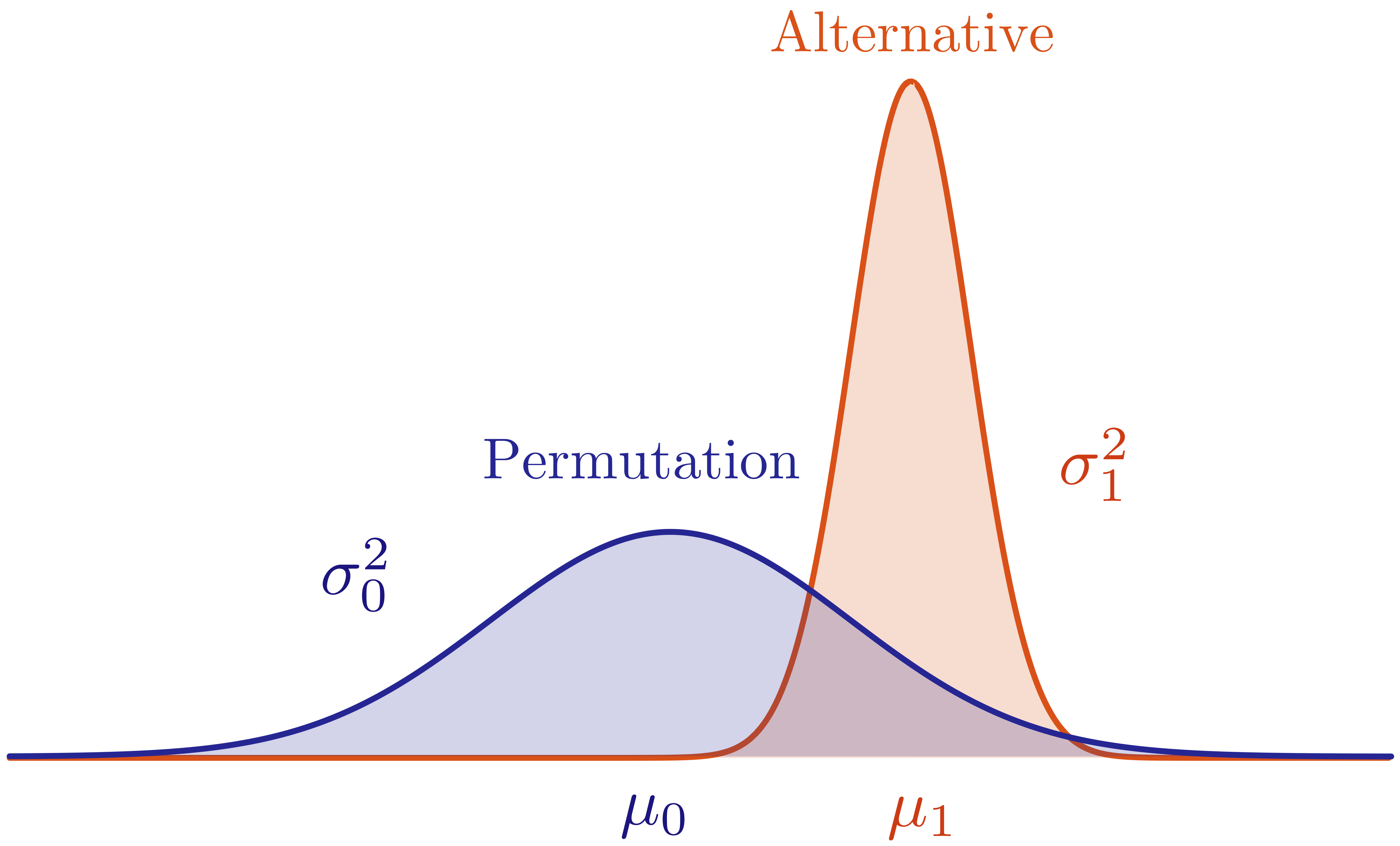
Intuition for the two moments method

- High power when signal $\mu_1 - \mu_0 \gg$ noise $\max\{\sigma_0, \sigma_1\}$



Intuition for the two moments method

- Suffer from **low power** when signal $\mu_1 - \mu_0 \ll$ noise $\max\{\sigma_0, \sigma_1\}$



Two moments method

Theorem [KBW 2022]

- The number of (random) permutations $B \gtrsim \alpha^{-2} \log(1/\beta)$

Two moments method

Theorem [KBW 2022]

Independent of the sample size!

- The number of (random) permutations $B \gtrsim \alpha^{-2}\log(1/\beta)$

Two moments method

Theorem [KBW 2022]

Independent of the sample size!

- The number of (random) permutations $B \gtrsim \alpha^{-2}\log(1/\beta)$
- For any $P \in \mathscr{P}_1$,

$$\mathbb{E}_P[T] - \mathbb{E}_{P,\pi}[T^\pi] \gtrsim \sqrt{\beta^{-1}\mathbb{V}_P[T]} + \sqrt{\alpha^{-1}\beta^{-1}\mathbb{V}_{P,\pi}[T^\pi]}$$

Two moments method

Theorem [KBW 2022]

Independent of the sample size!

- The number of (random) permutations $B \gtrsim \alpha^{-2} \log(1/\beta)$
- For any $P \in \mathscr{P}_1$,

$$\mathbb{E}_P[T] - \mathbb{E}_{P,\pi}[T^\pi] \gtrsim \sqrt{\beta^{-1}\mathbb{V}_P[T]} + \sqrt{\alpha^{-1}\beta^{-1}\mathbb{V}_{P,\pi}[T^\pi]}$$

Signal $\gtrsim$ Noise I + Noise II

(data) (data + permutation)

Two moments method

Theorem [KBW 2022]

Independent of the sample size!

- The number of (random) permutations $B \gtrsim \alpha^{-2} \log(1/\beta)$
- For any $P \in \mathscr{P}_1$,

$$\mathbb{E}_P[T] - \mathbb{E}_{P,\pi}[T^\pi] \gtrsim \sqrt{\beta^{-1}\mathbb{V}_P[T]} + \sqrt{\alpha^{-1}\beta^{-1}\mathbb{V}_{P,\pi}[T^\pi]}$$

Signal $\gtrsim$ Noise I + Noise II

(data) (data + permutation)

Uniform Type II error: $\sup_{P \in \mathscr{P}_1} \mathbb{P}_P(T \leq q_{1-\alpha}) \leq \beta$

Two moments method

[Punchline] Understanding the power of the permutation test boils down to understanding the **first two moments** of T and T^{π}.

Two moments method

[Punchline] Understanding the power of the permutation test boils down to understanding the **first two moments** of T and T^{π}.

[Improvement] If we have more information about T, we can further **improve/simplify** the conditions. We will illustrate this using U-statistics.

U-statistic for two-sample testing

- Suppose that we observe $\{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} P_Y$

U-statistic for two-sample testing

- Suppose that we observe $\{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} P_Y$
- Given a **kernel** $h(x, y)$,

$$U_n = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j) + \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(Y_i, Y_j)$$

$$-\frac{2}{n^2} \sum_{1 \le i \le n} \sum_{1 \le j \le n} h(X_i, Y_j)$$

U-statistic for two-sample testing

- Suppose that we observe $\{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} P_Y$
- Given a **kernel** $h(x, y)$,

$$U_n = \underbrace{\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h(X_i, X_j)}^{\text{Within Similarity}} + \underbrace{\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h(Y_i, Y_j)}^{\text{Within Similarity}} \\ - \underbrace{\frac{2}{n^2} \sum_{1 \leq i \leq n} \sum_{1 \leq j \leq n} h(X_i, Y_j)}_{\text{Between Similarity}}$$

U-statistic for two-sample testing

- Suppose that we observe $\{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} P_Y$
- Given a **kernel** $h(x, y)$,

$$U_n = \underbrace{\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j)}_{\text{Within Similarity}} + \underbrace{\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(Y_i, Y_j)}_{\text{Within Similarity}}$$

$$\approx (P_X - P_Y)^2 \qquad - \underbrace{\frac{2}{n^2} \sum_{1 \le i \le n} \sum_{1 \le j \le n} h(X_i, Y_j)}_{\text{Between Similarity}}$$

U-statistic for two-sample testing

- Suppose that we observe $\{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P_X$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} P_Y$
- Given a **kernel** $h(x, y)$,

$$U_n = \underbrace{\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j)}_{\text{Within Similarity}} + \underbrace{\frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(Y_i, Y_j)}_{\text{Within Similarity}}$$

$$\approx (P_X - P_Y)^2 \qquad \underbrace{- \frac{2}{n^2} \sum_{1 \le i \le n} \sum_{1 \le j \le n} h(X_i, Y_j)}_{\text{Between Similarity}}$$

Examples

MMD (Gretton et al., 2012)
Energy (Szekely & Rizzo, 2013)

U-statistic for independence testing

- Suppose that we observe $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} P_{XY}$

U-statistic for independence testing

- Suppose that we observe $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} P_{XY}$
- Given **kernels** $h_X(x_1, x_2)$ for X and $h_Y(y_1, y_2)$ for Y,

$$U_n = \frac{(n-2)!}{n!} \sum_{(i,j)\in I_2^n} h_X(X_i, X_j) h_Y(Y_i, Y_j)$$

$$+ \frac{(n-4)!}{n!} \sum_{(i,j,q,r)\in I_4^n} h_X(X_i, X_j) h_Y(Y_q, Y_r) - 2\frac{(n-3)!}{n!} \sum_{(i,j,q)\in I_3^n} h_X(X_i, X_j) h_Y(Y_i, Y_q)$$

U-statistic for independence testing

- Suppose that we observe $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} P_{XY}$
- Given **kernels** $h_X(x_1, x_2)$ for X and $h_Y(y_1, y_2)$ for Y,

$$U_n = \underbrace{\frac{(n-2)!}{n!} \sum_{(i,j) \in I_2^n} h_X(X_i, X_j) h_Y(Y_i, Y_j)}_{P_{XY}^2}$$

$$+ \underbrace{\frac{(n-4)!}{n!} \sum_{(i,j,q,r) \in I_4^n} h_X(X_i, X_j) h_Y(Y_q, Y_r)}_{P_X^2 P_Y^2} - 2 \underbrace{\frac{(n-3)!}{n!} \sum_{(i,j,q) \in I_3^n} h_X(X_i, X_j) h_Y(Y_i, Y_q)}_{P_{XY} P_X P_Y}$$

U-statistic for independence testing

- Suppose that we observe $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} P_{XY}$
- Given **kernels** $h_X(x_1, x_2)$ for X and $h_Y(y_1, y_2)$ for Y,

$$P_{XY}^2$$

$$U_n = \frac{(n-2)!}{n!} \sum_{(i,j) \in I_2^n} h_X(X_i, X_j) h_Y(Y_i, Y_j) \approx (P_{XY} - P_X P_Y)^2$$

$$+ \frac{(n-4)!}{n!} \sum_{(i,j,q,r) \in I_4^n} h_X(X_i, X_j) h_Y(Y_q, Y_r) - 2 \frac{(n-3)!}{n!} \sum_{(i,j,q) \in I_3^n} h_X(X_i, X_j) h_Y(Y_i, Y_q)$$

$$P_X^2 P_Y^2 \qquad P_{XY} P_X P_Y$$

U-statistic for independence testing

- Suppose that we observe $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} P_{XY}$
- Given **kernels** $h_X(x_1, x_2)$ for X and $h_Y(y_1, y_2)$ for Y,

$$P_{XY}^2$$

$$U_n = \boxed{\frac{(n-2)!}{n!} \sum_{(i,j) \in I_2^n} h_X(X_i, X_j) h_Y(Y_i, Y_j)} \approx (P_{XY} - P_X P_Y)^2$$

$$+ \boxed{\frac{(n-4)!}{n!} \sum_{(i,j,q,r) \in I_4^n} h_X(X_i, X_j) h_Y(Y_q, Y_r)} - 2 \boxed{\frac{(n-3)!}{n!} \sum_{(i,j,q) \in I_3^n} h_X(X_i, X_j) h_Y(Y_i, Y_q)}$$

$$P_X^2 P_Y^2 \qquad\qquad P_{XY} P_X P_Y$$

Examples

HSIC (Gretton et al., 2005)
Distance Covariance (Szekely et al., 2007)

Result for U-statistics

Recall that the **permutation test** becomes **powerful** when

$$\text{“}\ \mathbb{E}_P[U_n] - \mathbb{E}_{P,\pi}[U_n^\pi] \ \gtrsim\ \sqrt{\beta^{-1}\mathbb{V}_P[U_n]} + \sqrt{\alpha^{-1}\beta^{-1}\mathbb{V}_{P,\pi}[U_n^\pi]}\ \text{”}$$

Result for U-statistics

Recall that the **permutation test** becomes powerful when

$$\text{“}\ \mathbb{E}_P[U_n] - \boxed{\mathbb{E}_{P,\pi}[U_n^\pi]} \gtrsim \sqrt{\beta^{-1}\mathbb{V}_P[U_n]} + \sqrt{\alpha^{-1}\beta^{-1}\boxed{\mathbb{V}_{P,\pi}[U_n^\pi]}}\ \text{”}$$

Question: Can we further simplify **these two moments** involving π ?

Result for U-statistics

Recall that the **permutation test** becomes powerful when

$$\text{“}\ \mathbb{E}_P[U_n] - \boxed{\mathbb{E}_{P,\pi}[U_n^\pi]} \gtrsim \sqrt{\beta^{-1}\mathbb{V}_P[U_n]} + \sqrt{\alpha^{-1}\beta^{-1}\boxed{\mathbb{V}_{P,\pi}[U_n^\pi]}}\ \text{”}$$

Question: Can we further simplify **these two moments** involving π ?

Lemma [KBW 2022]

$\mathbb{E}_{P,\pi}[U_n^\pi] = 0$ and $\mathbb{V}_{P,\pi}[U_n^\pi] \lesssim \mathbb{V}_P[U_n]$

Centered

Noise II
(data + permutation)

Noise I
(data)

Result for U-statistics

Recall that the **permutation test** becomes powerful when

$$\text{“}\ \mathbb{E}_P[U_n] - \mathbb{E}_{P,\pi}[U_n^\pi] \gtrsim \sqrt{\beta^{-1}\mathbb{V}_P[U_n]} + \sqrt{\alpha^{-1}\beta^{-1}\mathbb{V}_{P,\pi}[U_n^\pi]}\ \text{”}$$

Simplifies to

$$\boxed{\mathbb{E}_P[U_n] \gtrsim \sqrt{\alpha^{-1}\beta^{-1}\mathbb{V}_P[U_n]}}$$

Lemma [KBW 2022]

$\mathbb{E}_{P,\pi}[U_n^\pi] = 0$ and $\mathbb{V}_{P,\pi}[U_n^\pi] \lesssim \mathbb{V}_P[U_n]$

Centered — Noise II *(data + permutation)* — Noise I *(data)*

Question. Are these **sufficient** conditions for the power guarantee also **necessary**?

Applications to minimax testing

- **Minimax analysis:** $H_0 : P = Q$ vs. H_1 : distance$(P, Q) \geq \epsilon_n$

Minimum ϵ_n for which **minimax power** is nontrivial?

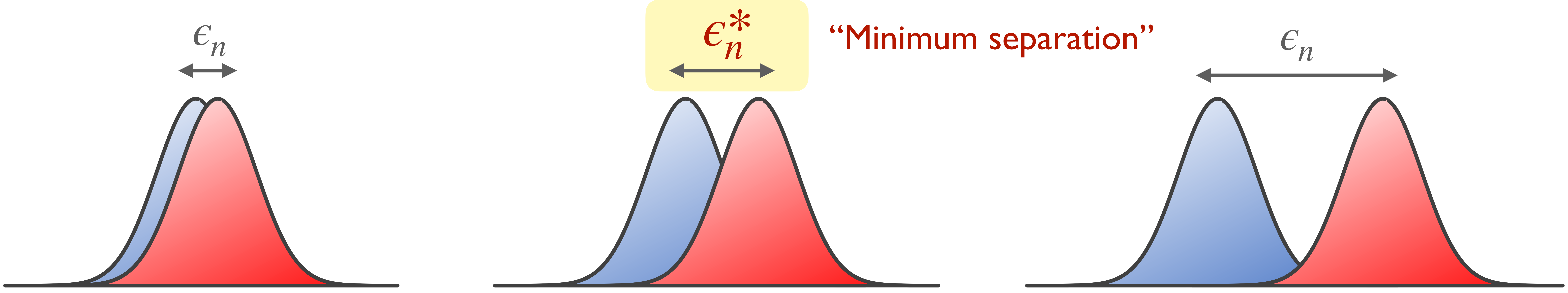
$$\underbrace{\sup_{\phi \in \Phi(\alpha)}}_{\text{Over all level } \alpha \text{ tests}} \underbrace{\inf_{P \in \mathscr{P}_1(\epsilon_n)} \mathbb{E}_P[\phi]}_{\text{Worst-case power}}$$

Applications to minimax testing

- **Minimax analysis:** $H_0 : P = Q$ vs. H_1 : distance$(P, Q) \geq \epsilon_n$

Minimum ϵ_n for which **minimax power** is nontrivial?

$$\underbrace{\sup_{\phi \in \Phi(\alpha)}}_{\text{Over all level } \alpha \text{ tests}} \underbrace{\inf_{P \in \mathscr{P}_1(\epsilon_n)} \mathbb{E}_P[\phi]}_{\text{Worst-case power}}$$



Minimax power ≤ 0.05 | Minimax power ≈ 0.8 | Minimax power $= 1$

Applications to minimax testing

- We call ϕ is minimax rate optimal if

$$\inf_{P \in \mathscr{P}_1(\epsilon_n)} \mathbb{E}_P[\phi] \geq 1 - \beta \quad \text{whenever } \epsilon_n^* \asymp \epsilon_n$$

Applications to minimax testing

- We call ϕ is minimax rate optimal if

$$\inf_{P \in \mathscr{P}_1(\epsilon_n)} \mathbb{E}_P[\phi] \geq 1 - \beta \quad \text{whenever } \epsilon_n^* \asymp \epsilon_n$$

- Initiated by Ingster (1987, 1994, 2000), Ingster and Suslina (2003)

I. Testing for multinomials: **(very incomplete)**

Paninski (2008), Chan et al. (2014), Bhattacharya & Valiant (2015), Diakonikolas & Kane (2016), Canonne et al. (2018), Balakrishnan & Wasserman (2019)

II. Testing for densities: **(very incomplete)**

Balasubramanian et al. (2017), Arias-Castro et al. (2018), Meynaoui et al. (2019), Li & Yuan (2019), Neykov et al. (2021), Berrett et al. (2021)

Previous tests depend on **unknown** constants

Example **Multinomial two-sample testing in ℓ_1-norm**

d: the number of bins
n: the sample size

- Chan et al. (2014) prove

$$\epsilon_n^* \asymp \max\left\{\frac{d^{1/2}}{n^{3/4}}, \frac{d^{1/4}}{n^{1/2}}\right\}$$

Previous tests depend on **unknown** constants

Example **Multinomial two-sample testing in ℓ_1-norm**

d: the number of bins
n: the sample size

- Chan et al. (2014) prove

$$\epsilon_n^* \asymp \max\left\{\frac{d^{1/2}}{n^{3/4}}, \frac{d^{1/4}}{n^{1/2}}\right\}$$

- The upper bound is based on the test

$$T_{\chi^2} > C\sqrt{\min\{n, d\}} \qquad \text{for some constant } C$$

Previous tests depend on **unknown** constants

Example **Density testing in ℓ_2-norm**

- Arias-Castro et al. (2018) prove

$$\epsilon_n^* \asymp n^{-\frac{2s}{4s+p}}$$

p: the dimension
s: Holder smoothness
n: the sample size

Previous tests depend on unknown constants

Example Density testing in ℓ_2-norm

- Arias-Castro et al. (2018) prove

$$\epsilon_n^* \asymp n^{-\frac{2s}{4s+p}}$$

p: the dimension
s: Holder smoothness
n: the sample size

- The upper bound is based on the test

$$T_{\mathrm{Bin}} > 2n + Cn^{\frac{4s+3p}{4s+p}} \quad \text{for some constant } C$$

Question. Can we prove the minimax results using a test that

i) tightly controls the type I error rate

ii) does not depend on unspecified constants

Question. Can we prove the minimax results using a test that

i) tightly controls the type I error rate

ii) does not depend on unspecified constants

Yes! Permutation tests

Two-sample testing

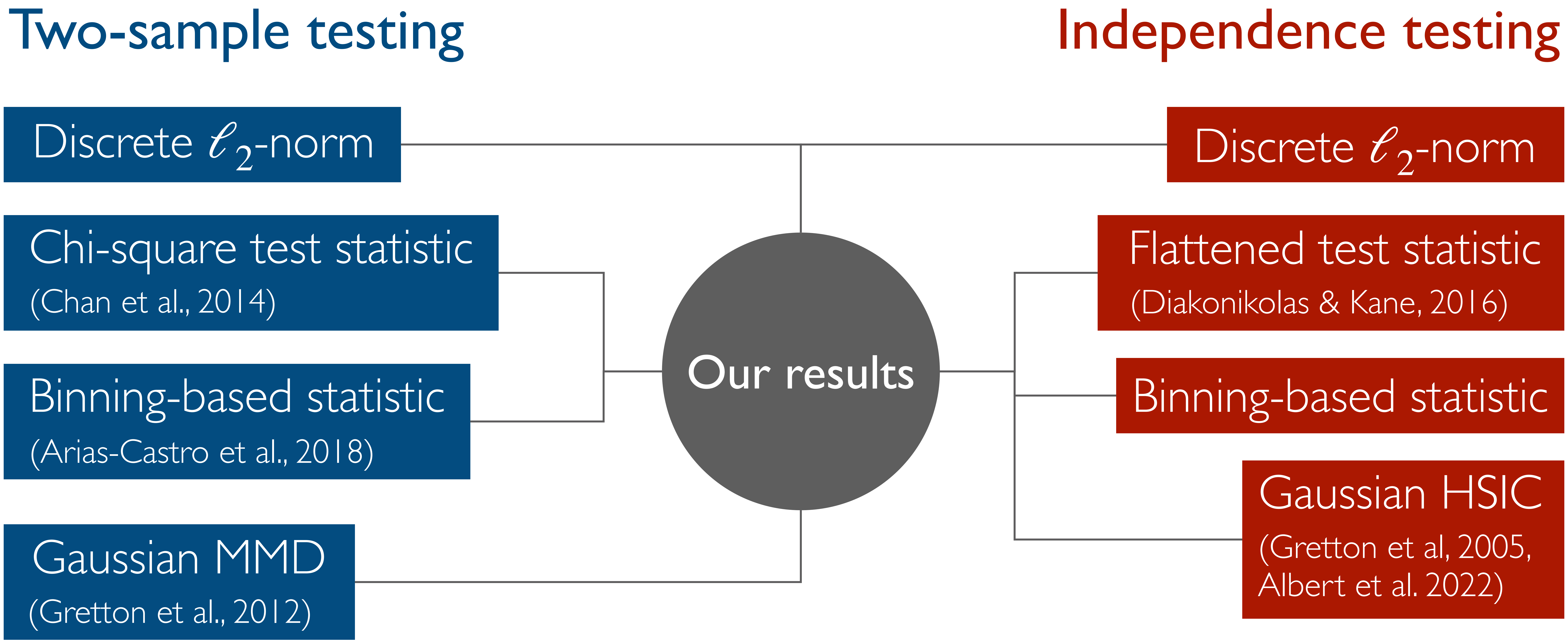
Discrete ℓ_2-norm

Chi-square test statistic
(Chan et al., 2014)

Binning-based statistic
(Arias-Castro et al., 2018)

Gaussian MMD
(Gretton et al., 2012)

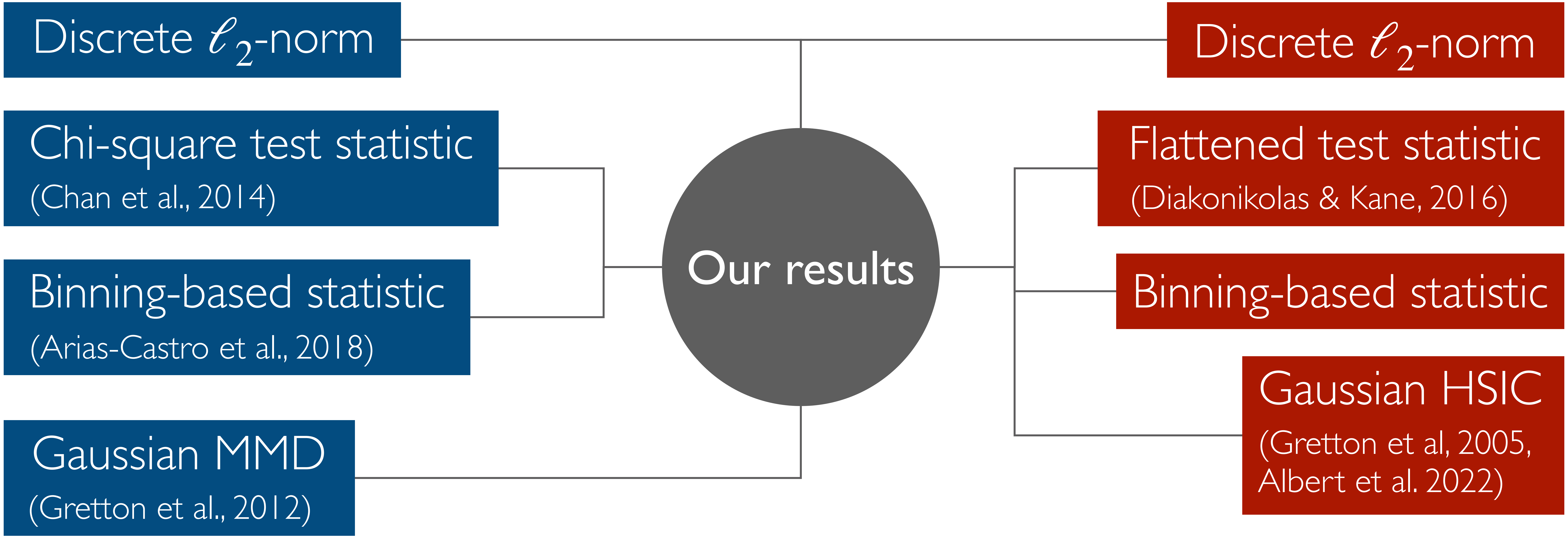
Our results



Two-sample testing
Independence testing
Discrete ℓ_2-norm
Chi-square test statistic
(Chan et al., 2014)
Binning-based statistic
(Arias-Castro et al., 2018)
Gaussian MMD
(Gretton et al., 2012)
Our results
Discrete ℓ_2-norm
Flattened test statistic
(Diakonikolas & Kane, 2016)
Binning-based statistic
Gaussian HSIC
(Gretton et al, 2005,
Albert et al. 2022)

Two-sample testing

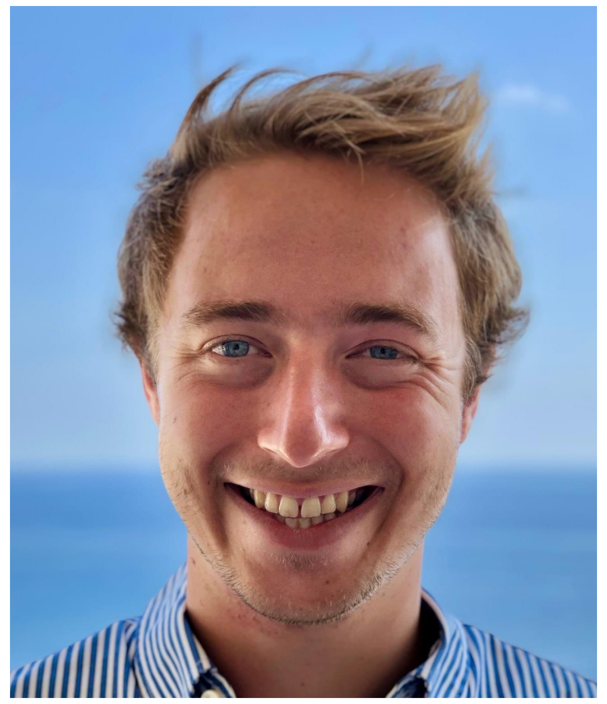
Independence testing



Permutation tests achieve **minimax optimality**

Follow-up

- Adaptivity
  - Schrab, Kim, Albert, Laurent, Guedj, Gretton (2023, *JMLR*)
- Computational-Power Trade-off
  - Schrab, Kim, Guedj, Gretton (2022, *NeurIPS*)
  - Choi, Kim (2024, *submitted*)
- Differential Privacy
  - Kim, Schrab (2024, *submitted*)
- Robustness
  - Schrab, Kim (2024, *submitted*)
- Conditional Independence
  - Kim, Neykov, Balakrishnan, Wasserman (2022, *AoS*)
  - Kim, Neykov, Balakrishnan, Wasserman (2024+, *EJS*)



Antonin Schrab

Benjamin Guedj

Arthur Gretton

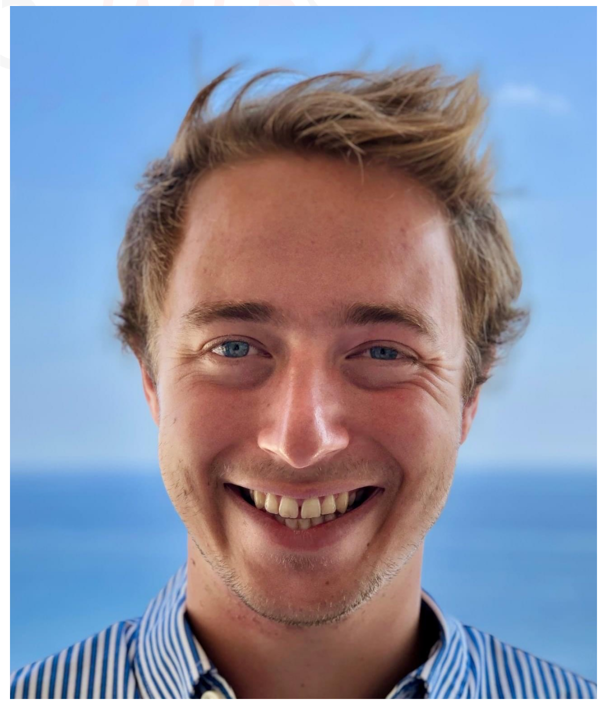
Mélisande Albert

Béatrice Laurent

Kernel two-sample test that **adapts to** unknown smoothness parameters

Follow-up

- Adaptivity
  - Schrab, Kim, Albert, Laurent, Guedj, Gretton (202
- Computational-Power Trade-off
  - Schrab, Kim, Guedj, Gretton (2022, *NeurIPS*)
  - Choi, Kim (2024, *submitted*)
- Differential Privacy
  - Kim, Schrab (2024, *submitted*)
- Robustness
  - Schrab, Kim (2024, *submitted*)
- Conditional Independence
  - Kim, Neykov, Balakrishnan, Wasserman (2022, *AoS*)
  - Kim, Neykov, Balakrishnan, Wasserman (2024+, *EJS*)


Antonin Schrab


Benjamin Guedj

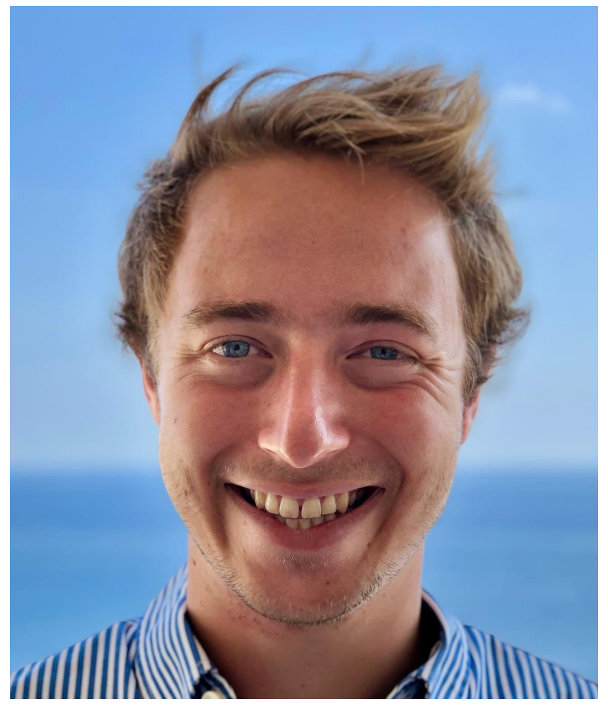

Arthur Gretton


Ikjun Choi

- Computational-power trade-offs in nonparametric testing
- Implemented via incomplete U-statistics and random Fourier features

Follow-up

- Adaptivity
  - Schrab, Kim, Albert, Laurent, Guedj, Gretton (2023, *JMLR*)
- Computational-Power Trade-off
  - Schrab, Kim, Guedj, Gretton (2022, *NeurIPS*)
  - Choi, Kim (2024, *submitted*)
- Differential Privacy
  - Kim, Schrab (2024, *submitted*)
- Robustness
  - Schrab, Kim (2024, *submitted*)
- Conditional Independence
  - Kim, Neykov, Balakrishnan, Wasserman (2022, *AoS*)
  - Kim, Neykov, Balakrishnan, Wasserman (2024+, *EJS*)

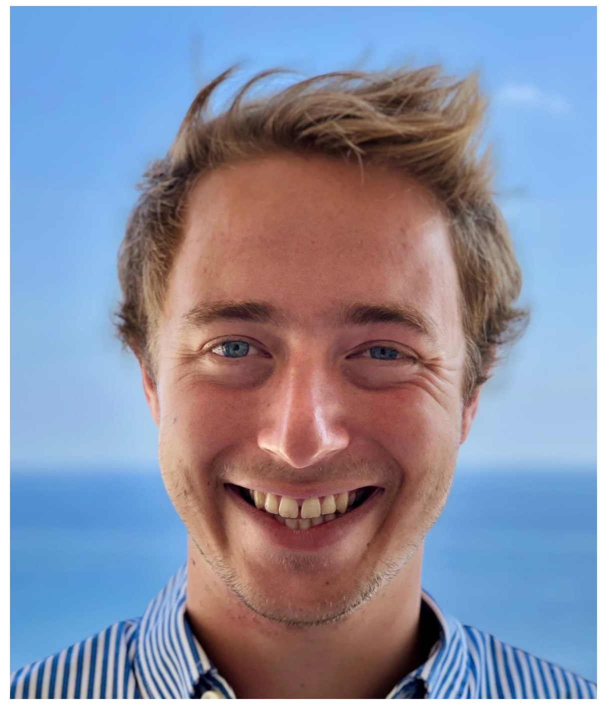


Antonin Schrab

Differentially private permutation tests applied to kernel methods

Follow-up

- Adaptivity
  - Schrab, Kim, Albert, Laurent, Guedj, Gretton (2023, *JMLR*)
- Computational-Power Trade-off
  - Schrab, Kim, Guedj, Gretton (2022, *NeurIPS*)
  - Choi, Kim (2024, *submitted*)
- Differential Privacy
  - Kim, Schrab (2024, *submitted*)
- Robustness
  - Schrab, Kim (2024, *submitted*)
- Conditional Independence
  - Kim, Neykov, Balakrishnan, Wasserman (2022, *AoS*)
  - Kim, Neykov, Balakrishnan, Wasserman (2024, *EJS+*)



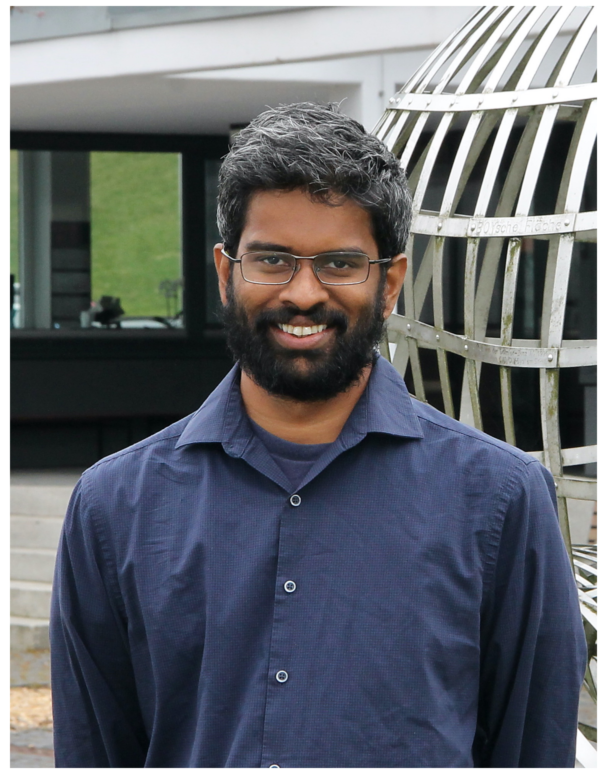
Antonin Schrab

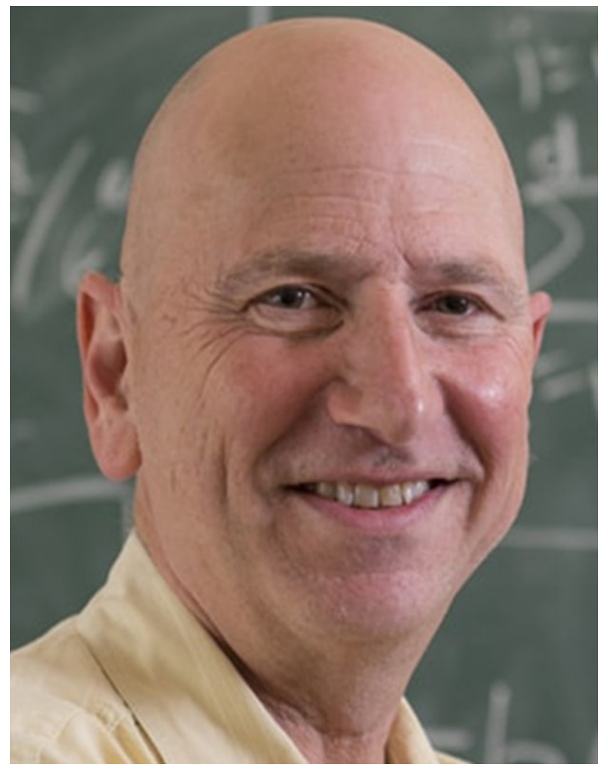
Permutation tests that are robust to data perturbation with optimal properties

Follow-up

- Adaptivity
  - Schrab, Kim, Albert, Laurent, Guedj, Gretton (2023, *JMLR*)
- Computational-Power Trade-off
  - Schrab, Kim, Guedj, Gretton (2022, *NeurIPS*)
  - Choi, Kim (2024, *submitted*)
- Differential Privacy
  - Kim, Schrab (2024, *submitted*)
- Robustness
  - Schrab, Kim (2024, *submitted*)
- Conditional Independence
  - Kim, Neykov, Balakrishnan, Wasserman (2022, *AoS*)
  - Kim, Neykov, Balakrishnan, Wasserman (2024+, *EJS*)


Matey Neykov


Siva Balakrishnan


Larry Wasserman

Local permutation tests for conditional independence with theoretical guarantees

Summary

- Permutation tests have **uniform, finite-sample** guarantees for any test statistic

- We have introduced a **flexible framework** for comparing distributions based on **classification and regression**

- We have provided **tools for analyzing the power** of permutation tests e.g. two moments method, exponential concentration bounds

- Permutation tests are **minimax optimal** in many problems

Summary

- Permutation tests have **uniform, finite-sample** guarantees for any test statistic

- We have introduced a **flexible framework** for comparing distributions based on **classification and regression**

- We have provided **tools for analyzing the power** of permutation tests e.g. two moments method, exponential concentration bounds

- Permutation tests are **minimax optimal** in many problems

Summary

- Permutation tests have **uniform, finite-sample** guarantees for any test statistic
- We have introduced a **flexible framework** for comparing distributions based on **classification and regression**
- We have provided **tools for analyzing the power** of permutation tests e.g. two moments method, exponential concentration bounds
- Permutation tests are **minimax optimal** in many problems

Summary

- Permutation tests have **uniform, finite-sample** guarantees for any test statistic
- We have introduced a **flexible framework** for comparing distributions based on **classification and regression**
- We have provided **tools for analyzing the power** of permutation tests e.g. two moments method, exponential concentration bounds
- Permutation tests are **minimax optimal** in many problems

Thank you!