# Hypothesis testing: randomisation, privacy and minimax optimality

Tom Berrett
University of Warwick

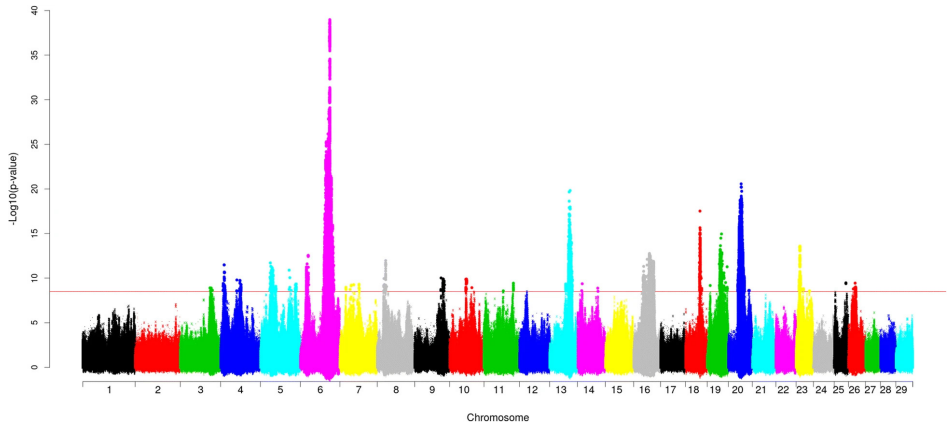
CUSO Winter School

1–4th February 2026

# Hypothesis testing

Testing hypotheses is at the heart of statistics. Basic inferential questions: making statements of *confidence* and *quantifying uncertainty*.

```
> attach(mtcars)
> summary(lm(mpg~disp+hp+drat+wt))

Call:
lm(formula = mpg ~ disp + hp + drat + wt)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5077 -1.9052 -0.5057  0.9821  5.6883

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.148738   6.293588   4.631  8.2e-05 ***
disp         0.003815   0.010805   0.353  0.72675
hp          -0.034784   0.011597  -2.999  0.00576 **
drat         1.768049   1.319779   1.340  0.19153
wt          -3.479668   1.078371  -3.227  0.00327 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.602 on 27 degrees of freedom
Multiple R-squared:  0.8376,    Adjusted R-squared:  0.8136
F-statistic: 34.82 on 4 and 27 DF,  p-value: 2.704e-10

>
```




In scientific application we often want to answer a qualitative question: is this variable important? is the new treatment better than the old one? Does my model explain what is happening?

# Property testing

As well as in statistics, modern research is done in computer science and information theory (e.g. Paninski, 2008; Diakonikolas & Kane, 2016; Canonne, 2020).

In *distribution testing* we see that properties of complex distributions can be tested with much less data than would be needed to estimate them. Are test and train distributions the same? has customer behaviour changed?

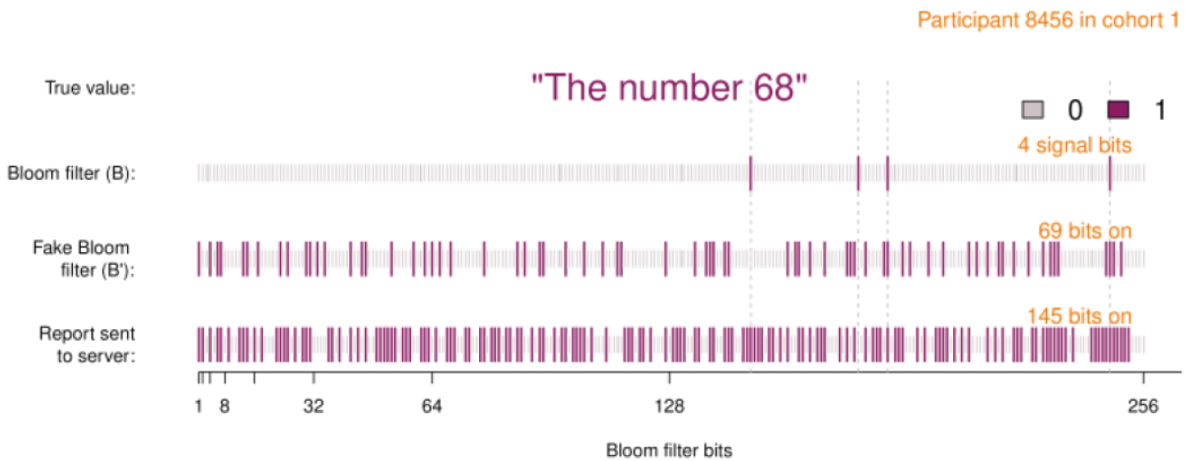



Figure: `RAPPOR` mechanism (Erlingsson, Pihur & Korolova, 2014)

When distributions are complex, traditional asymptotic approaches are typically not appropriate. Instead we take a finite-sample, minimax view.

# Minimax optimality theory

Hypothesis testing is the foundation for much modern statistical theory.

$$\mathcal{R}_n := \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\hat{\theta}, \theta).$$

Most minimax lower bounds are based on reductions from estimation to testing problems for minimax lower bounds.

If we can prove that $\mathcal{R}_n \asymp \psi_n$ we can

- Certify that our estimators are *rate optimal*
- Quantify the inherent difficulty of a statistical problem, study the effect on statistical utility of
  - computational/communication resources
  - contaminated/missing data
  - privacy/fairness constraints
  - ...

# Overview

1. Introduction

2. Simple null hypotheses
   - The likelihood ratio test and Neyman–Pearson
   - Composite alternatives: Goodness-of-fit testing

3. Permutation testing for composite nulls
   - Independence testing – minimax optimality
   - Conditional independence testing – non-uniform permutations

4. Local differential privacy
   - Two-point testing
   - Goodness-of-fit testing

1. Introduction

2. Simple null hypotheses
   - The likelihood ratio test and Neyman–Pearson
   - Composite alternatives: Goodness-of-fit testing

3. Permutation testing for composite nulls
   - Independence testing – minimax optimality
   - Conditional independence testing – non-uniform permutations

4. Local differential privacy
   - Two-point testing
   - Goodness-of-fit testing

# Simple hypothesis testing

One of the simplest and most studied statistical problems there is. Test

$$H_0 : P = P_0 \quad \text{vs.} \quad H_1 : P = P_1$$

given independent and indentically distributed data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$.

For us, a *test* is a function $\phi : \mathcal{X}^n \to [0, 1]$ with the interpretation that we reject $H_0$ with probability $\phi(X_1, \ldots, X_n)$.

Write $L(x) \equiv L(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{p_1(x_i)}{p_0(x_i)}$ where $p_0, p_1$ are densities of $P_0, P_1$ w.r.t. some dominating measure. *Likelihood Ratio Tests* are of the form

$$\phi_k(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } L(x) > k \\ c & \text{if } L(x) = k \\ 0 & \text{if } L(x) < k \end{cases},$$

where $k > 0$ and $c \in [0, 1]$ are chosen such that $\mathbb{E}_{P_0} \phi_k(X) = \alpha$, and $\alpha \in (0, 1)$ is the desired significance level.

# Simple hypothesis testing

One of the simplest and most studied statistical problems there is. Test

$$H_0 : P = P_0 \quad \text{vs.} \quad H_1 : P = P_1$$

given independent and indentically distributed data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$.

For us, a *test* is a function $\phi : \mathcal{X}^n \to [0, 1]$ with the interpretation that we reject $H_0$ with probability $\phi(X_1, \ldots, X_n)$.

Write $L(x) \equiv L(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{p_1(x_i)}{p_0(x_i)}$ where $p_0, p_1$ are densities of $P_0, P_1$ w.r.t. some dominating measure. *Likelihood Ratio Tests* are of the form

$$\phi_k(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } L(x) > k \\ c & \text{if } L(x) = k \\ 0 & \text{if } L(x) < k \end{cases},$$

where $k > 0$ and $c \in [0, 1]$ are chosen such that $\mathbb{E}_{P_0} \phi_k(X) = \alpha$, and $\alpha \in (0, 1)$ is the desired significance level.

# Neyman–Pearson Lemma

We are interested in LRTs because they are exactly *most powerful*.

**Theorem (Neyman–Pearson Lemma, Lehmann & Romano 2005)**

*Fix $\alpha \in (0,1)$ and let $\phi_k$ be a LRT such that $\mathbb{E}_{P_0}\phi_k(X) = \alpha$. Let $\phi$ be any test satisfying $\mathbb{E}_{P_0}\phi(X) \leq \alpha$. Then*

$$\mathbb{E}_{P_1}\phi(X) \leq \mathbb{E}_{P_1}\phi_k(X).$$

For fixed significance level $\alpha$, no *valid* test correctly rejects the null more often than LRTs.

It is rare that we are able to give such strong guarantees for a method (but this is a very simple problem).

# Neyman–Pearson Lemma

We are interested in LRTs because they are exactly *most powerful*.

**Theorem (Neyman–Pearson Lemma, Lehmann & Romano 2005)**

*Fix $\alpha \in (0,1)$ and let $\phi_k$ be a LRT such that $\mathbb{E}_{P_0}\phi_k(X) = \alpha$. Let $\phi$ be any test satisfying $\mathbb{E}_{P_0}\phi(X) \leq \alpha$. Then*

$$\mathbb{E}_{P_1}\phi(X) \leq \mathbb{E}_{P_1}\phi_k(X).$$

For fixed significance level $\alpha$, no *valid* test correctly rejects the null more often than LRTs.

It is rare that we are able to give such strong guarantees for a method (but this is a very simple problem).

# Optimal error rate

If we care about Type I and Type II errors equally, then $\phi_1$ is optimal and we can characterise its error rate through the Total Variation distance: for any test $\phi(X_1, \ldots, X_n)$ we have

$$\begin{aligned}
\mathbb{E}_{P_0}\phi + \mathbb{E}_{P_1}(1-\phi) &\geq 1 - \sup_{f:\mathcal{X}^n \to [0,1]} (\mathbb{E}_{P_1} f - \mathbb{E}_{P_0} f) \\
&= 1 - \{\mathbb{P}_{P_1}(L \geq 1) - \mathbb{P}_{P_0}(L \geq 1)\} \\
&= 1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \\
&= \mathbb{E}_{P_0}\phi_1 + \mathbb{E}_{P_1}(1-\phi_1).
\end{aligned}$$

There exist good tests of $H_0$ vs. $H_1$ if and only if $\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})$ is large.

When $n$ is large, $\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})$ is not easy to interpret.

# Optimal error rate

If we care about Type I and Type II errors equally, then $\phi_1$ is optimal and we can characterise its error rate through the Total Variation distance: for any test $\phi(X_1, \ldots, X_n)$ we have

$$\begin{aligned}
\mathbb{E}_{P_0}\phi + \mathbb{E}_{P_1}(1-\phi) &\geq 1 - \sup_{f:\mathcal{X}^n \to [0,1]} (\mathbb{E}_{P_1} f - \mathbb{E}_{P_0} f) \\
&= 1 - \{\mathbb{P}_{P_1}(L \geq 1) - \mathbb{P}_{P_0}(L \geq 1)\} \\
&= 1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \\
&= \mathbb{E}_{P_0}\phi_1 + \mathbb{E}_{P_1}(1-\phi_1).
\end{aligned}$$

There exist good tests of $H_0$ vs. $H_1$ if and only if $\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})$ is large.

When $n$ is large, $\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})$ is not easy to interpret.

# Optimal error rate

If we care about Type I and Type II errors equally, then $\phi_1$ is optimal and we can characterise its error rate through the Total Variation distance: for any test $\phi(X_1, \ldots, X_n)$ we have

$$\begin{aligned}
\mathbb{E}_{P_0}\phi + \mathbb{E}_{P_1}(1-\phi) &\geq 1 - \sup_{f:\mathcal{X}^n \to [0,1]} (\mathbb{E}_{P_1} f - \mathbb{E}_{P_0} f) \\
&= 1 - \{\mathbb{P}_{P_1}(L \geq 1) - \mathbb{P}_{P_0}(L \geq 1)\} \\
&= 1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \\
&= \mathbb{E}_{P_0}\phi_1 + \mathbb{E}_{P_1}(1-\phi_1).
\end{aligned}$$

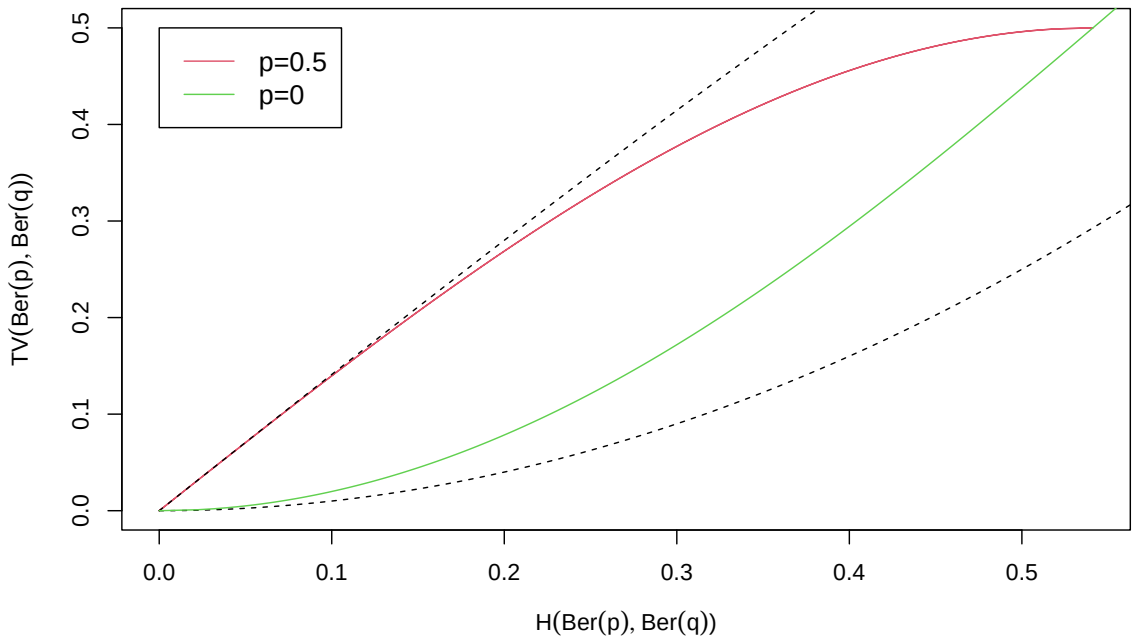
There exist good tests of $H_0$ vs. $H_1$ if and only if $\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})$ is large.

When $n$ is large, $\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})$ is not easy to interpret.

# Hellinger distance

The Hellinger distance $H^2(Q_1, Q_2) = 1 - \int \sqrt{q_1 q_2}$ is related to TV:

$$H^2(Q_1, Q_2) \leq \mathrm{TV}(Q_1, Q_2) \leq \sqrt{2} H(Q_1, Q_2) \left\{ 1 - \frac{1}{2} H^2(Q_1, Q_2) \right\}^{1/2}.$$



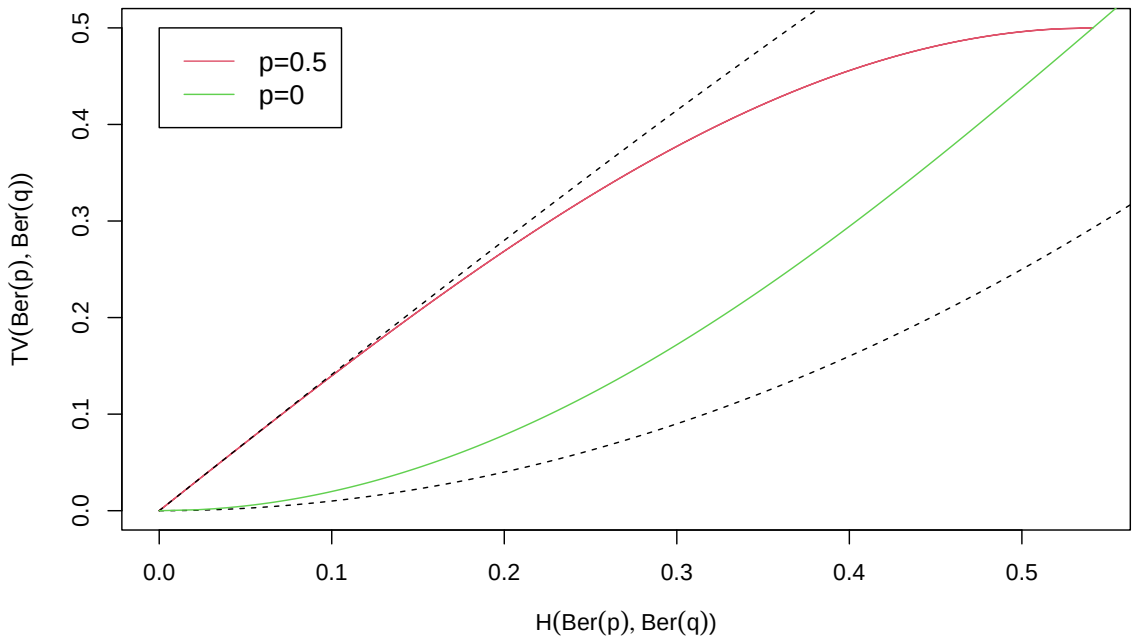

Hellinger is useful for its tensorisation property

$$H_n^2 := H^2(P_0^{\otimes n}, P_1^{\otimes n}) = 1 - \left\{ 1 - H^2(P_0, P_1) \right\}^n =: 1 - (1 - H^2)^n.$$

# Hellinger distance

The Hellinger distance $H^2(Q_1, Q_2) = 1 - \int \sqrt{q_1 q_2}$ is related to TV:

$$H^2(Q_1, Q_2) \leq \mathrm{TV}(Q_1, Q_2) \leq \sqrt{2} H(Q_1, Q_2) \left\{ 1 - \frac{1}{2} H^2(Q_1, Q_2) \right\}^{1/2}.$$




Hellinger is useful for its tensorisation property

$$H_n^2 := H^2(P_0^{\otimes n}, P_1^{\otimes n}) = 1 - \left\{ 1 - H^2(P_0, P_1) \right\}^n =: 1 - (1 - H^2)^n.$$

# Behaviour for large $n$

These properties can be used to show that

$$\frac{1}{2}\{1 - H^2(P_0, P_1)\}^{2n} \leq 1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq \{1 - H^2(P_0, P_1)\}^n.$$

Optimal error decays exponentially in $n$, with rate of decay $\asymp H^2$.

When studying test performance it is convenient to think about separation conditions or sample complexity. We can check that

**Hellinger separation**

- $H^2 \geq \frac{\log(\frac{1}{\delta})}{n} \implies \exists$ test with Type I + Type II error $\leq \delta$
- $H^2 < \frac{\log(\frac{1}{4\delta})}{4n} \leq \frac{1}{2} \implies$ all tests have Type I + Type II error $> \delta$.

There are good tests if and only if $P_0, P_1$ are separated in Hellinger dist.[1].

[1] If Type I and Type II error are weighted very differently things are more complicated (Pensia, Jog & Loh, 2024)

# Behaviour for large $n$

These properties can be used to show that

$$\frac{1}{2}\{1 - H^2(P_0, P_1)\}^{2n} \leq 1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq \{1 - H^2(P_0, P_1)\}^n.$$

Optimal error decays exponentially in $n$, with rate of decay $\asymp H^2$.

When studying test performance it is convenient to think about separation conditions or sample complexity. We can check that

## Hellinger separation

- $H^2 \geq \frac{\log(\frac{1}{\delta})}{n} \implies \exists$ test with Type I + Type II error $\leq \delta$
- $H^2 < \frac{\log(\frac{1}{4\delta})}{4n} \leq \frac{1}{2} \implies$ all tests have Type I + Type II error $> \delta$.

There are good tests if and only if $P_0, P_1$ are separated in Hellinger dist.[1].

[1] If Type I and Type II error are weighted very differently things are more complicated (Pensia, Jog & Loh, 2024)

# Behaviour for large $n$

These properties can be used to show that

$$\frac{1}{2}\{1 - H^2(P_0, P_1)\}^{2n} \leq 1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq \{1 - H^2(P_0, P_1)\}^n.$$

Optimal error decays exponentially in $n$, with rate of decay $\asymp H^2$.

When studying test performance it is convenient to think about separation conditions or sample complexity. We can check that

## Hellinger separation

- $H^2 \geq \frac{\log(\frac{1}{\delta})}{n} \implies \exists$ test with Type I + Type II error $\leq \delta$
- $H^2 < \frac{\log(\frac{1}{4\delta})}{4n} \leq \frac{1}{2} \implies$ all tests have Type I + Type II error $> \delta$.

There are good tests if and only if $P_0, P_1$ are separated in Hellinger dist.[1].

[1] If Type I and Type II error are weighted very differently things are more complicated (Pensia, Jog & Loh, 2024)

# Two–point testing – conclusion

- The Neyman–Pearson Lemma gives us the exactly optimal test for the classical model.
- We can give bounds on the power of this test, thus characterising the fundamental difficulty of the problem through the Hellinger distance.

Later, we will see that Hellinger is no longer the right distance for robust or private problems.

- Beyond this simple problem (e.g. when alternatives are *composite*) it is rare to find exactly optimal tests.
- If we fix a separation metric of interest, we can often find good tests and establish their rate optimality.

# Two–point testing – conclusion

- The Neyman–Pearson Lemma gives us the exactly optimal test for the classical model.
- We can give bounds on the power of this test, thus characterising the fundamental difficulty of the problem through the Hellinger distance.

Later, we will see that Hellinger is no longer the right distance for robust or private problems.

- Beyond this simple problem (e.g. when alternatives are *composite*) it is rare to find exactly optimal tests.
- If we fix a separation metric of interest, we can often find good tests and establish their rate optimality.

# Two–point testing – conclusion

- The Neyman–Pearson Lemma gives us the exactly optimal test for the classical model.
- We can give bounds on the power of this test, thus characterising the fundamental difficulty of the problem through the Hellinger distance.

Later, we will see that Hellinger is no longer the right distance for robust or private problems.

- Beyond this simple problem (e.g. when alternatives are *composite*) it is rare to find exactly optimal tests.
- If we fix a separation metric of interest, we can often find good tests and establish their rate optimality.

# Goodness-of-fit testing

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, a basic problem with composite alternative is

$$H_0 : P = P_0 \quad \text{vs.} \quad H_1 : P \neq P_0,$$

It is not possible to have uniformly powerful tests here, so we study power by introducing separation.




The minimal separation needed for acceptable error, e.g.

$$\rho^*(\phi) = \inf\left\{\rho \in (0,1) : \mathbb{E}_{P_0}(\phi) + \sup_{P : d(P,P_0) \geq \rho} \mathbb{E}(1-\phi) \leq 1/4\right\}$$

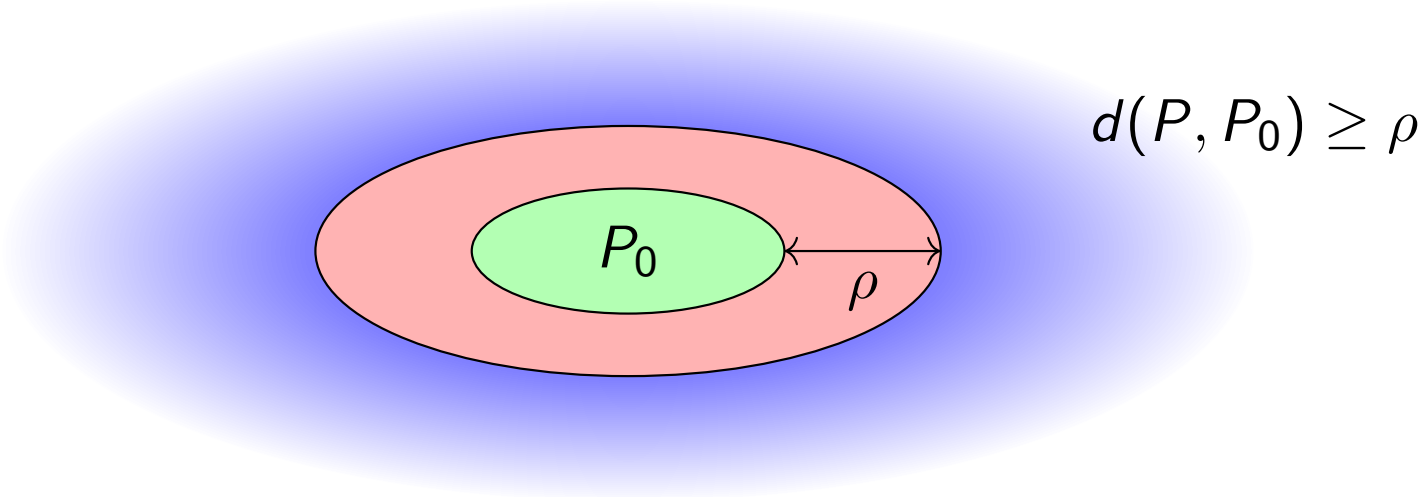
measures the power of $\phi$. The *minimax separation* is $\rho^* = \inf_\phi \rho^*(\phi)$.

# Goodness-of-fit testing

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, a basic problem with composite alternative is

$$H_0 : P = P_0 \quad \text{vs.} \quad H_1 : P \neq P_0,$$

It is not possible to have uniformly powerful tests here, so we study power by introducing separation.




The minimal separation needed for acceptable error, e.g.

$$\rho^*(\phi) = \inf\left\{\rho \in (0,1) : \mathbb{E}_{P_0}(\phi) + \sup_{P : d(P,P_0) \geq \rho} \mathbb{E}(1-\phi) \leq 1/4\right\}$$

measures the power of $\phi$. The *minimax separation* is $\rho^* = \inf_\phi \rho^*(\phi)$.

# Discrete uniformity testing

The minimax separation $\rho^* = \inf_\phi \rho^*(\phi)$ quantifies the difficulty of the testing problem.

Suppose $X_1, \dots, X_n \sim P$ take values in $[d] = \{1, \dots, d\}$ and we test

$$H_0 : P = \mathrm{Unif}([d]) \quad \text{vs.} \quad H_1(\rho) : \mathrm{TV}(P, \mathrm{Unif}([d])) \geq \rho.$$

**Theorem (e.g. Balakrishnan & Wasserman 2019)**

*In this setting, when $d^{1/2} \leq n$ there exist universal constants $c, C > 0$ such that*

$$\frac{c\, d^{1/4}}{n^{1/2}} \leq \rho^* \leq \frac{C\, d^{1/4}}{n^{1/2}},$$

*i.e. the minimax optimal separation rate is $d^{1/4}/n^{1/2}$.*

Note that the optimal rate for estimation of $P$ is $\mathrm{TV}(\hat{P}, P) \asymp (d/n)^{1/2}$. In this sense, testing is easier than estimation.

# Discrete uniformity testing

The minimax separation $\rho^* = \inf_\phi \rho^*(\phi)$ quantifies the difficulty of the testing problem.

Suppose $X_1, \dots, X_n \sim P$ take values in $[d] = \{1, \dots, d\}$ and we test

$$H_0 : P = \mathrm{Unif}([d]) \quad \text{vs.} \quad H_1(\rho) : \mathrm{TV}(P, \mathrm{Unif}([d])) \geq \rho.$$

**Theorem (e.g. Balakrishnan & Wasserman 2019)**

*In this setting, when $d^{1/2} \leq n$ there exist universal constants $c, C > 0$ such that*

$$\frac{c\, d^{1/4}}{n^{1/2}} \leq \rho^* \leq \frac{C\, d^{1/4}}{n^{1/2}},$$

*i.e. the minimax optimal separation rate is $d^{1/4}/n^{1/2}$.*

Note that the optimal rate for estimation of $P$ is $\mathrm{TV}(\hat{P}, P) \asymp (d/n)^{1/2}$. In this sense, testing is easier than estimation.

# Discrete uniformity testing

The minimax separation $\rho^* = \inf_\phi \rho^*(\phi)$ quantifies the difficulty of the testing problem.

Suppose $X_1, \ldots, X_n \sim P$ take values in $[d] = \{1, \ldots, d\}$ and we test

$$H_0 : P = \mathrm{Unif}([d]) \quad \text{vs.} \quad H_1(\rho) : \mathrm{TV}(P, \mathrm{Unif}([d])) \geq \rho.$$

**Theorem (e.g. Balakrishnan & Wasserman 2019)**

*In this setting, when $d^{1/2} \leq n$ there exist universal constants $c, C > 0$ such that*

$$\frac{c\, d^{1/4}}{n^{1/2}} \leq \rho^* \leq \frac{C\, d^{1/4}}{n^{1/2}},$$

*i.e. the minimax optimal separation rate is $d^{1/4}/n^{1/2}$.*

Note that the optimal rate for estimation of $P$ is $\mathrm{TV}(\hat{P}, P) \asymp (d/n)^{1/2}$. In this sense, testing is easier than estimation.

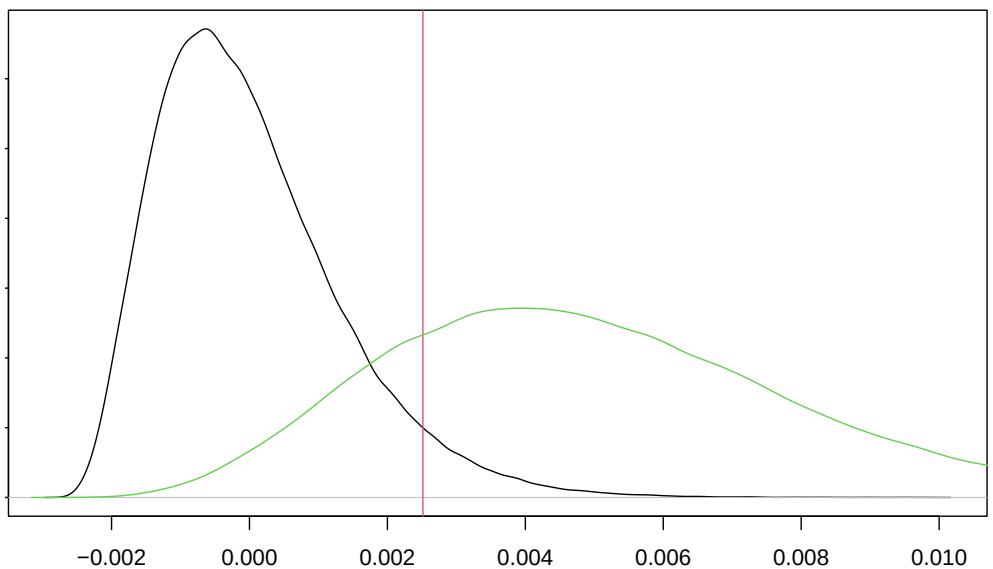
# Optimal tests

We can prove an upper bound on $\rho^*$ by choosing a test $\phi$ and bounding its error, since $\rho^* \leq \rho^*(\phi)$.

Define kernel $h(x, y) = \mathbb{1}_{\{x=y\}} - 1/d$ and test statistic

$$T = T(X_1, \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} h(X_{i_1}, X_{i_2}) = \frac{1}{n(n-1)} \sum_{j=1}^{d} N_j^2 + \text{const.}$$

For suitable critical value $C_{d,n} \asymp \frac{1}{n\sqrt{d}}$ we consider the test $\phi = \mathbb{1}_{\{T \geq C_{d,n}\}}$.




$d = 9, n = 333$

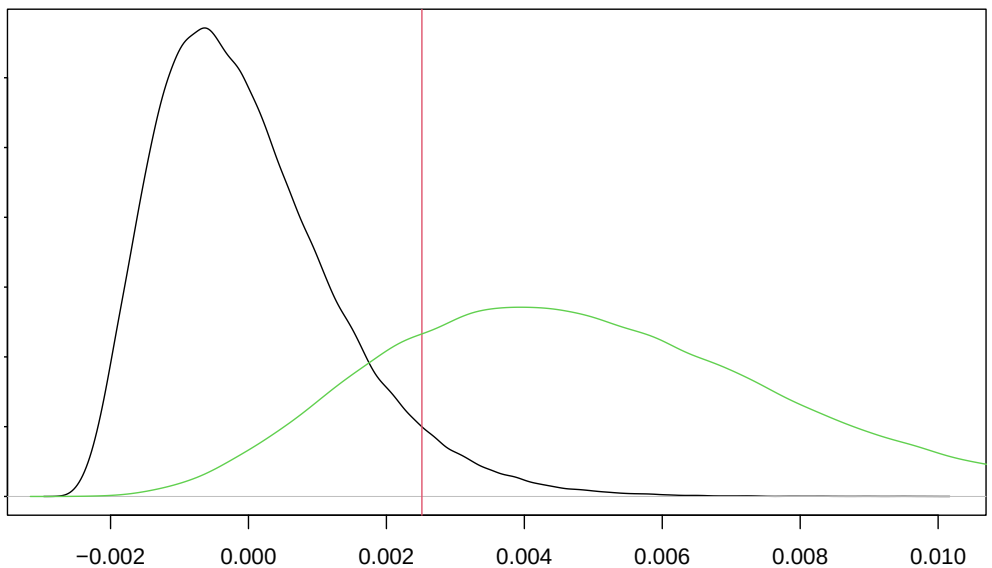
# Optimal tests

We can prove an upper bound on $\rho^*$ by choosing a test $\phi$ and bounding its error, since $\rho^* \leq \rho^*(\phi)$.

Define kernel $h(x, y) = \mathbb{1}_{\{x=y\}} - 1/d$ and test statistic

$$T = T(X_1, \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} h(X_{i_1}, X_{i_2}) = \frac{1}{n(n-1)} \sum_{j=1}^{d} N_j^2 + \text{const.}$$

For suitable critical value $C_{d,n} \asymp \frac{1}{n\sqrt{d}}$ we consider the test $\phi = \mathbb{1}_{\{T \geq C_{d,n}\}}$.




$d = 9, n = 333$

# Minimax upper bound

To analyse this test it suffices to bound the mean and variance of $T$. First,

$$\mathbb{E}_P(T) = \mathbb{E}\{h(X_1, X_2)\} = \sum_{j=1}^{d} p(j)^2 - 1/d = \sum_{j=1}^{d} \{p(j) - 1/d\}^2$$
$$= \|p - p_0\|_2^2.$$

Since $T$ is a second order $U$-statistic, we know (e.g. Serfling, 2009) that

$$\mathrm{Var}_P(T) = \frac{2(n-2)}{\binom{n}{2}} \mathrm{Var}\Big(\mathbb{E}\{h(X_1, X_2)|X_1\}\Big) + \frac{1}{\binom{n}{2}} \mathrm{Var}\big(h(X_1, X_2)\big)$$
$$= \frac{4(n-2)}{n(n-1)} \mathrm{Var}\big(p(X_1)\big) + \frac{2}{n(n-1)} \mathrm{Var}(\mathbb{1}_{\{X_1 = X_2\}}).$$

Under $H_0$ the first term vanishes and we get $\mathrm{Var}_P(T) \leq \frac{2}{dn(n-1)}$.

By Chebyshev's inequality we see that taking $C_{d,n} = 4/\{dn(n-1)\}^{1/2}$ gives a test with Type I error $\leq 1/8$.

# Minimax upper bound

To analyse this test it suffices to bound the mean and variance of $T$. First,

$$\mathbb{E}_P(T) = \mathbb{E}\{h(X_1, X_2)\} = \sum_{j=1}^{d} p(j)^2 - 1/d = \sum_{j=1}^{d} \{p(j) - 1/d\}^2$$
$$= \|p - p_0\|_2^2.$$

Since $T$ is a second order $U$-statistic, we know (e.g. Serfling, 2009) that

$$\mathrm{Var}_P(T) = \frac{2(n-2)}{\binom{n}{2}} \mathrm{Var}\Big(\mathbb{E}\{h(X_1, X_2)|X_1\}\Big) + \frac{1}{\binom{n}{2}} \mathrm{Var}\big(h(X_1, X_2)\big)$$
$$= \frac{4(n-2)}{n(n-1)} \mathrm{Var}\big(p(X_1)\big) + \frac{2}{n(n-1)} \mathrm{Var}\big(\mathbb{1}_{\{X_1 = X_2\}}\big).$$

Under $H_0$ the first term vanishes and we get $\mathrm{Var}_P(T) \leq \frac{2}{dn(n-1)}$.

By Chebyshev's inequality we see that taking $C_{d,n} = 4/\{dn(n-1)\}^{1/2}$ gives a test with Type I error $\leq 1/8$.

# Minimax upper bound

To analyse this test it suffices to bound the mean and variance of $T$. First,

$$\mathbb{E}_P(T) = \mathbb{E}\{h(X_1, X_2)\} = \sum_{j=1}^{d} p(j)^2 - 1/d = \sum_{j=1}^{d} \{p(j) - 1/d\}^2$$
$$= \|p - p_0\|_2^2.$$

Since $T$ is a second order $U$-statistic, we know (e.g. Serfling, 2009) that

$$\mathrm{Var}_P(T) = \frac{2(n-2)}{\binom{n}{2}} \mathrm{Var}\Big(\mathbb{E}\{h(X_1, X_2)|X_1\}\Big) + \frac{1}{\binom{n}{2}} \mathrm{Var}\big(h(X_1, X_2)\big)$$
$$= \frac{4(n-2)}{n(n-1)} \mathrm{Var}\big(p(X_1)\big) + \frac{2}{n(n-1)} \mathrm{Var}(\mathbb{1}_{\{X_1 = X_2\}}).$$

Under $H_0$ the first term vanishes and we get $\mathrm{Var}_P(T) \leq \frac{2}{dn(n-1)}$.

By Chebyshev's inequality we see that taking $C_{d,n} = 4/\{dn(n-1)\}^{1/2}$ gives a test with Type I error $\leq 1/8$.

# Minimax upper bound

More generally, we can see that

$$\mathrm{Var}_P(T) \lesssim \frac{\mathbb{E}_P(T)}{n}(1/d + 1/n) + \frac{\{\mathbb{E}_P(T)\}^{3/2}}{n} + \frac{1}{dn^2}.$$

Thus, when $\mathbb{E}_P(T) = \|p - p_0\|_2^2 \geq 2C_{d,n} \asymp 1/(n\sqrt{d})$ we have

$$\frac{\mathrm{Var}_P(T)}{(\mathbb{E}_P T)^2} \lesssim \frac{\sqrt{d}}{n} + \frac{1}{\sqrt{d}} + \frac{d^{1/4}}{\sqrt{n}} + 1 \lesssim 1.$$

By Chebyshev's inequality, there exists universal constant $B > 0$ such that Type II error $\leq 1/8$ when $\|p - p_0\|_2^2 \geq B/(n\sqrt{d})$.

Since $\|p - p_0\|_2^2 \geq (4/d)\mathrm{TV}(P, P_0)^2$, we have

**Upper bound**

$$\rho^*(\phi) \lesssim \frac{d^{1/4}}{n^{1/2}}$$

# Minimax upper bound

More generally, we can see that

$$\mathrm{Var}_P(T) \lesssim \frac{\mathbb{E}_P(T)}{n}(1/d + 1/n) + \frac{\{\mathbb{E}_P(T)\}^{3/2}}{n} + \frac{1}{dn^2}.$$

Thus, when $\mathbb{E}_P(T) = \|p - p_0\|_2^2 \geq 2C_{d,n} \asymp 1/(n\sqrt{d})$ we have

$$\frac{\mathrm{Var}_P(T)}{(\mathbb{E}_P T)^2} \lesssim \frac{\sqrt{d}}{n} + \frac{1}{\sqrt{d}} + \frac{d^{1/4}}{\sqrt{n}} + 1 \lesssim 1.$$

By Chebyshev's inequality, there exists universal constant $B > 0$ such that Type II error $\leq 1/8$ when $\|p - p_0\|_2^2 \geq B/(n\sqrt{d})$.

Since $\|p - p_0\|_2^2 \geq (4/d)\mathrm{TV}(P, P_0)^2$, we have

**Upper bound**

$$\rho^*(\phi) \lesssim \frac{d^{1/4}}{n^{1/2}}$$

# Minimax upper bound

More generally, we can see that

$$\mathrm{Var}_P(T) \lesssim \frac{\mathbb{E}_P(T)}{n}(1/d + 1/n) + \frac{\{\mathbb{E}_P(T)\}^{3/2}}{n} + \frac{1}{dn^2}.$$

Thus, when $\mathbb{E}_P(T) = \|p - p_0\|_2^2 \geq 2C_{d,n} \asymp 1/(n\sqrt{d})$ we have

$$\frac{\mathrm{Var}_P(T)}{(\mathbb{E}_P T)^2} \lesssim \frac{\sqrt{d}}{n} + \frac{1}{\sqrt{d}} + \frac{d^{1/4}}{\sqrt{n}} + 1 \lesssim 1.$$

By Chebyshev's inequality, there exists universal constant $B > 0$ such that Type II error $\leq 1/8$ when $\|p - p_0\|_2^2 \geq B/(n\sqrt{d})$.

Since $\|p - p_0\|_2^2 \geq (4/d)\mathrm{TV}(P, P_0)^2$, we have

**Upper bound**

$$\rho^*(\phi) \lesssim \frac{d^{1/4}}{n^{1/2}}$$

# Broad strategy

We proved an upper bound by choosing a test and analysing its risk:

- $U$-statistic unbiased estimator of $\|p - p_0\|_2^2$, with tractable variance;
- $\|p - p_0\|_2 \geq \|p - p_0\|_1$ so TV separation implies $L_2$ separation.

We prove lower bounds by constructing 'hard' instances of the problem, where all tests would fail.



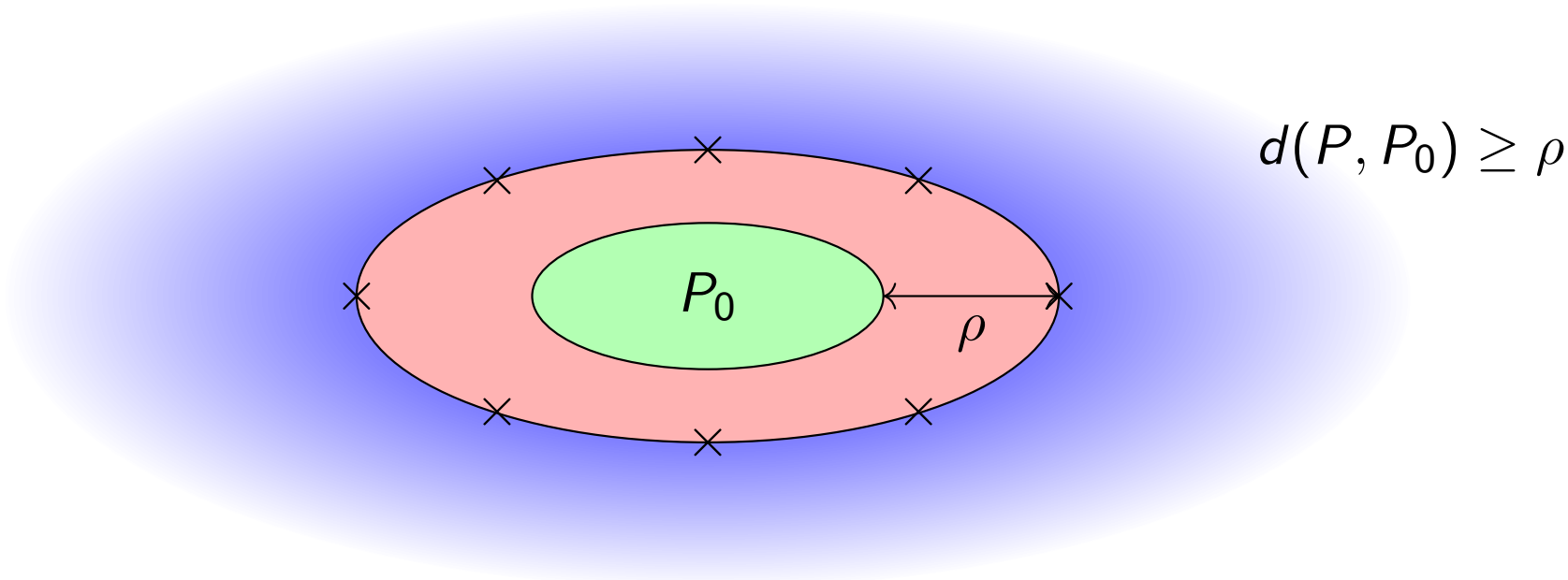

$d(P, P_0) \geq \rho$

# Broad strategy

We proved an upper bound by choosing a test and analysing its risk:

- $U$-statistic unbiased estimator of $\|p - p_0\|_2^2$, with tractable variance;
- $\|p - p_0\|_2 \geq \|p - p_0\|_1$ so TV separation implies $L_2$ separation.

We prove lower bounds by constructing 'hard' instances of the problem, where all tests would fail.



# Minimax lower bound

If $\pi$ is any (prior) distribution over $\mathcal{P}_1(\rho) = \{P : \mathrm{TV}(P, P_0) \geq \rho\}$, we have

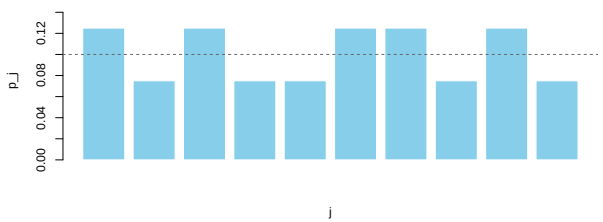
$$\inf_\phi \Big\{ \mathbb{E}_{P_0}(\phi) + \sup_{P \in \mathcal{P}_1(\rho)} \mathbb{E}_P(1-\phi) \Big\} \geq \inf_\phi \Big\{ \mathbb{E}_{P_0}(\phi) + \mathbb{E}_\pi \mathbb{E}_P(1-\phi) \Big\}$$
$$= 1 - \mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n})),$$

where $\mathbb{E}_\pi(P^{\otimes n})$ is the mixture distribution on $[d]^n$ with mass function

$$(x_1, \ldots, x_n) \mapsto \mathbb{E}_\pi \Big[ \prod_{i=1}^n p(x_i) \Big].$$

Construct $\pi$ (Paninski, 2008) by generating $\xi_1, \ldots, \xi_{d/2} \sim \mathrm{Rademacher}$, and setting

$$p(j) = \begin{cases} (1/d)(1 + 2\rho\xi_{j/2}) & j \text{ even} \\ (1/d)(1 - 2\rho\xi_{(j+1)/2}) & j \text{ odd} \end{cases}$$




By construction, $\mathrm{TV}(P, P_0) = (1/2)\|p - p_0\|_1 = \rho$ with prob. 1

# Minimax lower bound

If $\pi$ is any (prior) distribution over $\mathcal{P}_1(\rho) = \{P : \mathrm{TV}(P, P_0) \geq \rho\}$, we have

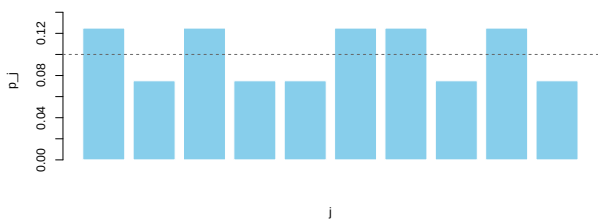
$$\inf_\phi \Big\{ \mathbb{E}_{P_0}(\phi) + \sup_{P \in \mathcal{P}_1(\rho)} \mathbb{E}_P(1-\phi) \Big\} \geq \inf_\phi \Big\{ \mathbb{E}_{P_0}(\phi) + \mathbb{E}_\pi \mathbb{E}_P(1-\phi) \Big\}$$
$$= 1 - \mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n})),$$

where $\mathbb{E}_\pi(P^{\otimes n})$ is the mixture distribution on $[d]^n$ with mass function

$$(x_1, \ldots, x_n) \mapsto \mathbb{E}_\pi \Big[ \prod_{i=1}^n p(x_i) \Big].$$

Construct $\pi$ (Paninski, 2008) by generating $\xi_1, \ldots, \xi_{d/2} \sim \mathrm{Rademacher}$, and setting

$$p(j) = \begin{cases} (1/d)(1 + 2\rho\xi_{j/2}) & j \text{ even} \\ (1/d)(1 - 2\rho\xi_{(j+1)/2}) & j \text{ odd} \end{cases}$$




By construction, $\mathrm{TV}(P, P_0) = (1/2)\|p - p_0\|_1 = \rho$ with prob. 1

# Minimax lower bound

Let $p'$ be an independent copy of $p$. Taking $\rho = \{\frac{d\log(1+\delta^2/4)}{16n^2}\}^{1/4}$ we have

$$1 + 4\mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))^2 \leq 1 + \chi^2(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))$$

$$= \sum_{x_1,\ldots,x_n} \frac{\{\mathbb{E}_\pi(\prod_{i=1}^n p(x_i))\}^2}{(1/d)^n} = \mathbb{E}_\pi\left[\sum_{x_1,\ldots,x_n} \frac{\prod_{i=1}^n \{p(x_i)p'(x_i)\}}{(1/d)^n}\right]$$

$$= \mathbb{E}_\pi\left[\left\{d\sum_{x=1}^d p(x)p'(x)\right\}^n\right] = \mathbb{E}_\pi\left[\left\{1 + \frac{8\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right\}^n\right]$$

$$\leq \mathbb{E}_\pi\left[\exp\left(\frac{8n\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right)\right] = \mathbb{E}_\pi^{d/2}\left[\exp\left(\frac{8n\rho^2}{d}\xi_1\right)\right]$$

$$= \cosh^{d/2}(8n\rho^2/d) \leq \exp\left(\frac{16n^2\rho^4}{d}\right) = 1 + 4\delta^2.$$

**Lower bound**

$$\rho^*(\phi) \geq \rho \gtrsim d^{1/4}/n^{1/2}$$

# Minimax lower bound

Let $p'$ be an independent copy of $p$. Taking $\rho = \{\frac{d \log(1+\delta^2/4)}{16n^2}\}^{1/4}$ we have

$$1 + 4\mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))^2 \leq 1 + \chi^2(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))$$

$$= \sum_{x_1,\ldots,x_n} \frac{\{\mathbb{E}_\pi(\prod_{i=1}^n p(x_i))\}^2}{(1/d)^n} = \mathbb{E}_\pi\left[\sum_{x_1,\ldots,x_n} \frac{\prod_{i=1}^n \{p(x_i)p'(x_i)\}}{(1/d)^n}\right]$$

$$= \mathbb{E}_\pi\left[\left\{d\sum_{x=1}^d p(x)p'(x)\right\}^n\right] = \mathbb{E}_\pi\left[\left\{1 + \frac{8\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right\}^n\right]$$

$$\leq \mathbb{E}_\pi\left[\exp\left(\frac{8n\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right)\right] = \mathbb{E}_\pi^{d/2}\left[\exp\left(\frac{8n\rho^2}{d}\xi_1\right)\right]$$

$$= \cosh^{d/2}(8n\rho^2/d) \leq \exp\left(\frac{16n^2\rho^4}{d}\right) = 1 + 4\delta^2.$$

**Lower bound**

$$\rho^*(\phi) \geq \rho \gtrsim d^{1/4}/n^{1/2}$$

# Minimax lower bound

Let $p'$ be an independent copy of $p$. Taking $\rho = \{\frac{d\log(1+\delta^2/4)}{16n^2}\}^{1/4}$ we have

$$1 + 4\mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))^2 \le 1 + \chi^2(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))$$

$$= \sum_{x_1,\ldots,x_n} \frac{\{\mathbb{E}_\pi(\prod_{i=1}^n p(x_i))\}^2}{(1/d)^n} = \mathbb{E}_\pi\left[\sum_{x_1,\ldots,x_n} \frac{\prod_{i=1}^n \{p(x_i)p'(x_i)\}}{(1/d)^n}\right]$$

$$= \mathbb{E}_\pi\left[\left\{d\sum_{x=1}^d p(x)p'(x)\right\}^n\right] = \mathbb{E}_\pi\left[\left\{1 + \frac{8\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right\}^n\right]$$

$$\le \mathbb{E}_\pi\left[\exp\left(\frac{8n\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right)\right] = \mathbb{E}_\pi^{d/2}\left[\exp\left(\frac{8n\rho^2}{d}\xi_1\right)\right]$$

$$= \cosh^{d/2}(8n\rho^2/d) \le \exp\left(\frac{16n^2\rho^4}{d}\right) = 1 + 4\delta^2.$$

**Lower bound**

$$\rho^*(\phi) \ge \rho \gtrsim d^{1/4}/n^{1/2}$$

# Minimax lower bound

Let $p'$ be an independent copy of $p$. Taking $\rho = \{\frac{d\log(1+\delta^2/4)}{16n^2}\}^{1/4}$ we have

$$1 + 4\mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))^2 \le 1 + \chi^2(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))$$

$$= \sum_{x_1,\ldots,x_n} \frac{\{\mathbb{E}_\pi(\prod_{i=1}^n p(x_i))\}^2}{(1/d)^n} = \mathbb{E}_\pi\left[\sum_{x_1,\ldots,x_n} \frac{\prod_{i=1}^n \{p(x_i)p'(x_i)\}}{(1/d)^n}\right]$$

$$= \mathbb{E}_\pi\left[\left\{d\sum_{x=1}^d p(x)p'(x)\right\}^n\right] = \mathbb{E}_\pi\left[\left\{1 + \frac{8\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right\}^n\right]$$

$$\le \mathbb{E}_\pi\left[\exp\left(\frac{8n\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right)\right] = \mathbb{E}_\pi^{d/2}\left[\exp\left(\frac{8n\rho^2}{d}\xi_1\right)\right]$$

$$= \cosh^{d/2}(8n\rho^2/d) \le \exp\left(\frac{16n^2\rho^4}{d}\right) = 1 + 4\delta^2.$$

**Lower bound**

$$\rho^*(\phi) \ge \rho \gtrsim d^{1/4}/n^{1/2}$$

# Minimax lower bound

Let $p'$ be an independent copy of $p$. Taking $\rho = \{\frac{d\log(1+\delta^2/4)}{16n^2}\}^{1/4}$ we have

$$
\begin{aligned}
1 + 4\mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))^2 &\le 1 + \chi^2(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n})) \\
&= \sum_{x_1,\ldots,x_n} \frac{\{\mathbb{E}_\pi(\prod_{i=1}^n p(x_i))\}^2}{(1/d)^n} = \mathbb{E}_\pi\left[\sum_{x_1,\ldots,x_n} \frac{\prod_{i=1}^n \{p(x_i)p'(x_i)\}}{(1/d)^n}\right] \\
&= \mathbb{E}_\pi\left[\left\{d\sum_{x=1}^d p(x)p'(x)\right\}^n\right] = \mathbb{E}_\pi\left[\left\{1 + \frac{8\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right\}^n\right] \\
&\le \mathbb{E}_\pi\left[\exp\left(\frac{8n\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right)\right] = \mathbb{E}_\pi^{d/2}\left[\exp\left(\frac{8n\rho^2}{d}\xi_1\right)\right] \\
&= \cosh^{d/2}(8n\rho^2/d) \le \exp\left(\frac{16n^2\rho^4}{d}\right) = 1 + 4\delta^2.
\end{aligned}
$$

**Lower bound**

$$\rho^*(\phi) \ge \rho \gtrsim d^{1/4}/n^{1/2}$$

# Minimax lower bound

Let $p'$ be an independent copy of $p$. Taking $\rho = \{\frac{d\log(1+\delta^2/4)}{16n^2}\}^{1/4}$ we have

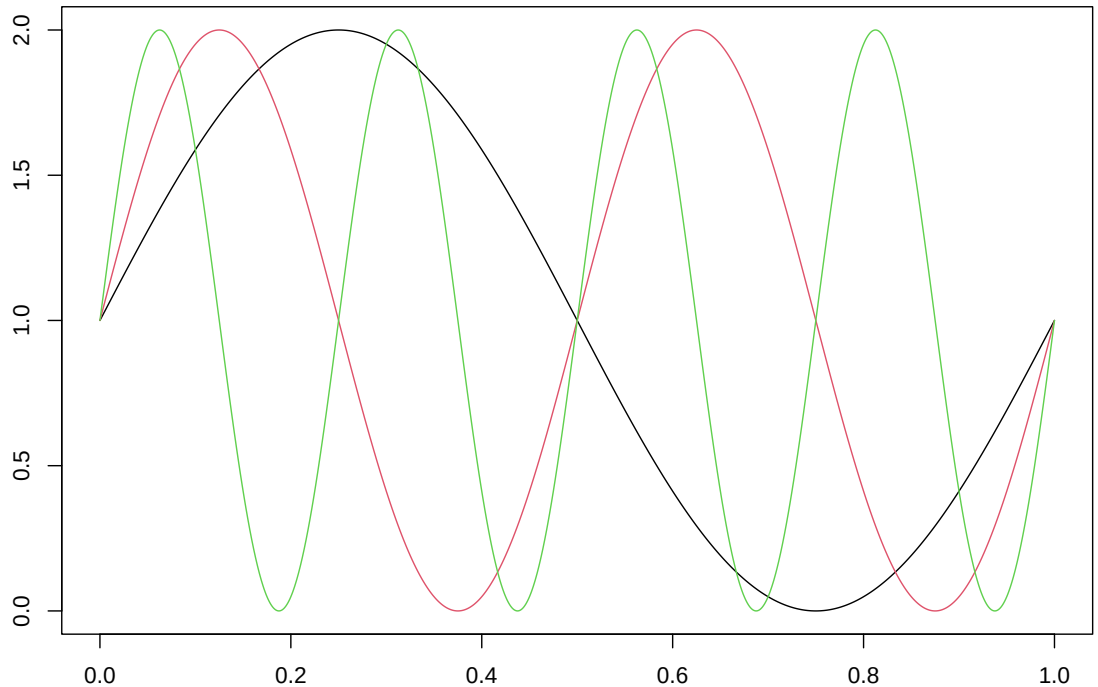
$$
\begin{aligned}
1 + 4\mathrm{TV}(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n}))^2 &\le 1 + \chi^2(P_0^{\otimes n}, \mathbb{E}_\pi(P^{\otimes n})) \\
&= \sum_{x_1,\ldots,x_n} \frac{\{\mathbb{E}_\pi(\prod_{i=1}^n p(x_i))\}^2}{(1/d)^n} = \mathbb{E}_\pi\left[\sum_{x_1,\ldots,x_n} \frac{\prod_{i=1}^n \{p(x_i)p'(x_i)\}}{(1/d)^n}\right] \\
&= \mathbb{E}_\pi\left[\left\{d\sum_{x=1}^d p(x)p'(x)\right\}^n\right] = \mathbb{E}_\pi\left[\left\{1 + \frac{8\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right\}^n\right] \\
&\le \mathbb{E}_\pi\left[\exp\left(\frac{8n\rho^2}{d}\sum_{j=1}^{d/2}\xi_j\xi_j'\right)\right] = \mathbb{E}_\pi^{d/2}\left[\exp\left(\frac{8n\rho^2}{d}\xi_1\right)\right] \\
&= \cosh^{d/2}(8n\rho^2/d) \le \exp\left(\frac{16n^2\rho^4}{d}\right) = 1 + 4\delta^2.
\end{aligned}
$$

**Lower bound**

$$\rho^*(\phi) \ge \rho \gtrsim d^{1/4}/n^{1/2}$$

# Continuous case

Similar ideas can be used to find optimal tests for continuous settings, though there is an important conceptual difference.

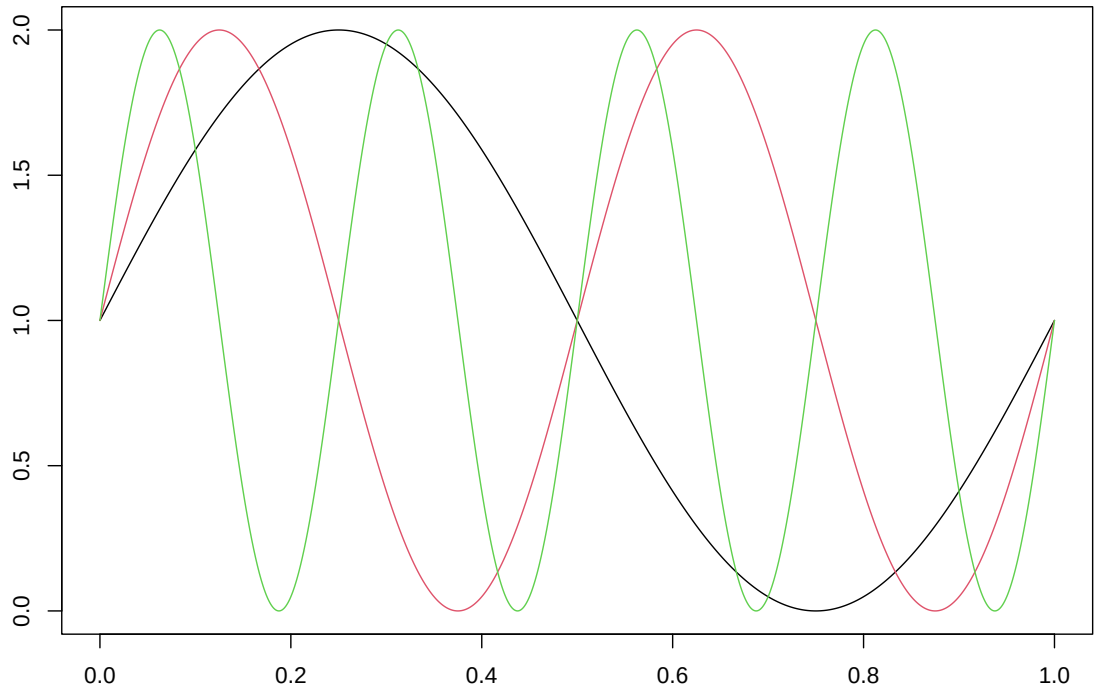



Even fixing the magnitude of departure (e.g. through $\|p - p_0\|_2$), there are infinitely many different ways $H_0$ can be violated.

Janssen (2000) confirms the impossibility of valid tests with uniform power against alternatives of the form $\{p : \|p - p_0\|_2 \geq \rho\}$.

# Continuous case

Similar ideas can be used to find optimal tests for continuous settings, though there is an important conceptual difference.




Even fixing the magnitude of departure (e.g. through $\|p - p_0\|_2$), there are infinitely many different ways $H_0$ can be violated.

Janssen (2000) confirms the impossibility of valid tests with uniform power against alternatives of the form $\{p : \|p - p_0\|_2 \geq \rho\}$.

# Continous case – smoothness

To give tests with uniform power guarantees we must impose structure on the alternative. It is common to make smoothness assumptions such as

$$|p(x) - p(y)| \leq L\|x - y\|_2^\beta \quad \text{for all } x, y \in [0,1]^d$$

for some $\beta \in (0,1], L > 0$. Write $p \in \mathcal{H}(\beta, L)$.

Our problem becomes: given $X_1, \ldots, X_n \sim P$ on $[0,1]^d$, test

$$H_0 : P = \mathrm{Unif}([0,1]^d) \quad \text{vs.} \quad H_1(\rho) : \|p - p_0\|_2 \geq \rho, p \in \mathcal{H}(\beta, L)$$

**Theorem (Arias-Castro et al. 2018)**

*In this setting, there exist constants $c, C > 0$ depending only on $d, \beta, L$ such that*

$$c\, n^{-\frac{2\beta}{4\beta+d}} \leq \rho^* \leq C\, n^{-\frac{2\beta}{4\beta+d}}.$$

Again, this rate is faster than the optimal estimation rate of $n^{-\frac{\beta}{2\beta+d}}$.

# Continous case – smoothness

To give tests with uniform power guarantees we must impose structure on the alternative. It is common to make smoothness assumptions such as

$$|p(x) - p(y)| \leq L\|x - y\|_2^\beta \quad \text{for all } x, y \in [0,1]^d$$

for some $\beta \in (0,1], L > 0$. Write $p \in \mathcal{H}(\beta, L)$.

Our problem becomes: given $X_1, \ldots, X_n \sim P$ on $[0,1]^d$, test

$$H_0 : P = \mathrm{Unif}([0,1]^d) \quad \text{vs.} \quad H_1(\rho) : \|p - p_0\|_2 \geq \rho, p \in \mathcal{H}(\beta, L)$$

**Theorem (Arias-Castro et al. 2018)**

*In this setting, there exist constants $c, C > 0$ depending only on $d, \beta, L$ such that*

$$c\, n^{-\frac{2\beta}{4\beta+d}} \leq \rho^* \leq C\, n^{-\frac{2\beta}{4\beta+d}}.$$

Again, this rate is faster than the optimal estimation rate of $n^{-\frac{\beta}{2\beta+d}}$.

# Instance optimality

So far we have considered GoF testing with uniform nulls. Recent work (Diakonikolas & Kane, 2016; Valiant & Valiant, 2017; Balakrishnan & Wasserman, 2019) has shown that the difficulty of the problem depends intricately on $P_0$.

**Theorem (Chhor & Carpentier 2022)**

*Assuming that $p_0(1) \geq p_0(2) \geq \ldots \geq p_0(d)$, we have*

$$\rho^* \asymp \sqrt{\frac{\|(p_0^{-\max})_{\leq I}\|_{2/3}}{n}} + \|p_{>I}\|_1 + \frac{1}{n}$$

*where $I = \min\{J : \sum_{i>J} p_0(i)^2 \leq cn^{-2}\}$ and $p_0^{-\max} = (p_0(2), \ldots, p_0(d))$.*

Roughly speaking, the difficulty is determined by the '2/3-norm' of $p_0$:

- $P_0 = \mathrm{Unif}([d])$ is the hardest problem, with $\frac{d^{1/4}}{n^{1/2}}$ a global upper bound
- highly structured distributions (e.g. a Dirac mass) are easier to test.

# Instance optimality

So far we have considered GoF testing with uniform nulls. Recent work (Diakonikolas & Kane, 2016; Valiant & Valiant, 2017; Balakrishnan & Wasserman, 2019) has shown that the difficulty of the problem depends intricately on $P_0$.

**Theorem (Chhor & Carpentier 2022)**

*Assuming that* $p_0(1) \geq p_0(2) \geq \ldots \geq p_0(d)$, *we have*

$$\rho^* \asymp \sqrt{\frac{\|(p_0^{-\max})_{\leq I}\|_{2/3}}{n}} + \|p_{>I}\|_1 + \frac{1}{n}$$

*where* $I = \min\{J : \sum_{i>J} p_0(i)^2 \leq cn^{-2}\}$ *and* $p_0^{-\max} = (p_0(2), \ldots, p_0(d))$.

Roughly speaking, the difficulty is determined by the '2/3-norm' of $p_0$:

- $P_0 = \mathrm{Unif}([d])$ is the hardest problem, with $\frac{d^{1/4}}{n^{1/2}}$ a global upper bound
- highly structured distributions (e.g. a Dirac mass) are easier to test.

# Goodness-of-fit testing – conclusion

- Separation rates give a convenient way to study the power of tests – how strong does a signal need to be before we can detect it?
- $U$-statistics are often a convenient basis for powerful tests.
- Lower bounds can be proved by constructing alternatives that are difficult to detect.

However, the picture is more complex than for point hypotheses:

- We typically do not know the exactly optimal test.
- In general (e.g. for continuous data) uniform power is impossible, and we must make (smoothness) assumptions.
- The choice of test may vary according to which assumptions we make and what separation metric we choose.

Here $H_0$ was simple, meaning that calibrating tests is (in principle) easy. Next, we will consider problems with composite nulls.

# Goodness-of-fit testing – conclusion

- Separation rates give a convenient way to study the power of tests – how strong does a signal need to be before we can detect it?
- $U$-statistics are often a convenient basis for powerful tests.
- Lower bounds can be proved by constructing alternatives that are difficult to detect.

However, the picture is more complex than for point hypotheses:

- We typically do not know the exactly optimal test.
- In general (e.g. for continuous data) uniform power is impossible, and we must make (smoothness) assumptions.
- The choice of test may vary according to which assumptions we make and what separation metric we choose.

Here $H_0$ was simple, meaning that calibrating tests is (in principle) easy. Next, we will consider problems with composite nulls.

# Goodness-of-fit testing – conclusion

- Separation rates give a convenient way to study the power of tests – how strong does a signal need to be before we can detect it?
- $U$-statistics are often a convenient basis for powerful tests.
- Lower bounds can be proved by constructing alternatives that are difficult to detect.

However, the picture is more complex than for point hypotheses:

- We typically do not know the exactly optimal test.
- In general (e.g. for continuous data) uniform power is impossible, and we must make (smoothness) assumptions.
- The choice of test may vary according to which assumptions we make and what separation metric we choose.

Here $H_0$ was simple, meaning that calibrating tests is (in principle) easy. Next, we will consider problems with composite nulls.

1. Introduction

2. Simple null hypotheses
   - The likelihood ratio test and Neyman–Pearson
   - Composite alternatives: Goodness-of-fit testing

3. Permutation testing for composite nulls
   - Independence testing – minimax optimality
   - Conditional independence testing – non-uniform permutations

4. Local differential privacy
   - Two-point testing
   - Goodness-of-fit testing

# Composite null hypotheses

So far we have seen how to construct powerful tests in settings where the null hypothesis is *simple*. Given $T$ we can find $C_\alpha$ such that

$$\mathbb{P}_{H_0}(T \geq C_\alpha) \leq \alpha.$$

When $H_0 : P \in \mathcal{P}_0$ is *composite*, it is harder to guarantee validity. We will say that a test $\phi$ is *uniformly valid* if

$$\sup_{P \in \mathcal{P}_0} \mathbb{E}_P(\phi) \leq \alpha.$$

Two canonical problems are:

**Two-sample testing**

Given $X_1, \ldots, X_n \sim f$ and $Y_1, \ldots, Y_m \sim g$, we test $H_0 : f = g$.

**Independence testing**

Given $(X_1, Y_1), \ldots, (X_n, Y_n) \sim f$ we test $H_0 : f(x, y) = f_X(x) f_Y(y)\ \forall x, y$.

# Independence testing

Independence testing is one of the most studied problems in statistics. With discrete data we usually see a contingency table, e.g.

| | Middle school or lower | High school | Bachelor's | Master's | PhD or higher |
|---|---|---|---|---|---|
| Never married | 18 | 36 | 21 | 9 | 6 |
| Married | 12 | 36 | 45 | 36 | 21 |
| Divorced | 6 | 9 | 9 | 3 | 3 |
| Widowed | 3 | 9 | 9 | 6 | 3 |

Source: https://www.spss-tutorials.com/chi-square-independence-test/.

Writing $p_{jk} = \mathbb{P}(X = j, Y = k)$, we test

$$H_0 : p_{jk} = r_j q_k \quad \text{for some } (r_j), (q_k).$$

# The $\chi^2$ test

It is standard practice to use Pearson's $\chi^2$ test, where we use the fact that

$$T = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(o_{jk} - e_{jk})^2}{e_{jk}} \xrightarrow{\mathrm{d}} \chi^2_{(J-1)(K-1)}$$

whenever $H_0$ holds. We have a test $\phi = \mathbb{1}\{T \geq (\chi^2_{(J-1)(K-1)})^{-1}(1-\alpha)\}$ with *pointwise asymptotic validity*: $\limsup_n \mathbb{E}_P(\phi) \leq \alpha$ for each $P \in \mathcal{P}_0$.

| $j/k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | $o_{11}$ | $o_{12}$ | $o_{13}$ | $o_{14}$ | $o_{15}$ |
| 2 | $o_{21}$ | $o_{22}$ | $o_{23}$ | $o_{24}$ | $o_{25}$ |
| 3 | $o_{31}$ | $o_{32}$ | $o_{33}$ | $o_{34}$ | $o_{35}$ |
| 4 | $o_{41}$ | $o_{42}$ | $o_{43}$ | $o_{44}$ | $o_{45}$ |

$$e_{jk} = \frac{o_{j+} \cdot o_{+k}}{n}$$

This convergence is not uniform and validity at finite sample sizes is not guaranteed.

# The $\chi^2$ test

It is standard practice to use Pearson's $\chi^2$ test, where we use the fact that

$$T = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(o_{jk} - e_{jk})^2}{e_{jk}} \xrightarrow{\mathrm{d}} \chi^2_{(J-1)(K-1)}$$

whenever $H_0$ holds. We have a test $\phi = \mathbb{1}\{T \geq (\chi^2_{(J-1)(K-1)})^{-1}(1-\alpha)\}$ with *pointwise asymptotic validity*: $\limsup_n \mathbb{E}_P(\phi) \leq \alpha$ for each $P \in \mathcal{P}_0$.

| $j/k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | $o_{11}$ | $o_{12}$ | $o_{13}$ | $o_{14}$ | $o_{15}$ |
| 2 | $o_{21}$ | $o_{22}$ | $o_{23}$ | $o_{24}$ | $o_{25}$ |
| 3 | $o_{31}$ | $o_{32}$ | $o_{33}$ | $o_{34}$ | $o_{35}$ |
| 4 | $o_{41}$ | $o_{42}$ | $o_{43}$ | $o_{44}$ | $o_{45}$ |

$$e_{jk} = \frac{o_{j+} \cdot o_{+k}}{n}$$

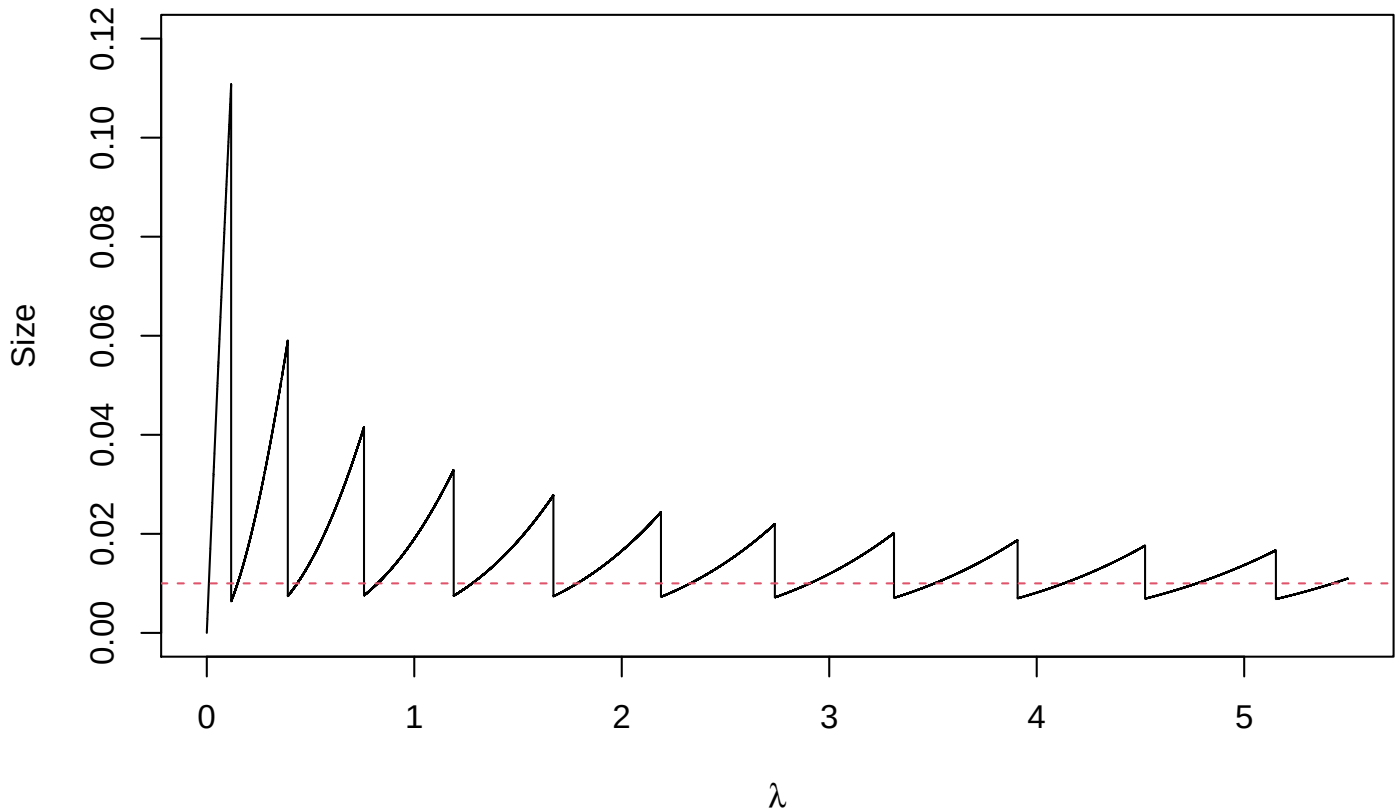
This convergence is not uniform and validity at finite sample sizes is not guaranteed.

# Asymptotic null distributions

For fixed $\lambda > 0$ set $p = \sqrt{\lambda/n}$ and consider the null distribution given by:

| $j/k$ | 1 | 2 |
|---|---|---|
| 1 | $p^2$ | $p(1-p)$ |
| 2 | $p(1-p)$ | $(1-p)^2$ |

Comparing $T$ to the 99th quantile of the $\chi^2_1$ distribution, for large $n$ we get



# Permutation tests

The issue is exacerbated in more complex problems, where

- asymptotic approximations may not be known; or
- rates of convergence depend on unverifiable smoothness assumptions.

Fortunately, for some problems permutation tests exist and give exact Type I error guarantees.

# Exact tests

Permutation testing is a classical idea (Fisher, 1935; Pitman, 1938). The simplest example is the *Lady tasting tea* experiment, using an *exact test*.


Figure: Source: e10v.me/tea-tasting-rdd/

Can the taster tell whether milk or tea was first? We can calculate the exact significance of any outcome.

# Exact tests

Thought of as an independence test: data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \{0, 1\}^2$

| $X/Y$ | Guess tea first | Guess milk first | Total |
|---|---|---|---|
| Tea first | $n_{11}$ | $n_{10}$ | $n_{1+}$ |
| Milk first | $n_{01}$ | $n_{00}$ | $n_{0+}$ |
| Total | $n_{+1}$ | $n_{+0}$ | n |

Permutations $(X_{\pi(1)}, Y_1), \ldots, (X_{\pi(n)}, Y_n)$ preserve the row/column totals and lead to $n_{11}^{\pi}$ with a Hypergeometric distribution.

Compare $n_{11}$ to $(n_{11}^{\pi} : \pi \in \mathcal{S}_n)$, rejecting $H_0 : X \perp\!\!\!\perp Y$ if it is sufficiently extreme.

$$p = (n!)^{-1} \sum_{\pi \in \mathcal{S}_n} \mathbb{1}_{\{n_{11}^{\pi} \geq n_{11}\}}.$$

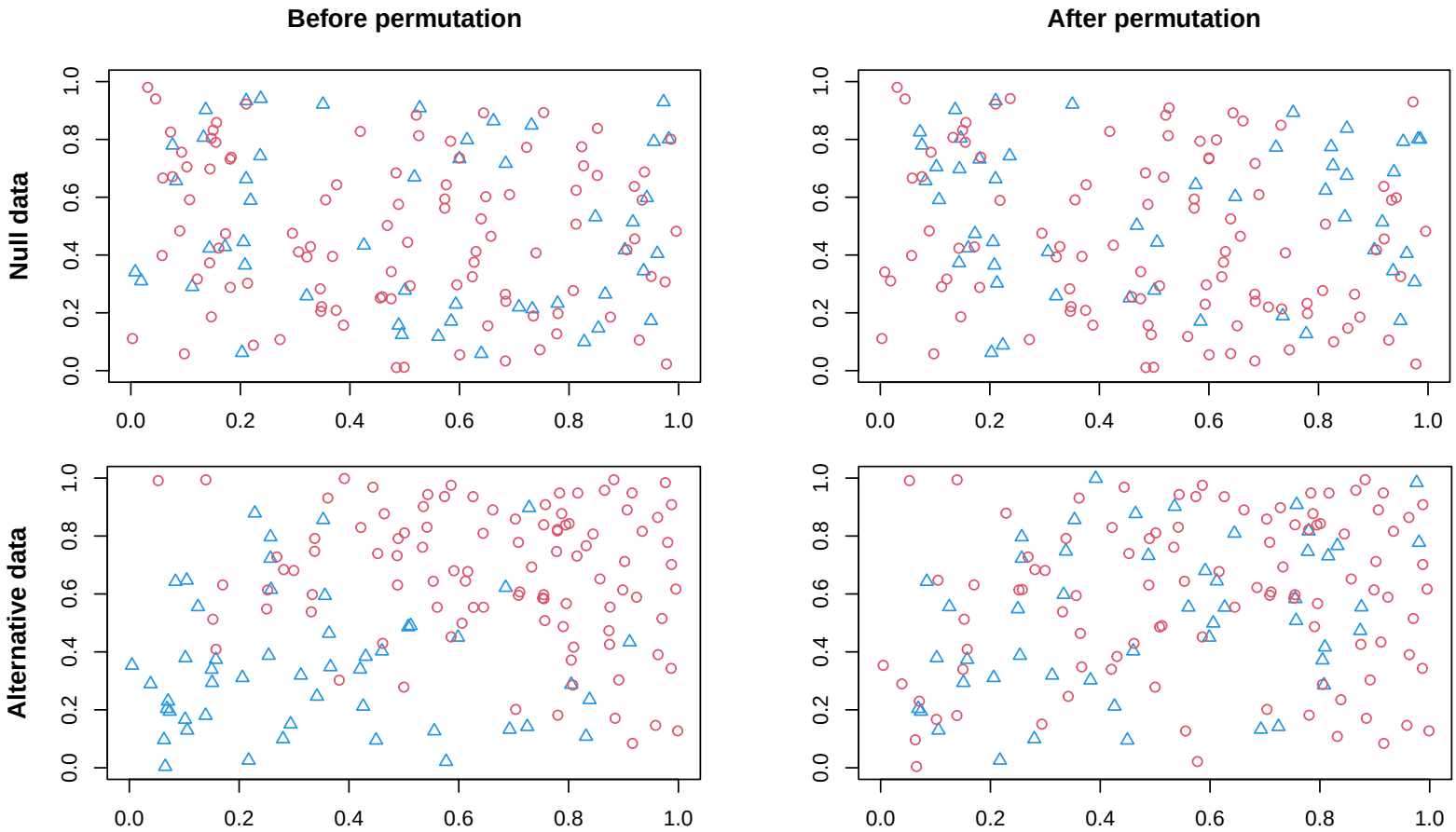
Strong Type I error guarantees: $\mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$.

# Beyond discrete data – Two-sample testing

## Two-sample testing

Given $X_1, \ldots, X_n \sim f$ and $Y_1, \ldots, Y_m \sim g$, we test $H_0 : f = g$.

$$\mathbf{Z} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m), \quad \mathbf{Z}_\pi = (Z_{\pi(1)}, \ldots, Z_{\pi(m+n)}) \text{ for } \pi \in \mathcal{S}_{n+m}.$$

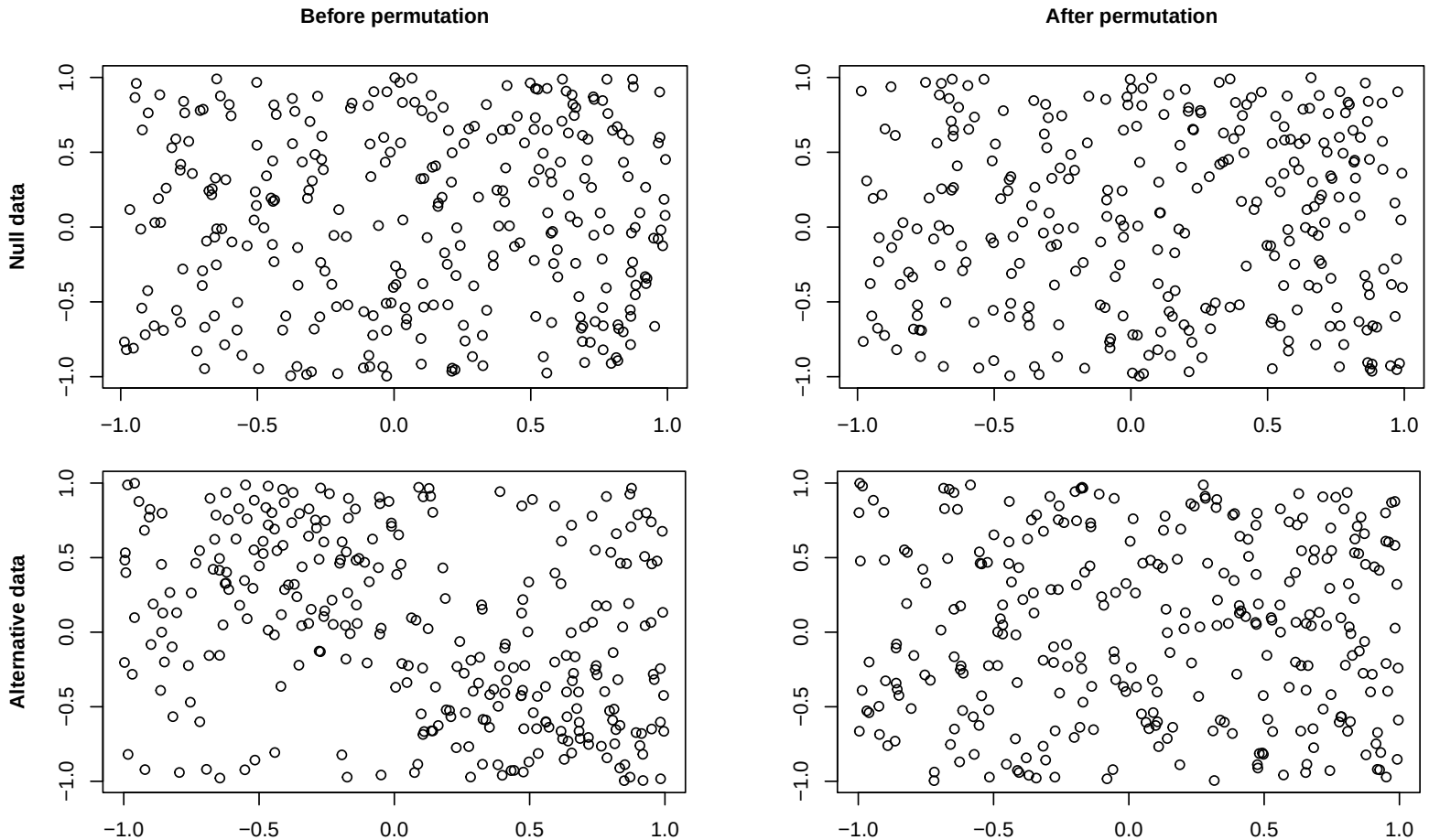


# Independence testing

## Independence testing

Given $(X_1, Y_1), \ldots, (X_n, Y_n) \sim f$ we test $H_0 : f(x, y) = f_X(x) f_Y(y)\ \forall x, y$.

$$\mathbf{Z} = (X_1, Y_1, \ldots, X_n, Y_n), \quad \mathbf{Z}_\pi = (X_{\pi(1)}, Y_1, \ldots, X_{\pi(n)}, Y_n) \text{ for } \pi \in \mathcal{S}_n.$$



# Type I error control

Typically infeasible to use all permutations, and permutation distributions (e.g. Hypergeometric) are not known, so we sample uniformly.

For any test statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ and $\Pi_1, \ldots, \Pi_B \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{S}_n)$ we calculate the p-value

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{1}\{T(\mathbf{Z}) \leq T(\mathbf{Z}_{\Pi_b})\}}{1 + B} = \frac{\text{rank}(T(\mathbf{Z}))}{1 + B}.$$

**Theorem**

*Suppose that $H_0$ holds. Then*

$$\mathbb{P}(p \leq \alpha) \leq \frac{\lfloor \alpha(B+1) \rfloor}{B+1} \leq \alpha$$

*for all $\alpha \in [0, 1]$.*

# Type I error control

Typically infeasible to use all permutations, and permutation distributions (e.g. Hypergeometric) are not known, so we sample uniformly.

For any test statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ and $\Pi_1, \ldots, \Pi_B \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{S}_n)$ we calculate the p-value

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{1}\{T(\mathbf{Z}) \leq T(\mathbf{Z}_{\Pi_b})\}}{1 + B} = \frac{\text{rank}(T(\mathbf{Z}))}{1 + B}.$$

**Theorem**

*Suppose that $H_0$ holds. Then*

$$\mathbb{P}(p \leq \alpha) \leq \frac{\lfloor \alpha(B+1) \rfloor}{B+1} \leq \alpha$$

*for all $\alpha \in [0, 1]$.*

# Type I error control – proof

Under $H_0$, if we condition on the multi-set $\mathcal{X}_n := \{X_1, \ldots, X_n\}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ then $\mathbf{Z} = (X_1, Y_1, \ldots, X_n, Y_n)$ is a uniform assignment.

So, given $\mathcal{X}_n$ and $\mathbf{Y}$, the datasets $\mathbf{Z}, \mathbf{Z}_{\Pi_1}, \ldots, \mathbf{Z}_{\Pi_B}$ are conditionally i.i.d.

So $\mathbf{Z}, \mathbf{Z}_{\Pi_1}, \ldots, \mathbf{Z}_{\Pi_B}$ are exchangeable.

So

$$
\begin{aligned}
&(B+1)\mathbb{P}(p \leq \alpha) = (B+1)\mathbb{P}(\mathrm{rank}(T(\mathbf{Z})) \leq \alpha(B+1)) \\
&= \mathbb{E}\left[\mathbb{1}\{\mathrm{rank}(T(\mathbf{Z}) \leq \alpha(B+1)\} + \sum_{b=1}^{B} \mathbb{1}\{\mathrm{rank}(T(\mathbf{Z}_{\Pi_b}) \leq \alpha(B+1)\}\right] \\
&\leq \lfloor \alpha(B+1) \rfloor.
\end{aligned}
$$

# Asymptotic power

While Type I error control is simple, power results are difficult to prove.

**Theorem (Hoeffding 1952)**

*Say $(X_1, Y_1) \sim P_n$ for some sequence $(P_n)$. Suppose that*

$$\mathbb{P}(T(\mathbf{Z}_{\Pi_1}) \leq u, T(\mathbf{Z}_{\Pi_2}) \leq v) \to R(u)R(v)$$

*and $\mathbb{P}(T(\mathbf{Z}) \leq u) \to H(u)$ for distribution functions $H, R$. If $H$ and $R$ are continuous at $R^{-1}(1-\alpha)$ and $R$ is strictly increasing at $R^{-1}(1-\alpha)$, then, if $B \to \infty$, we have*

$$\mathbb{P}(p \leq \alpha) \to 1 - H\big(R^{-1}(1-\alpha)\big)$$

*as $n \to \infty$.*

Used for results on power against sequences $(P_n)$ of local alternatives.

# Asymptotic power

While Type I error control is simple, power results are difficult to prove.

**Theorem (Hoeffding 1952)**

*Say $(X_1, Y_1) \sim P_n$ for some sequence $(P_n)$. Suppose that*

$$\mathbb{P}(T(\mathbf{Z}_{\Pi_1}) \leq u, T(\mathbf{Z}_{\Pi_2}) \leq v) \to R(u)R(v)$$

*and $\mathbb{P}(T(\mathbf{Z}) \leq u) \to H(u)$ for distribution functions $H, R$. If $H$ and $R$ are continuous at $R^{-1}(1-\alpha)$ and $R$ is strictly increasing at $R^{-1}(1-\alpha)$, then, if $B \to \infty$, we have*

$$\mathbb{P}(p \leq \alpha) \to 1 - H\big(R^{-1}(1-\alpha)\big)$$

*as $n \to \infty$.*

Used for results on power against sequences $(P_n)$ of local alternatives.

# Permutation tests

Permutation tests:

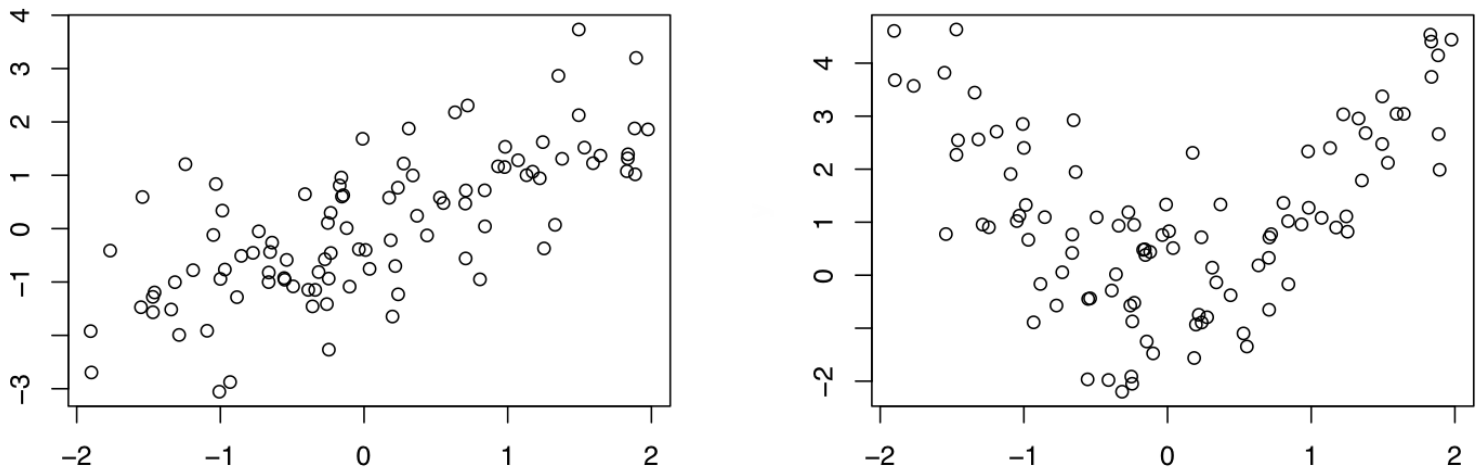
- require no assumptions for non-asymptotic Type I error control;
- can be used with any test statistic.

In the rest of the lecture we will see that they:

- can achieve minimax rate optimal power
- can be adapted for problems with non-exchangeable null hypotheses.

# Independence testing

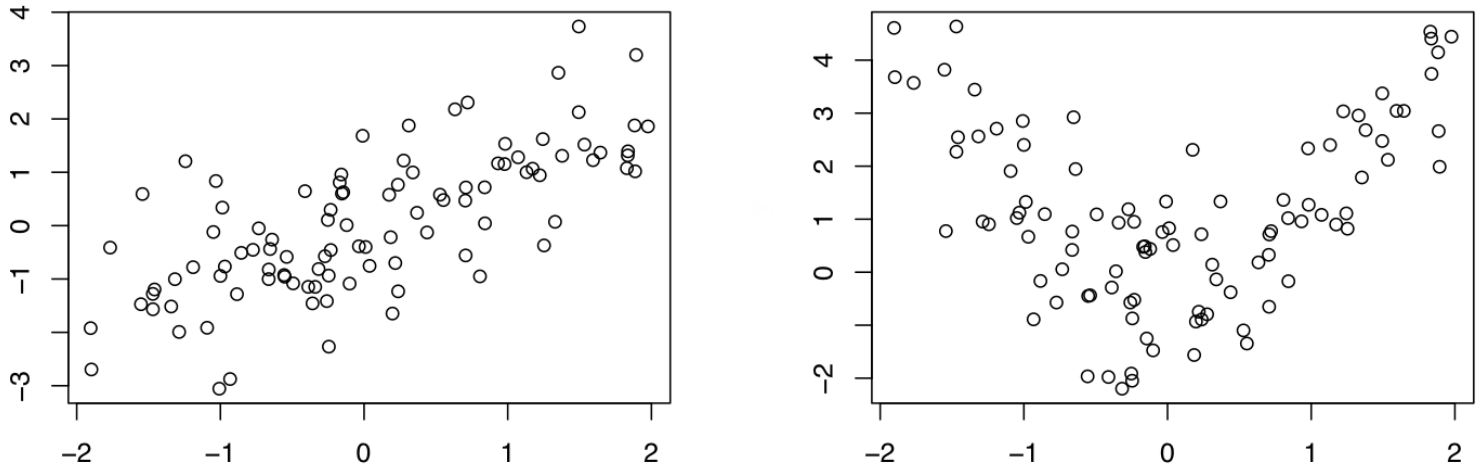
There is a vast literature on testing independence of continuous variables, with classical measure such as Pearson's correlation (e.g. Pearson, 1920), Kendall's tau (Kendall, 1938), Hoeffding's D (Hoeffding, 1948). These are limited to linear or monotonic dependence, or bivariate settings.




Modern datasets exhibit complex dependence not captured by classical measures. More flexible measures and tests include HSIC (Gretton et al., 2005), Distance covariance (Székely, Rizzo & Bakirov, 2007), mutual information (B. & Samworth, 2019), multivariate rank-based tests (Deb & Sen, 2023; Chatterjee, 2021), etc...

# Independence testing

There is a vast literature on testing independence of continuous variables, with classical measure such as Pearson's correlation (e.g. Pearson, 1920), Kendall's tau (Kendall, 1938), Hoeffding's D (Hoeffding, 1948). These are limited to linear or monotonic dependence, or bivariate settings.



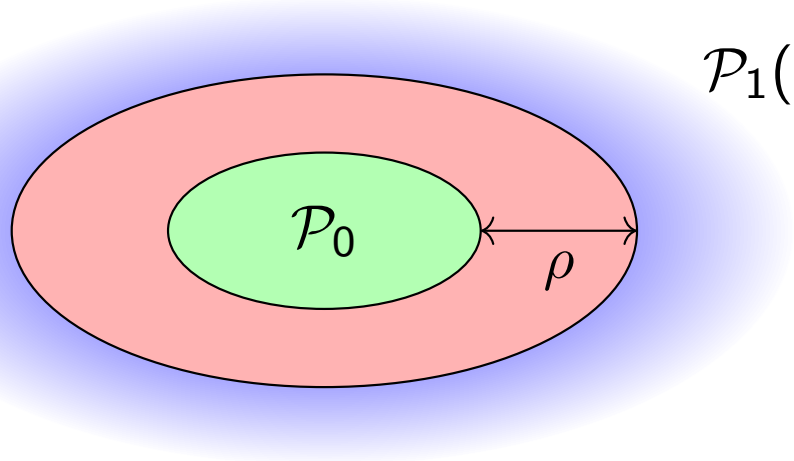

Modern datasets exhibit complex dependence not captured by classical measures. More flexible measures and tests include HSIC (Gretton et al., 2005), Distance covariance (Székely, Rizzo & Bakirov, 2007), mutual information (B. & Samworth, 2019), multivariate rank-based tests (Deb & Sen, 2023; Chatterjee, 2021), etc...

# Independence testing – problem statement

## Independence testing

Given $(X_1, Y_1), \ldots, (X_n, Y_n) \sim f$ we test $H_0 : f(x, y) = f_X(x) f_Y(y)\ \forall x, y$.

Given a suitable (e.g. smooth) class $\mathcal{P}_1$, define the alternatives




$$\mathcal{P}_1(\rho) = \left\{ P \in \mathcal{P}_1 : D(P) \geq \rho^2 \right\}$$

$$D(f) = \int (f - f_X f_Y)^2$$

For a test $\phi$, we define the worst-case risk

$$\mathcal{R}_n(\mathcal{P}_1(\rho), \phi) = \sup_{P \in \mathcal{P}_0} \mathbb{E}_P(\phi) + \sup_{P \in \mathcal{P}_1(\rho)} \mathbb{E}_P(1 - \phi),$$

and the associated separation radius $\rho_n^*(\mathcal{P}_1, \phi)$.

# Continuous case – basis expansion

Suppose $\mathcal{X} \times \mathcal{Y} = [0,1]^{d_X + d_Y}$. As we mentioned before, we will need smoothness assumptions for uniform power.

Let $(p_{jk})_{j \in \mathcal{J}, k \in \mathcal{K}} = (p_j^X p_k^Y)_{j \in \mathbb{N}^{d_X}, k \in \mathbb{N}^{d_Y}}$ be the Fourier basis and for density $f$ define

$$a_{jk} = \int p_{jk} f \, d\mu, \quad a_{j\bullet} = \int p_j^X f_X \, d\mu_X, \quad a_{\bullet k} = \int p_k^Y f_Y \, d\mu_Y$$

Define the Sobolev smoothness of $f - f_X f_Y$ by

$$S_{s_X, s_Y}(f) = \sum_{j \in \mathbb{N}^{d_X}, k \in \mathbb{N}^{d_Y}} (\|j\|_1^{2s_X} \vee \|k\|_1^{2s_Y}) \{a_{jk}(f) - a_{j\bullet}(f) a_{\bullet k}(f)\}^2.$$

We will consider classes of the form

$$\mathcal{P}_1 = \left\{ P : S_{s_X, s_Y}(f) \leq r^2, \max(\|f\|_\infty, \|f_X\|_\infty, \|f_Y\|_\infty) \leq A \right\}.$$

# Continuous case – basis expansion

Suppose $\mathcal{X} \times \mathcal{Y} = [0,1]^{d_X + d_Y}$. As we mentioned before, we will need smoothness assumptions for uniform power.

Let $(p_{jk})_{j \in \mathcal{J}, k \in \mathcal{K}} = (p_j^X p_k^Y)_{j \in \mathbb{N}^{d_X}, k \in \mathbb{N}^{d_Y}}$ be the Fourier basis and for density $f$ define

$$a_{jk} = \int p_{jk} f \, d\mu, \quad a_{j\bullet} = \int p_j^X f_X \, d\mu_X, \quad a_{\bullet k} = \int p_k^Y f_Y \, d\mu_Y$$

Define the Sobolev smoothness of $f - f_X f_Y$ by

$$S_{s_X, s_Y}(f) = \sum_{j \in \mathbb{N}^{d_X}, k \in \mathbb{N}^{d_Y}} (\|j\|_1^{2s_X} \vee \|k\|_1^{2s_Y}) \{a_{jk}(f) - a_{j\bullet}(f) a_{\bullet k}(f)\}^2.$$

We will consider classes of the form

$$\mathcal{P}_1 = \left\{ P : S_{s_X, s_Y}(f) \leq r^2, \max(\|f\|_\infty, \|f_X\|_\infty, \|f_Y\|_\infty) \leq A \right\}.$$

# Continuous case – minimax rates

Suppose $\mathcal{X} \times \mathcal{Y} = [0,1]^{d_X + d_Y}$ with $(s_X, s_Y)$-Sobolev smoothness.

**Theorem (B., Kontoyiannis, & Samworth 2021)**

*Fix $\gamma \in (0,1)$. Writing $d = d_X + d_Y$ and $s = d/(d_X/s_X + d_Y/s_Y)$ we have*

$$\inf_{\phi} \rho_n^*(\mathcal{P}_1, \phi) \asymp n^{-2s/(4s+d)}.$$

*Moreover, this rate is attained by a permutation test.*

- Similar rates for univariate problems known since Ingster (1989).
- In the multivariate setting, Albert et al. (2022) finds minimax rates with a test using an oracle critical value.
- In independent work, Kim, Balakrishnan & Wasserman (2022) proved the minimax rate optimality of HSIC-based permutation tests. Also consider two-sample testing.

# Upper bound

As for goodness-of-fit testing, $U$-statistics are useful.

For a subset $\mathcal{M} \subseteq \mathbb{N}^{d_X+d_Y}$ we write

$$h\big((x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\big)$$
$$= \sum_{(j,k)\in\mathcal{M}} \{p_{jk}(x_1,y_1)p_{jk}(x_2,y_2) - 2p_{jk}(x_1,y_1)p_{jk}(x_2,y_3) + p_{jk}(x_1,y_2)p_{jk}(x_3,y_4)\}$$

and take test statistic

$$\hat{D}_n = \frac{1}{4!\binom{n}{4}} \sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{distinct}}} h\big((X_{i_1},Y_{i_1}),(X_{i_2},Y_{i_2}),(X_{i_3},Y_{i_3}),(X_{i_4},Y_{i_4})\big).$$

This is an unbiased estimator of a truncated version of $D(f)$:

$$\mathbb{E}(\hat{D}_n) = \sum_{(j,k)\in\mathcal{M}} (a_{jk}^2 - 2a_{jk}a_{j\bullet}a_{\bullet k} + a_{j\bullet}^2 a_{\bullet k}^2) = \sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2.$$

# Upper bound

As for goodness-of-fit testing, $U$-statistics are useful.

For a subset $\mathcal{M} \subseteq \mathbb{N}^{d_X+d_Y}$ we write

$$h\big((x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\big)$$
$$= \sum_{(j,k)\in\mathcal{M}} \{p_{jk}(x_1,y_1)p_{jk}(x_2,y_2) - 2p_{jk}(x_1,y_1)p_{jk}(x_2,y_3) + p_{jk}(x_1,y_2)p_{jk}(x_3,y_4)\}$$

and take test statistic

$$\hat{D}_n = \frac{1}{4!\binom{n}{4}} \sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{distinct}}} h\big((X_{i_1},Y_{i_1}),(X_{i_2},Y_{i_2}),(X_{i_3},Y_{i_3}),(X_{i_4},Y_{i_4})\big).$$

This is an unbiased estimator of a truncated version of $D(f)$:

$$\mathbb{E}(\hat{D}_n) = \sum_{(j,k)\in\mathcal{M}} (a_{jk}^2 - 2a_{jk}a_{j\bullet}a_{\bullet k} + a_{j\bullet}^2 a_{\bullet k}^2) = \sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2.$$

# Upper bound

As for goodness-of-fit testing, $U$-statistics are useful.

For a subset $\mathcal{M} \subseteq \mathbb{N}^{d_X+d_Y}$ we write

$$\begin{aligned} &h\big((x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\big) \\ &= \sum_{(j,k)\in\mathcal{M}} \{p_{jk}(x_1,y_1)p_{jk}(x_2,y_2) - 2p_{jk}(x_1,y_1)p_{jk}(x_2,y_3) \\ &\qquad\qquad + p_{jk}(x_1,y_2)p_{jk}(x_3,y_4)\} \end{aligned}$$

and take test statistic

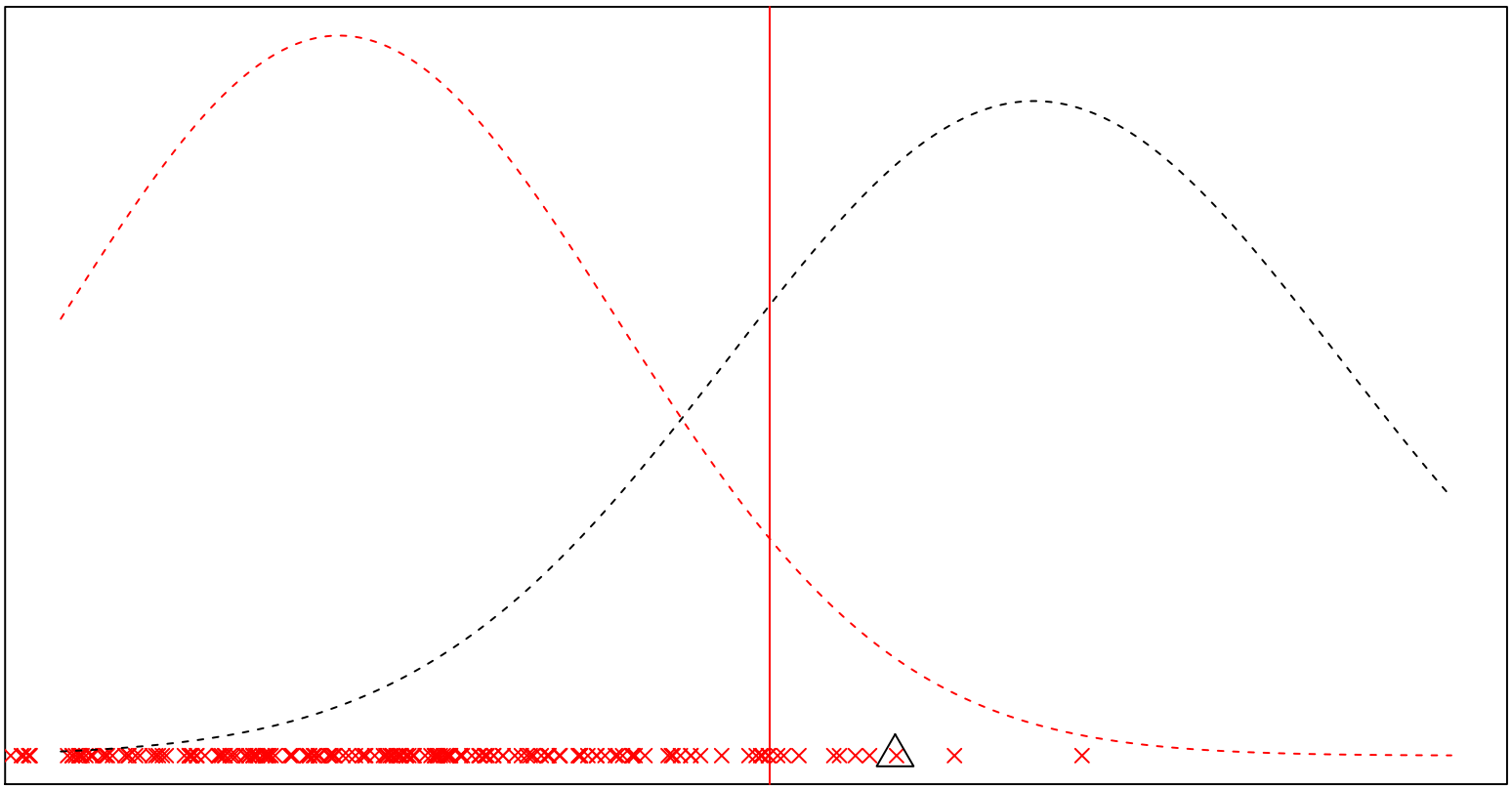
$$\hat{D}_n = \frac{1}{4!\binom{n}{4}} \sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{distinct}}} h\big((X_{i_1},Y_{i_1}),(X_{i_2},Y_{i_2}),(X_{i_3},Y_{i_3}),(X_{i_4},Y_{i_4})\big).$$

This is an unbiased estimator of a truncated version of $D(f)$:

$$\mathbb{E}(\hat{D}_n) = \sum_{(j,k)\in\mathcal{M}} (a_{jk}^2 - 2a_{jk}a_{j\bullet}a_{\bullet k} + a_{j\bullet}^2 a_{\bullet k}^2) = \sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2.$$

# Upper bound

To analyse the power of the permutation test, it suffices to control the first and second moments of $\hat{D}_n$ and its permuted versions $\hat{D}_n^{(1)}, \ldots, \hat{D}_n^{(B)}$.



Can then use Chebyshev's inequality to show that $\hat{D}_n > \hat{D}_n^{(1)}$ with high probability.

# Upper bound – original statistic

By construction, we have $\mathbb{E}_f(\hat{D}_n) = \sum_{(j,k)\in\mathcal{M}}(a_{jk} - a_{j\bullet}a_{\bullet k})^2$. Bias is due to truncation to $\mathcal{M}$. For suitable choices we have

$$|\mathbb{E}_f(\hat{D}) - D(f)| = \sum_{(j,k)\notin\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2$$
$$\leq \frac{r^2}{\inf\{\|j\|_1^{2s_X} \vee \|k\|_1^{2s_Y} : (j,k) \notin \mathcal{M}\}} \lesssim \frac{r^2}{|\mathcal{M}|^{2s/d}}.$$

Variance can be controlled using $U$-statistic results (Serfling, 2009), and

$$\mathrm{Var}_f(\hat{D}_n) \lesssim \frac{D(f)}{n} + \frac{|\mathcal{M}|}{n^2}.$$

As for goodness-of-fit testing, our statistic converges quicker under the null hypothesis with a rate that depends on the problem complexity.

Choosing $|\mathcal{M}| \asymp_r n^{\frac{2d}{4s+d}}$ gives convergence at rate $n^{-\frac{4s}{4s+d}}$.

# Upper bound – original statistic

By construction, we have $\mathbb{E}_f(\hat{D}_n) = \sum_{(j,k)\in\mathcal{M}}(a_{jk} - a_{j\bullet}a_{\bullet k})^2$. Bias is due to truncation to $\mathcal{M}$. For suitable choices we have

$$|\mathbb{E}_f(\hat{D}) - D(f)| = \sum_{(j,k)\notin\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2$$
$$\leq \frac{r^2}{\inf\{\|j\|_1^{2s_X} \vee \|k\|_1^{2s_Y} : (j,k) \notin \mathcal{M}\}} \lesssim \frac{r^2}{|\mathcal{M}|^{2s/d}}.$$

Variance can be controlled using $U$-statistic results (Serfling, 2009), and

$$\mathrm{Var}_f(\hat{D}_n) \lesssim \frac{D(f)}{n} + \frac{|\mathcal{M}|}{n^2}.$$

As for goodness-of-fit testing, our statistic converges quicker under the null hypothesis with a rate that depends on the problem complexity.

Choosing $|\mathcal{M}| \asymp_r n^{\frac{2d}{4s+d}}$ gives convergence at rate $n^{-\frac{4s}{4s+d}}$.

# Upper bound – original statistic

By construction, we have $\mathbb{E}_f(\hat{D}_n) = \sum_{(j,k)\in\mathcal{M}}(a_{jk} - a_{j\bullet}a_{\bullet k})^2$. Bias is due to truncation to $\mathcal{M}$. For suitable choices we have

$$|\mathbb{E}_f(\hat{D}) - D(f)| = \sum_{(j,k)\notin\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2$$
$$\leq \frac{r^2}{\inf\{\|j\|_1^{2s_X} \vee \|k\|_1^{2s_Y} : (j,k) \notin \mathcal{M}\}} \lesssim \frac{r^2}{|\mathcal{M}|^{2s/d}}.$$

Variance can be controlled using $U$-statistic results (Serfling, 2009), and

$$\mathrm{Var}_f(\hat{D}_n) \lesssim \frac{D(f)}{n} + \frac{|\mathcal{M}|}{n^2}.$$

As for goodness-of-fit testing, our statistic converges quicker under the null hypothesis with a rate that depends on the problem complexity.

Choosing $|\mathcal{M}| \asymp_r n^{\frac{2d}{4s+d}}$ gives convergence at rate $n^{-\frac{4s}{4s+d}}$.

# Upper bound – permuted statistic

Under permutation it is harder to analyse the statistic, but it can be done.

If $\Pi \sim \mathrm{Unif}(\mathcal{S}_n)$ we have, for example, $(\Pi(1), \Pi(2)) \overset{\mathrm{d}}{=} (\Pi(1), \Pi(3))$, and we can check

$$p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_2) \overset{\mathrm{d}}{=} p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(3)}, Y_2)$$

$$p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_3) \overset{\mathrm{d}}{=} p_{jk}(X_{\Pi(1)}, Y_2)p_{jk}(X_{\Pi(3)}, Y_4).$$

Hence

$$\mathbb{E}(\hat{D}_n^{(1)}) = \sum_{(j,k)\in\mathcal{M}} \mathbb{E}\{p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_2) - 2p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_3) + p_{jk}(X_{\Pi(1)}, Y_2)p_{jk}(X_{\Pi(3)}, Y_4)\} = 0.$$

We can also prove that $\mathrm{Var}_f(\hat{D}_n^{(1)}) \lesssim \frac{|\mathcal{M}|}{n^2} \lesssim_r n^{-\frac{4s}{4s+d}}$.

Up to constants, the moments of $\hat{D}_n^{(1)}$ behave like those of $\hat{D}_n$ under $H_0$.

# Upper bound – permuted statistic

Under permutation it is harder to analyse the statistic, but it can be done.

If $\Pi \sim \mathrm{Unif}(\mathcal{S}_n)$ we have, for example, $(\Pi(1), \Pi(2)) \overset{\mathrm{d}}{=} (\Pi(1), \Pi(3))$, and we can check

$$p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_2) \overset{\mathrm{d}}{=} p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(3)}, Y_2)$$
$$p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_3) \overset{\mathrm{d}}{=} p_{jk}(X_{\Pi(1)}, Y_2)p_{jk}(X_{\Pi(3)}, Y_4).$$

Hence

$$\mathbb{E}(\hat{D}_n^{(1)}) = \sum_{(j,k)\in\mathcal{M}} \mathbb{E}\{p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_2) - 2p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_3) + p_{jk}(X_{\Pi(1)}, Y_2)p_{jk}(X_{\Pi(3)}, Y_4)\} = 0.$$

We can also prove that $\mathrm{Var}_f(\hat{D}_n^{(1)}) \lesssim \frac{|\mathcal{M}|}{n^2} \lesssim_r n^{-\frac{4s}{4s+d}}$.

Up to constants, the moments of $\hat{D}_n^{(1)}$ behave like those of $\hat{D}_n$ under $H_0$.

# Upper bound – permuted statistic

Under permutation it is harder to analyse the statistic, but it can be done. If $\Pi \sim \mathrm{Unif}(\mathcal{S}_n)$ we have, for example, $(\Pi(1), \Pi(2)) \stackrel{\mathrm{d}}{=} (\Pi(1), \Pi(3))$, and we can check

$$p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_2) \stackrel{\mathrm{d}}{=} p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(3)}, Y_2)$$
$$p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_3) \stackrel{\mathrm{d}}{=} p_{jk}(X_{\Pi(1)}, Y_2)p_{jk}(X_{\Pi(3)}, Y_4).$$

Hence

$$\mathbb{E}(\hat{D}_n^{(1)}) = \sum_{(j,k)\in\mathcal{M}} \mathbb{E}\{p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_2) - 2p_{jk}(X_{\Pi(1)}, Y_1)p_{jk}(X_{\Pi(2)}, Y_3) + p_{jk}(X_{\Pi(1)}, Y_2)p_{jk}(X_{\Pi(3)}, Y_4)\} = 0.$$

We can also prove that $\mathrm{Var}_f(\hat{D}_n^{(1)}) \lesssim \frac{|\mathcal{M}|}{n^2} \lesssim_r n^{-\frac{4s}{4s+d}}$.

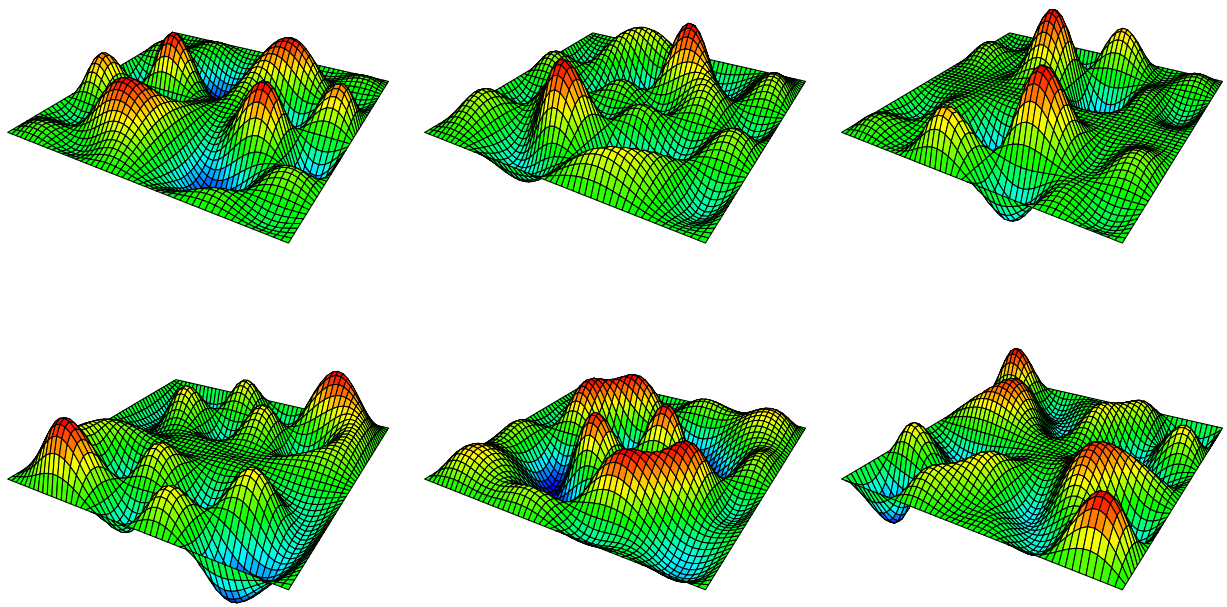
Up to constants, the moments of $\hat{D}_n^{(1)}$ behave like those of $\hat{D}_n$ under $H_0$.

# Lower bound

As for goodness-of-fit, we define a prior over $\mathcal{P}_1(\rho)$. Here we take

$$f(x,y) = 1 + \sum_{(j,k)\in\mathbb{N}^{d_X+d_Y}} w_{jk}\xi_{jk}p_{jk}(x,y).$$

for $(\xi_{jk})$ independent Rademacher and $(w_{jk})$ appropriate weights.



The marginal distribution of the data with this prior can be shown to be indistinguishable from the uniform distribution.

# Summary so far

- With composite nulls, Type I error control can be difficult.
- For some problems, random permutation can take any test statistic and give us a uniformly valid test.
- Theoretical analysis is more complex, but for independence testing we can prove that appropriate permutation tests are rate optimal.

Not discussed here:

- Choice of optimal truncation set depends on unknown smoothness, but we can be adaptive.
- We can prove stronger distributional results about $\hat{D}_n^{(1)}$ to get approximate power functions.
- Similar results also hold for two-sample testing (Kim, Balakrishnan & Wasserman, 2022).

This works for independence and two-sample testing, but these are 'nice' problems.

# Summary so far

- With composite nulls, Type I error control can be difficult.
- For some problems, random permutation can take any test statistic and give us a uniformly valid test.
- Theoretical analysis is more complex, but for independence testing we can prove that appropriate permutation tests are rate optimal.

Not discussed here:

- Choice of optimal truncation set depends on unknown smoothness, but we can be adaptive.
- We can prove stronger distributional results about $\hat{D}_n^{(1)}$ to get approximate power functions.
- Similar results also hold for two-sample testing (Kim, Balakrishnan & Wasserman, 2022).

This works for independence and two-sample testing, but these are 'nice' problems.

# Summary so far

- With composite nulls, Type I error control can be difficult.
- For some problems, random permutation can take any test statistic and give us a uniformly valid test.
- Theoretical analysis is more complex, but for independence testing we can prove that appropriate permutation tests are rate optimal.

Not discussed here:

- Choice of optimal truncation set depends on unknown smoothness, but we can be adaptive.
- We can prove stronger distributional results about $\hat{D}_n^{(1)}$ to get approximate power functions.
- Similar results also hold for two-sample testing (Kim, Balakrishnan & Wasserman, 2022).

This works for independence and two-sample testing, but these are 'nice' problems.

1. Introduction

2. Simple null hypotheses
   - The likelihood ratio test and Neyman–Pearson
   - Composite alternatives: Goodness-of-fit testing

3. Permutation testing for composite nulls
   - Independence testing – minimax optimality
   - Conditional independence testing – non-uniform permutations

4. Local differential privacy
   - Two-point testing
   - Goodness-of-fit testing

# Non-exchangeable nulls

Validity of traditional permutation tests relies on certain exchangeabilities under $H_0$, but many interesting problems do not have these.

## Conditional independence testing

Given an i.i.d. sample $(X_1, Y_1, W_1), \ldots, (X_n, Y_n, W_n)$ we aim to test

$$H_0 : X \perp\!\!\!\perp Y | W.$$

## Conditional two-sample testing

Given $(X_1, W_1), \ldots, (X_n, W_n) \sim f$ and $(Y_1, W'_1), \ldots, (Y_m, W'_m) \sim g$, we want to test

$$H_0 : X|\{W = w\} \stackrel{\mathrm{d}}{=} Y|\{W = w\} \quad \forall w.$$

Note that $f_W$ and $g_W$ may be different.

# Non-exchangeable nulls

Validity of traditional permutation tests relies on certain exchangeabilities under $H_0$, but many interesting problems do not have these.

## Conditional independence testing

Given an i.i.d. sample $(X_1, Y_1, W_1), \ldots, (X_n, Y_n, W_n)$ we aim to test

$$H_0 : X \perp\!\!\!\perp Y | W.$$

## Conditional two-sample testing

Given $(X_1, W_1), \ldots, (X_n, W_n) \sim f$ and $(Y_1, W'_1), \ldots, (Y_m, W'_m) \sim g$, we want to test

$$H_0 : X|\{W = w\} \stackrel{\mathrm{d}}{=} Y|\{W = w\} \quad \forall w.$$

Note that $f_W$ and $g_W$ may be different.

# Non-exchangeable nulls

Validity of traditional permutation tests relies on certain exchangeabilities under $H_0$, but many interesting problems do not have these.

## Conditional independence testing

Given an i.i.d. sample $(X_1, Y_1, W_1), \ldots, (X_n, Y_n, W_n)$ we aim to test

$$H_0 : X \perp\!\!\!\perp Y | W.$$

## Conditional two-sample testing

Given $(X_1, W_1), \ldots, (X_n, W_n) \sim f$ and $(Y_1, W'_1), \ldots, (Y_m, W'_m) \sim g$, we want to test

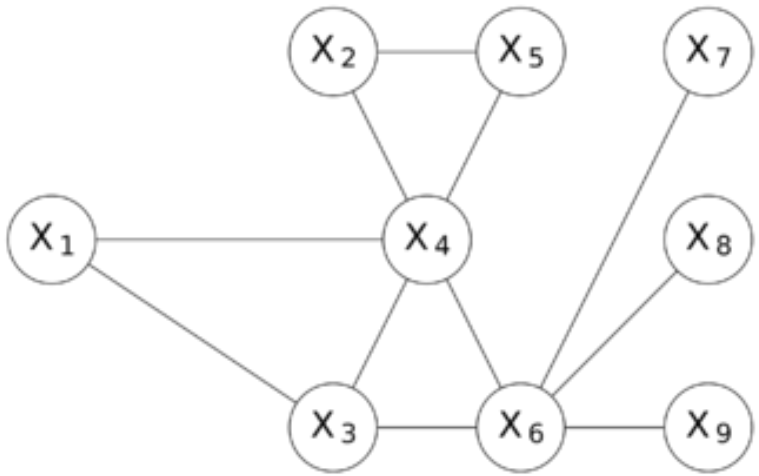
$$H_0 : X|\{W = w\} \stackrel{\mathrm{d}}{=} Y|\{W = w\} \quad \forall w.$$

Note that $f_W$ and $g_W$ may be different.

# Conditional independence testing

Conditional independence is central to statistical theory and modelling (Dawid, 1979).

We say that $X \perp\!\!\!\perp Y|W$ if the conditional density of $(X, Y)|W$ factorises as $f_{XY|W} = f_{X|W} f_{Y|W}$. Equivalently, if $f_{Y|X,W} = f_{Y|W}$.



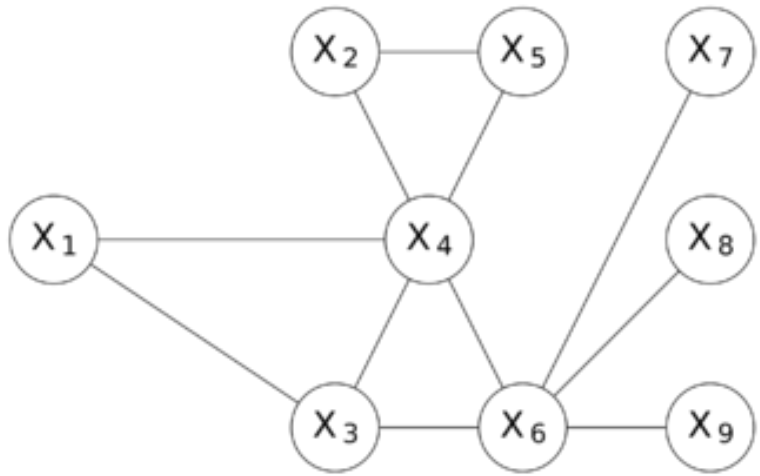

Tests of conditional independence are frequently used in data analysis:

- Constructing graphical models.
- Variable selection in regression: $X$ can be removed from model $Y \sim (X, W)$ if $X \perp\!\!\!\perp Y|W$.

# Conditional independence testing

Conditional independence is central to statistical theory and modelling (Dawid, 1979).

We say that $X \perp\!\!\!\perp Y|W$ if the conditional density of $(X, Y)|W$ factorises as $f_{XY|W} = f_{X|W} f_{Y|W}$. Equivalently, if $f_{Y|X,W} = f_{Y|W}$.




Tests of conditional independence are frequently used in data analysis:

- Constructing graphical models.
- Variable selection in regression: $X$ can be removed from model $Y \sim (X, W)$ if $X \perp\!\!\!\perp Y|W$.

# Hardness of conditional independence testing

Such conditional problems are significantly more complex. E.g. for conditional independence testing:

**Hardness result (Shah & Peters, 2020)**

For $n \in \mathbb{N}$ and $\alpha \in (0, 1)$, let $\psi_n$ be a test with $\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha$. Then

$$\sup_{Q \in \mathcal{Q}} \mathbb{P}_Q(\psi_n = 1) \leq \alpha,$$

where $\mathcal{P}$ is the class of continuous null $(X \perp\!\!\!\perp Y | W)$ distributions of $(X, Y, W)$ and $\mathcal{Q}$ is the class of continuous alternatives $(X \not\perp\!\!\!\perp Y | W)$.

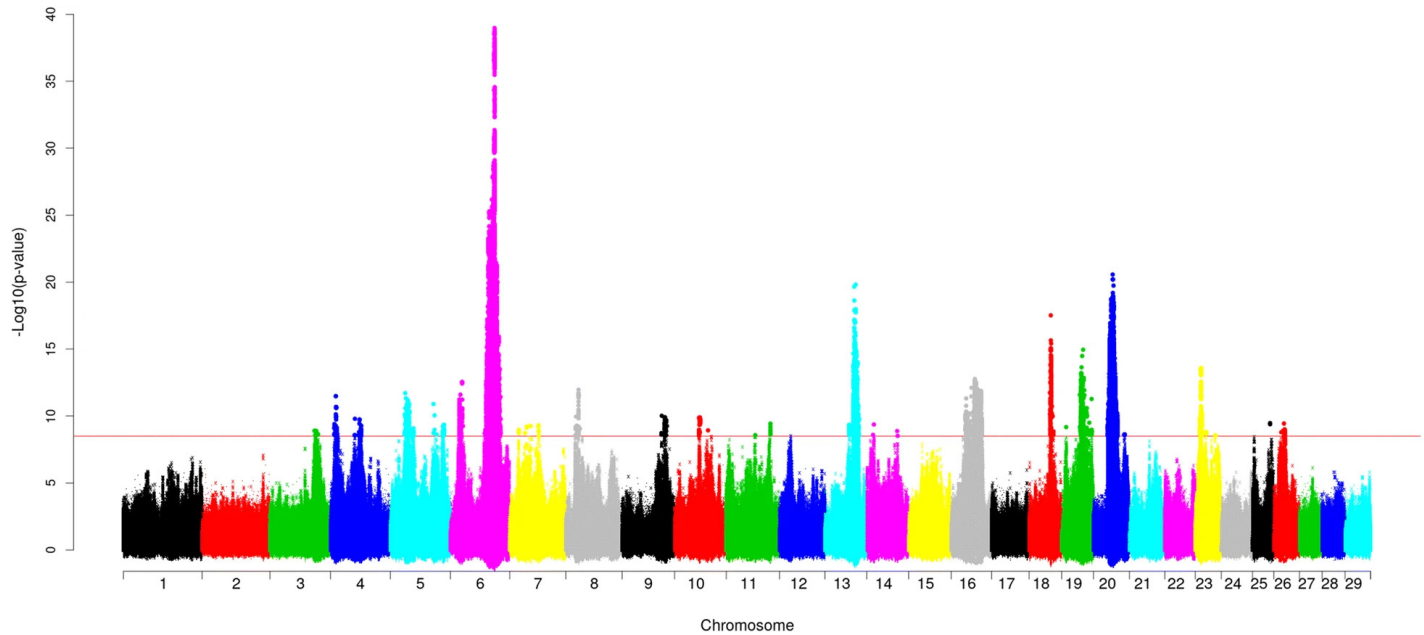
The null hypothesis is so complex that any test with uniform Type I error control must be trivial.

We will need to make assumptions to make any progress.

# Model-X framework

In many settings we have more information that can help.

In the 'Model-X' framework (Candès et al., 2018) we assume that we know (an approximation to) the conditional distribution $Q(\cdot|w)$ of $X|W = w$.




In some cases $(X, W)$ data is abundant while labeled data $(X, Y, W)$ is relatively scarce.

# The conditional permutation test

This additional information allows the construction of a permutation test.

Generate[2] $\Pi_1, \ldots, \Pi_B \in \mathcal{S}_n$ independently from the distribution

$$\mathbb{P}(\Pi = \pi | \mathbf{X}, \mathbf{Y}, \mathbf{W}) \propto \prod_{i=1}^{n} q(X_{\pi(i)} | W_i),$$

where $q(\cdot|w)$ is the density of $X|\{W = w\}$, and set $X_i^{(b)} = X_{\Pi_b(i)}$.

Generates $\mathbf{X}^{(b)}$ from $Q^n(\cdot|\mathbf{W})$ *conditional on the event that the marginal distribution is preserved.*[3]

[2] Simulation is non-trivial, but can be done. More later...

[3] Without this condition, we have the CRT of Candès et al. (2018)

# The conditional permutation test

This additional information allows the construction of a permutation test.

Generate[2] $\Pi_1, \ldots, \Pi_B \in \mathcal{S}_n$ independently from the distribution

$$\mathbb{P}(\Pi = \pi | \mathbf{X}, \mathbf{Y}, \mathbf{W}) \propto \prod_{i=1}^{n} q(X_{\pi(i)} | W_i),$$

where $q(\cdot|w)$ is the density of $X|\{W = w\}$, and set $X_i^{(b)} = X_{\Pi_b(i)}$.

Generates $\mathbf{X}^{(b)}$ from $Q^n(\cdot|\mathbf{W})$ *conditional on the event that the marginal distribution is preserved.*[3]

[2] Simulation is non-trivial, but can be done. More later...

[3] Without this condition, we have the CRT of Candès et al. (2018)

# The conditional permutation test

This additional information allows the construction of a permutation test.

Generate[2] $\Pi_1, \ldots, \Pi_B \in \mathcal{S}_n$ independently from the distribution

$$\mathbb{P}(\Pi = \pi | \mathbf{X}, \mathbf{Y}, \mathbf{W}) \propto \prod_{i=1}^{n} q(X_{\pi(i)} | W_i),$$

where $q(\cdot|w)$ is the density of $X|\{W = w\}$, and set $X_i^{(b)} = X_{\Pi_b(i)}$.

Generates $\mathbf{X}^{(b)}$ from $Q^n(\cdot|\mathbf{W})$ *conditional on the event that the marginal distribution is preserved*.[3]

[2] Simulation is non-trivial, but can be done. More later...

[3] Without this condition, we have the CRT of Candès et al. (2018)

# Validity of the CPT

$$\mathbb{P}(\Pi = \pi | \mathbf{X}, \mathbf{Y}, \mathbf{W}) \propto \prod_{i=1}^{n} q(X_{\pi(i)} | W_i),$$

This distribution ensures that $(\mathbf{X}, \mathbf{Y}, \mathbf{W}), (\mathbf{X}^{(1)}, \mathbf{Y}, \mathbf{W}), \ldots (\mathbf{X}^{(B)}, \mathbf{Y}, \mathbf{W})$ are exchangeable under $H_0$.

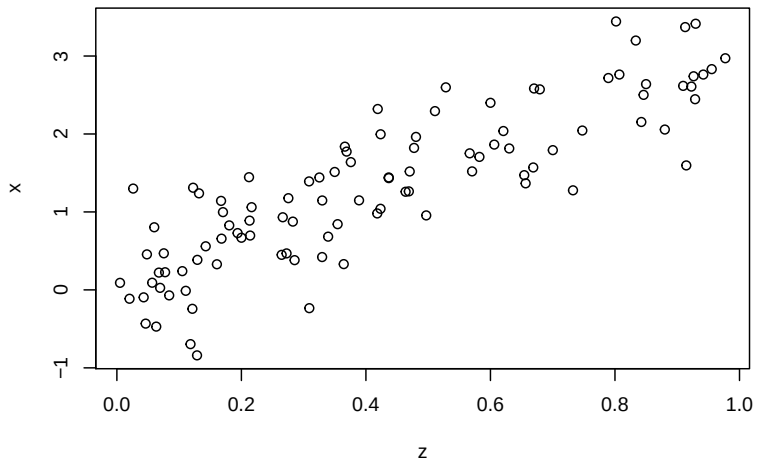
**CPT (B. et al., 2020)**

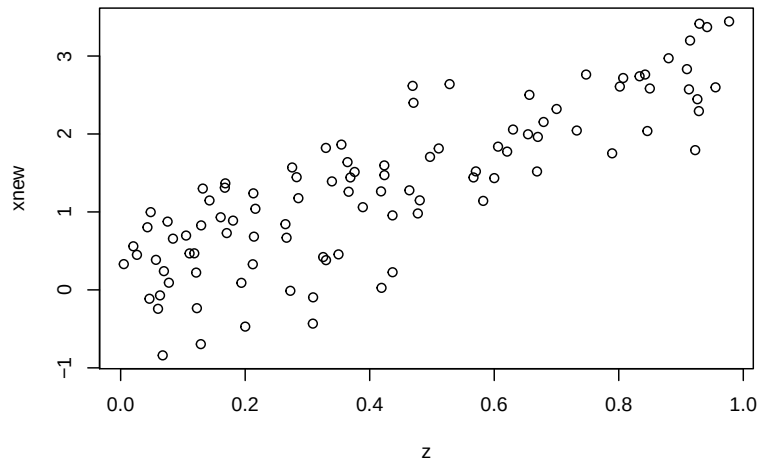
Reject if $$p = \frac{1 + \sum_{b=1}^{B} \mathbb{1}\{T(\mathbf{X}^{(b)}, \mathbf{Y}, \mathbf{W}) \geq T(\mathbf{X}, \mathbf{Y}, \mathbf{W})\}}{1 + B} \leq \alpha.$$

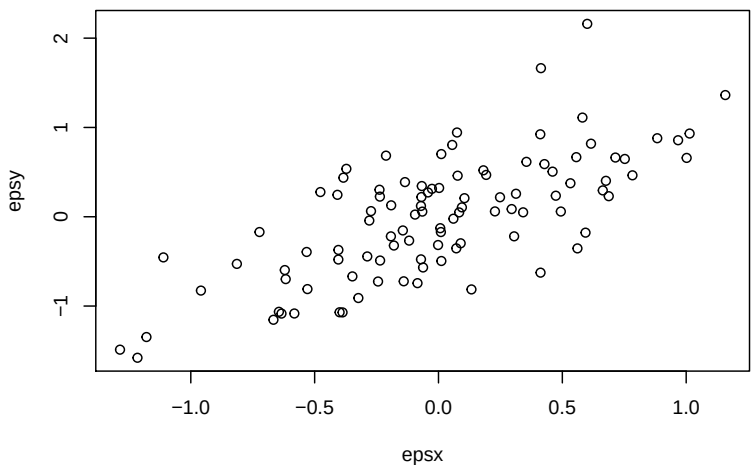
Non-uniform permutation distributions also in testing exchangeability (Ramdas et al., 2022) and weighted conformal prediction (Tibshirani et al., 2019).

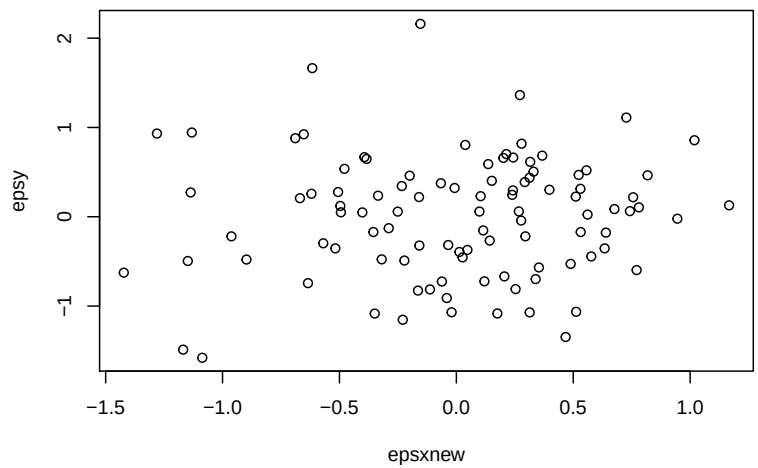
# The conditional permutation test

Consider $X|W \sim \mathcal{N}(\beta^T W, \sigma^2)$ and $Y|X, W \sim \mathcal{N}(\beta^T W + \gamma X, \sigma^2)$. We can compare the residuals $\hat{\epsilon}_X$ and $\hat{\epsilon}_Y$ after regressing $X$ and $Y$ on $W$.



# Robustness

The input $Q(\cdot|z)$ will generally be different to the true conditional distribution $Q_*(\cdot|z)$, and the Type I error control may not hold exactly.

**Theorem**

*Under $H_0$, for any test statistic $T$ and significance level $\alpha \in [0,1]$ we have that*

$$\mathbb{P}(p \leq \alpha | \mathbf{Y}, \mathbf{W}) \leq \alpha + \mathrm{TV}\big(Q_*^n(\cdot|\mathbf{W}), Q^n(\cdot|\mathbf{W})\big).$$

If we can choose $Q$ with $\mathrm{TV}\big(Q_*^n(\cdot|\mathbf{W}), Q^n(\cdot|\mathbf{W})\big) = o(1)$ then we have an approximately valid test.

# Sampling

We construct Markov chains whose stationary distribution is the desired distribution, allowing for (approximate) sampling.

---

**Algorithm** Parallelized pairwise sampler for the CPT

---

**Input:** Initial permutation $\Pi^{[0]}$, integer $S \geq 1$.

**for** $s = 1, 2, \ldots, S$ **do**

Sample uniformly without replacement from $\{1, \ldots, n\}$ to obtain disjoint pairs

$$(i_{s,1}, j_{s,1}), \ldots, (i_{s,\lfloor n/2 \rfloor}, j_{s,\lfloor n/2 \rfloor}).$$

Draw independent Bernoulli variables $B_{s,1}, \ldots, B_{s,\lfloor n/2 \rfloor}$ with odds ratios

$$\frac{\mathbb{P}(B_{s,k} = 1)}{\mathbb{P}(B_{s,k} = 0)} = \frac{q(X_{(\Pi^{[s-1]}(j_{s,k}))}|Z_{i_{s,k}}) \cdot q(X_{(\Pi^{[s-1]}(i_{s,k}))}|Z_{j_{s,k}})}{q(X_{(\Pi^{[s-1]}(i_{s,k}))}|Z_{i_{s,k}}) \cdot q(X_{(\Pi^{[s-1]}(j_{s,k}))}|Z_{j_{s,k}})}.$$

Define $\Pi^{[s]}$ by swapping $\Pi^{[s-1]}(i_{s,k})$ and $\Pi^{[s-1]}(j_{s,k})$ for each $k$ with $B_{s,k} = 1$.

**end for**

---

# Sampling

We construct Markov chains whose stationary distribution is the desired distribution, allowing for (approximate) sampling.

---

**Algorithm** Parallelized pairwise sampler for the CPT

---

**Input:** Initial permutation $\Pi^{[0]}$, integer $S \geq 1$.

**for** $s = 1, 2, \ldots, S$ **do**

Sample uniformly without replacement from $\{1, \ldots, n\}$ to obtain disjoint pairs

$$(i_{s,1}, j_{s,1}), \ldots, (i_{s,\lfloor n/2 \rfloor}, j_{s,\lfloor n/2 \rfloor}).$$

Draw independent Bernoulli variables $B_{s,1}, \ldots, B_{s,\lfloor n/2 \rfloor}$ with odds ratios

$$\frac{\mathbb{P}(B_{s,k} = 1)}{\mathbb{P}(B_{s,k} = 0)} = \frac{q(X_{(\Pi^{[s-1]}(j_{s,k}))} | Z_{i_{s,k}}) \cdot q(X_{(\Pi^{[s-1]}(i_{s,k}))} | Z_{j_{s,k}})}{q(X_{(\Pi^{[s-1]}(i_{s,k}))} | Z_{i_{s,k}}) \cdot q(X_{(\Pi^{[s-1]}(j_{s,k}))} | Z_{j_{s,k}})}.$$

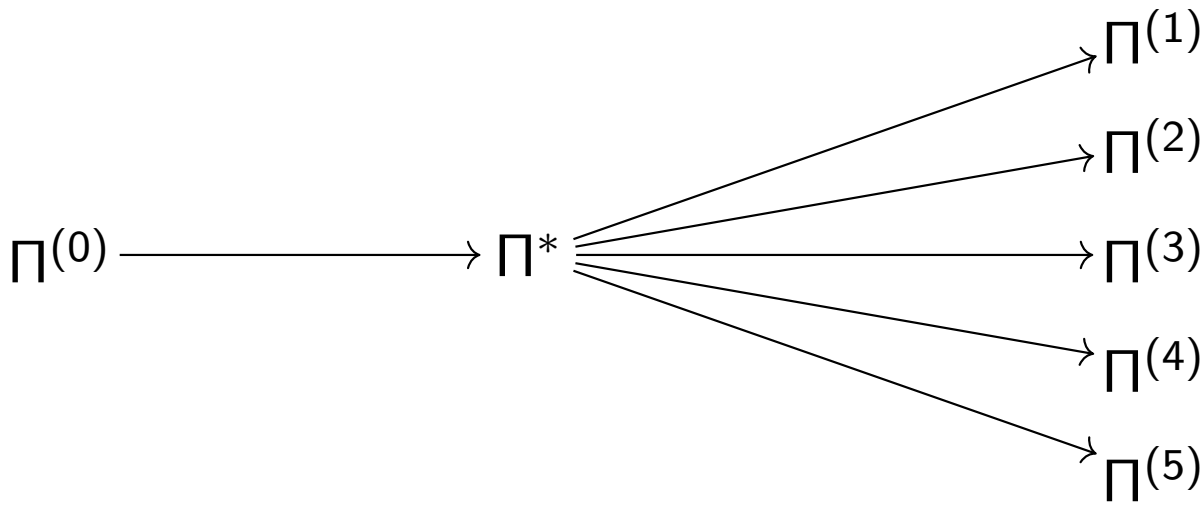
Define $\Pi^{[s]}$ by swapping $\Pi^{[s-1]}(i_{s,k})$ and $\Pi^{[s-1]}(j_{s,k})$ for each $k$ with $B_{s,k} = 1$.

**end for**

---

# Parallelised algorithm

We overcome (weak) dependence between permutations using a star-shaped sampler (Besag & Clifford, 1989).




This structure means that $(\mathbf{X}, \mathbf{Y}, \mathbf{W}), (\mathbf{X}^{(1)}, \mathbf{Y}, \mathbf{W}), \ldots (\mathbf{X}^{(B)}, \mathbf{Y}, \mathbf{W})$ are exchangeable under $H_0$, so we still have validity.

# Capital bikeshare data set

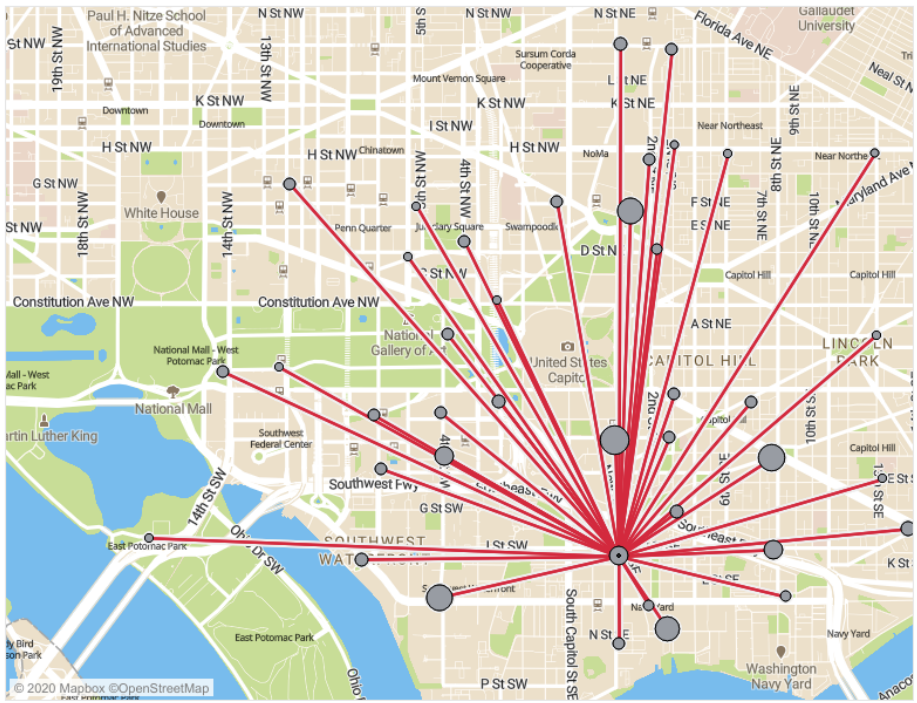
We implemented the CPT on the Capital bikeshare data set.

capital bikeshare TOP DESTINATIONS BY STATION

Select Starting Station: 1st & K St SE | Weekend/Weekday: All | Show Top Destinations: 40

**1st & K St SE**

Top 40 Destinations : All Days




**Trip Counts from 1st & K St SE**

Top 40 Destinations: All Days

*Mouse over to view route on map. All times in minutes.*

| Destination | Total Trips | Avg. Ride Time | Fastest Time |
|---|---|---|---|
| 3rd & D St SE | 268 | 5.9 | 3.6 |
| 4th & C St SW | 264 | 13.6 | 6.1 |
| 2nd & G St NE | 262 | 13.2 | 7.7 |
| 4th & D St NW / Judiciary Square | 260 | 13.8 | 7.8 |
| Eastern Market / 7th & North Caroli.. | 259 | 9.9 | 4.9 |
| North Capitol St & F St NW | 257 | 13.6 | 8.2 |
| Metro Center / 12th & G St NW | 255 | 20.9 | 11.4 |
| 1st & N St SE | 250 | 14.5 | 1.5 |
| 4th St & Madison Dr NW | 245 | 13.2 | 6.6 |
| Jefferson Dr & 14th St SW | 237 | 28.7 | 9.8 |
| L'Enfant Plaza / 7th & C St SW | 237 | 10.7 | 6.2 |
| 7th & E St SW | 228 | 9.9 | 5.8 |
| 3rd & M St NE | 223 | 16.6 | 10.6 |
| 3rd St & Pennsylvania Ave SE | 220 | 10.5 | 4.4 |
| 2nd St & Massachusetts Ave NE | 210 | 13.9 | 7.3 |
| M St & New Jersey Ave SE | 208 | 18.7 | 1.4 |
| Potomac Ave & 8th St SE | 205 | 6.7 | 3.6 |
| Lincoln Park / 13th & East Capitol St .. | 165 | 14.4 | 7.9 |
| Constitution Ave & 2nd St NW/DOL | 163 | 13.3 | 7.7 |

All 2017 Capital Bikeshare trips. Data via Capital Bikeshare (https://s3.amazonaws.com/capitalbikeshare-data/index.html)

This records for each ride the start and end times and locations and a user type that can be 'Member' or 'Casual'.

# Capital bikeshare data set

## Problem

Test $X \perp\!\!\!\perp Y | W$, where

- $X$ is the duration of the ride
- $W$ the start and end locations and time of day
- $Y$ the user type, date **or** day of the week.

For our conditional distribution we use

$$Q(\cdot|w) = \mathcal{N}(\hat{\mu}(w), \hat{\sigma}^2(w)),$$

where $\hat{\mu}(w)$ and $\hat{\sigma}^2(w)$ are calculated using the training data for each combination of start and end location with a kernel weighting times of day.

- Test data: all rides taken on weekdays in Oct 2011. Sample size n=7,346 rides (after screening).
- Training data: all rides taken on weekdays in Sep and Nov 2011. Sample size N=149,912 rides.

# Capital bikeshare data set

## Problem

Test $X \perp\!\!\!\perp Y | W$, where

- $X$ is the duration of the ride
- $W$ the start and end locations and time of day
- $Y$ the user type, date **or** day of the week.

For our conditional distribution we use

$$Q(\cdot|w) = \mathcal{N}(\hat{\mu}(w), \hat{\sigma}^2(w)),$$

where $\hat{\mu}(w)$ and $\hat{\sigma}^2(w)$ are calculated using the training data for each combination of start and end location with a kernel weighting times of day.

- Test data: all rides taken on weekdays in Oct 2011. Sample size n=7,346 rides (after screening).
- Training data: all rides taken on weekdays in Sep and Nov 2011. Sample size N=149,912 rides.

# Capital bikeshare data set

Writing $R_i = X_i - \hat{\mu}(W_i)$ we take $T$ to be the correlation between **R** and **Y** when $Y$ is the user type. When $Y$ is the day of the week we take $T$ to be

$$\max_{y \in \{\text{Mon},\ldots,\text{Fri}\}} \Big| \text{Corr}\big(\mathbf{R}, (\mathbb{1}\{Y_1 = y\}, \ldots, \mathbb{1}\{Y_n = y\})\big) \Big|.$$

With $M = 100$ we obtain the following average p-values over ten trials of the experiment:

| Variable $Y$ | CPT |
|---|---|
| User type | 0.0010 (0.0000) |
| Date | 0.1146 (0.0032) |
| Day of week | 0.1980 (0.0037) |

We see good performance in practice, but theoretical power results are rare.

# Capital bikeshare data set

Writing $R_i = X_i - \hat{\mu}(W_i)$ we take $T$ to be the correlation between $\mathbf{R}$ and $\mathbf{Y}$ when $Y$ is the user type. When $Y$ is the day of the week we take $T$ to be

$$\max_{y \in \{\text{Mon},\ldots,\text{Fri}\}} \Big| \mathrm{Corr}\big(\mathbf{R}, (\mathbb{1}\{Y_1 = y\}, \ldots, \mathbb{1}\{Y_n = y\})\big) \Big|.$$

With $M = 100$ we obtain the following average p-values over ten trials of the experiment:

| Variable $Y$ | CPT |
|---|---|
| User type | 0.0010 (0.0000) |
| Date | 0.1146 (0.0032) |
| Day of week | 0.1980 (0.0037) |

We see good performance in practice, but theoretical power results are rare.

# Conclusion

Permutation testing is a robust and powerful approach that can be widely used.

- Minimax rate optimal power guarantees for independence and two-sample testing.
- With additional knowledge and non-uniform permutations we can introduce conditional independence tests with Type I error guarantees.

1. Introduction

2. Simple null hypotheses
   - The likelihood ratio test and Neyman–Pearson
   - Composite alternatives: Goodness-of-fit testing

3. Permutation testing for composite nulls
   - Independence testing – minimax optimality
   - Conditional independence testing – non-uniform permutations

4. Local differential privacy
   - Two-point testing
   - Goodness-of-fit testing

# Beyond classical models

So far we have been considering the classical model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ and answering questions about $P$. We have:

- found (rate) optimal tests of various hypotheses;
- given tight characterisations of the power of optimal tests, or the fundamental difficulties of the problems.

In modern data science, we may have other concerns than simply minimising statistical error. We may also have to deal with

- limited computational/communication resources
- contaminated/missing data
- privacy/fairness constraints
- ...

These considerations may need to balanced with statistical accuracy. With minimax optimality results we can quantify trade-offs.

# Beyond classical models

So far we have been considering the classical model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ and answering questions about $P$. We have:

- found (rate) optimal tests of various hypotheses;
- given tight characterisations of the power of optimal tests, or the fundamental difficulties of the problems.

In modern data science, we may have other concerns than simply minimising statistical error. We may also have to deal with

- limited computational/communication resources
- contaminated/missing data
- privacy/fairness constraints
- ...

These considerations may need to balanced with statistical accuracy. With minimax optimality results we can quantify trade-offs.

# Personal data

In this lecture we will consider *differential privacy*.

The collection and use of personal data is increasingly common in modern society.

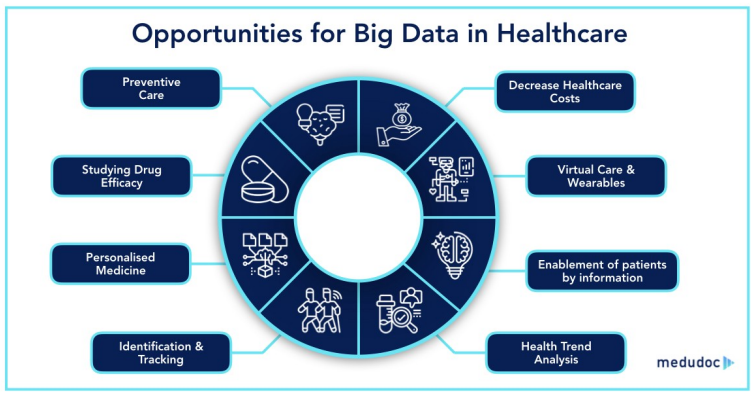
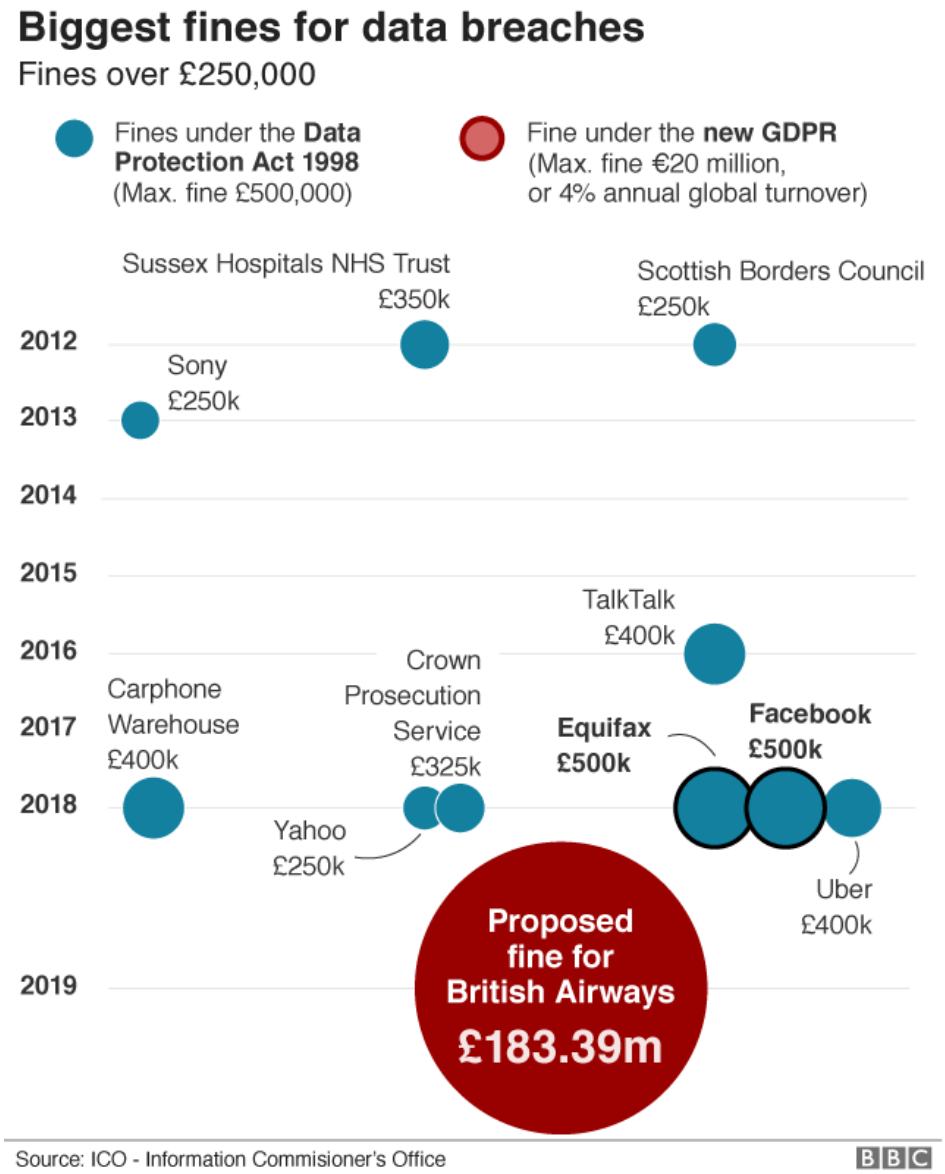



Figure: Source: `medium.com`

Sensitive information is routinely collected in modern technology, the census, medicine, finance...

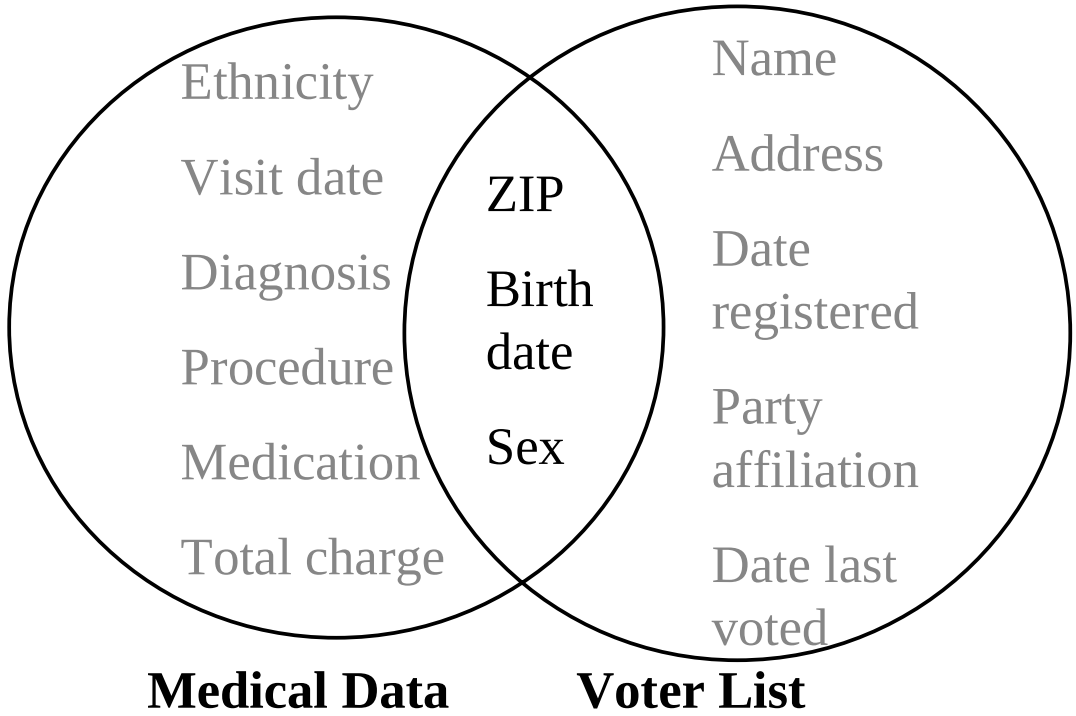
# Data breaches

At the same time, data breaches have become a feature of everyday life.



# Traditional anonymisation is not enough

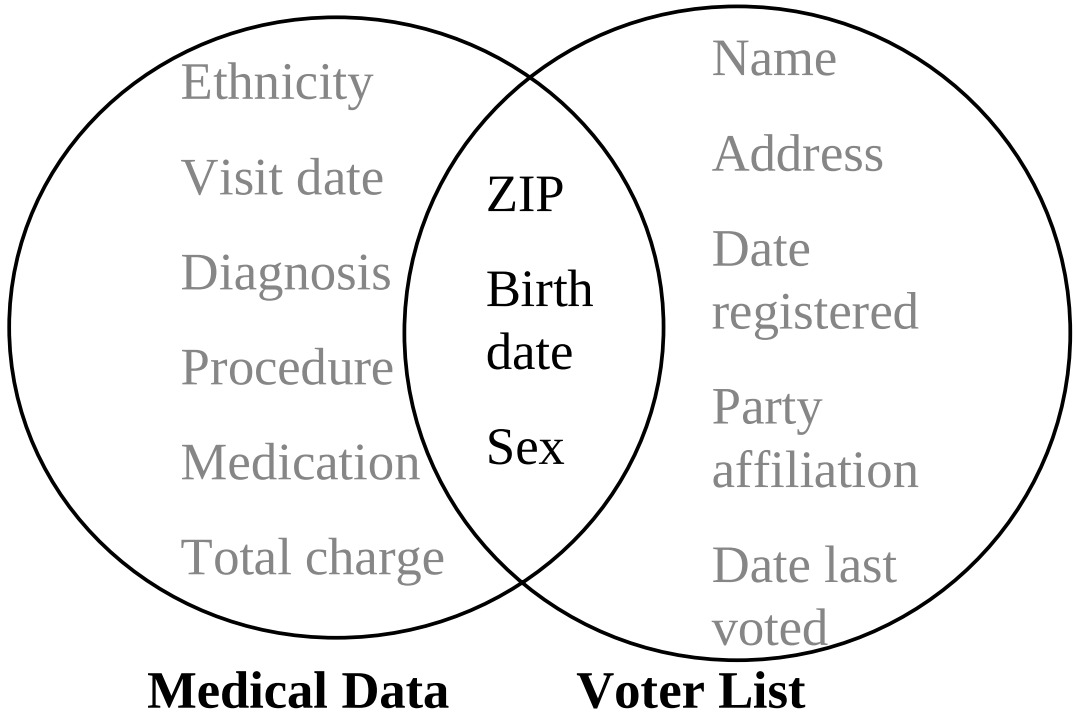
Removing names/addresses is insufficient to prevent re-identification.




Through 'anonymised' state medical records and publicly available voter registration lists, Sweeney (2002) found the medical records of the governor of Massachusetts.

# Traditional anonymisation is not enough

Removing names/addresses is insufficient to prevent re-identification.




Through 'anonymised' state medical records and publicly available voter registration lists, Sweeney (2002) found the medical records of the governor of Massachusetts.

# Summary statistics may also be insecure

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0, I_d)$ is used to calculate $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, which is published.

Let $X_1', \ldots, X_n' \overset{\text{i.i.d.}}{\sim} N(0, I_d)$ be data from individuals not in the sample.

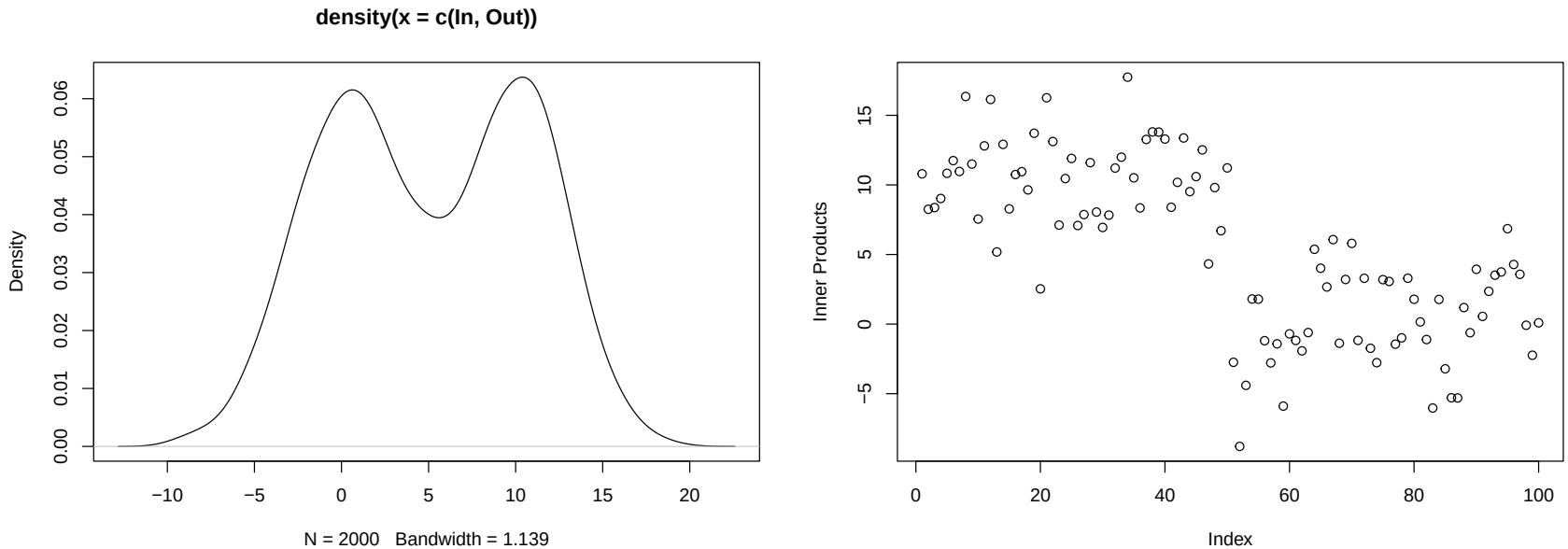



Figure: $n = 1000, \quad d = 10000$

Distributions of $\bar{X}_n^T X_i$ and $\bar{X}_n^T X_i'$ can be very different (e.g. Homer et al., 2008).

# Privacy mechanisms

*Privacy mechanisms* are algorithms taking input $\mathbf{X} = (X_1, \ldots, X_n)$ and producing randomised publishable $\mathbf{Z}$.

Formally, they are sets of conditional distributions $Q = \{Q(\cdot|\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ such that

$$\mathbf{Z}|\{\mathbf{X} = \mathbf{x}\} \sim Q(\cdot|\mathbf{x}).$$

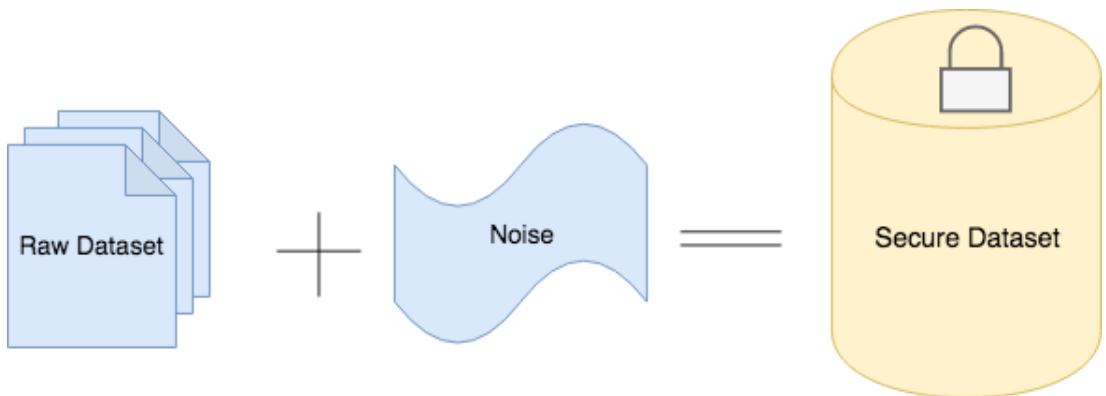



Figure: Source: `medium.com`

We choose $Q$ to preserve the most pertinent information. *How much* noise should we add? *What type* of noise?

# Randomised response

The basic idea is present in Warner (1965). Say $X_1, \ldots, X_n$ are i.i.d. binary variables and we want to estimate $\pi = \mathbb{P}(X_1 = 1)$.

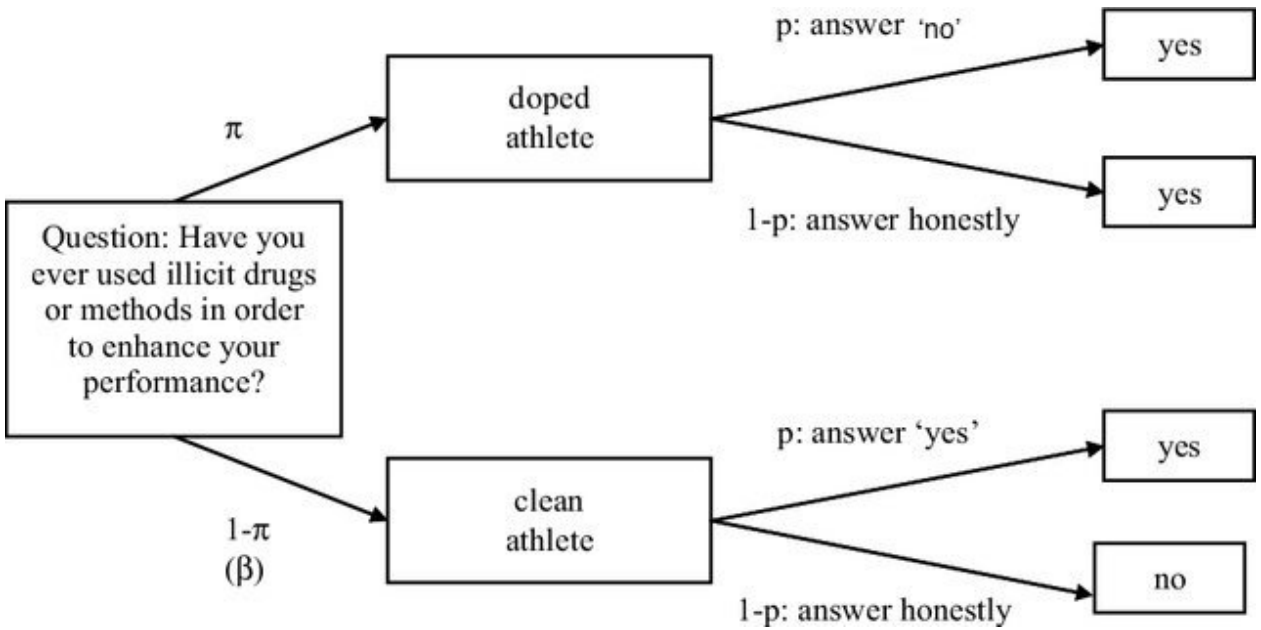



Figure: Source: Adapted from Pitsch & Emrich (2012)

# Randomised response

Ask individual $i$ to calculate and transmit

$$Z_i = \begin{cases} X_i & \text{with probability } p \\ 1 - X_i & \text{otherwise} \end{cases},$$

where $p$ controls the amount of noise. Typically $p > 1/2$.

Given $Z_1, \ldots, Z_n$, the MLE for $\theta$ is

$$\hat{\theta}_n = \frac{1}{n(2p-1)} \sum_{i=1}^{n} \{Z_i - (1-p)\}.$$

We have $\mathbb{E}\hat{\theta}_n = \theta$ and

$$\mathrm{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n(2p-1)^2},$$

an inflation of the non-private ($p = 1$) variance.

- *How much* noise should we add?
- Beyond binary data: *What type* of noise?

# Randomised response

Ask individual $i$ to calculate and transmit

$$Z_i = \begin{cases} X_i & \text{with probability } p \\ 1 - X_i & \text{otherwise} \end{cases},$$

where $p$ controls the amount of noise. Typically $p > 1/2$.

Given $Z_1, \ldots, Z_n$, the MLE for $\theta$ is

$$\hat{\theta}_n = \frac{1}{n(2p-1)} \sum_{i=1}^{n} \{Z_i - (1-p)\}.$$

We have $\mathbb{E}\hat{\theta}_n = \theta$ and

$$\mathrm{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n(2p-1)^2},$$

an inflation of the non-private ($p = 1$) variance.

- *How much* noise should we add?
- Beyond binary data: *What type* of noise?

# Randomised response

Ask individual $i$ to calculate and transmit

$$Z_i = \begin{cases} X_i & \text{with probability } p \\ 1 - X_i & \text{otherwise} \end{cases},$$

where $p$ controls the amount of noise. Typically $p > 1/2$.

Given $Z_1, \ldots, Z_n$, the MLE for $\theta$ is

$$\hat{\theta}_n = \frac{1}{n(2p-1)} \sum_{i=1}^{n} \{Z_i - (1-p)\}.$$

We have $\mathbb{E}\hat{\theta}_n = \theta$ and

$$\mathrm{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n(2p-1)^2},$$

an inflation of the non-private $(p = 1)$ variance.

- *How much* noise should we add?
- Beyond binary data: *What type* of noise?

# Differential privacy

Privacy mechanism $Q$ is called $\alpha$-*differentially private* (Dwork et al., 2006) if

$$\sup_A \frac{Q(A|\mathbf{x})}{Q(A|\mathbf{x}')} = \sup_A \frac{\mathbb{P}(\mathbf{Z} \in A|\mathbf{X} = \mathbf{x})}{\mathbb{P}(\mathbf{Z} \in A|\mathbf{X} = \mathbf{x}')} \leq e^{\alpha}$$

for all $\mathbf{x}, \mathbf{x}'$ such that $d_{\mathrm{Ham}}(\mathbf{x}, \mathbf{x}') := \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq x_i'\}} \leq 1$. Assume $\alpha \leq 1$.

Large scale applications include

- Google Chrome (Erlingsson, Pihur & Korolova, 2014);
- Apple in iOS and macOS (Tang et al., 2017);
- Microsoft (Ding, Kulkarni & Yekhanin, 2017);
- Uber (Near, 2018);
- US Census (Dwork, 2019).

Can also be used to demonstrate GDPR compliance (Cohen & Nissim, 2020).

# Differential privacy

Privacy mechanism $Q$ is called $\alpha$-*differentially private* (Dwork et al., 2006) if

$$\sup_A \frac{Q(A|\mathbf{x})}{Q(A|\mathbf{x}')} = \sup_A \frac{\mathbb{P}(\mathbf{Z} \in A|\mathbf{X} = \mathbf{x})}{\mathbb{P}(\mathbf{Z} \in A|\mathbf{X} = \mathbf{x}')} \leq e^{\alpha}$$

for all $\mathbf{x}, \mathbf{x}'$ such that $d_{\mathrm{Ham}}(\mathbf{x}, \mathbf{x}') := \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq x_i'\}} \leq 1$. Assume $\alpha \leq 1$.

Large scale applications include

- Google Chrome (Erlingsson, Pihur & Korolova, 2014);
- Apple in iOS and macOS (Tang et al., 2017);
- Microsoft (Ding, Kulkarni & Yekhanin, 2017);
- Uber (Near, 2018);
- US Census (Dwork, 2019).

Can also be used to demonstrate GDPR compliance (Cohen & Nissim, 2020).

# A statistical interpretation

Suppose that the data $X_1, \ldots, X_n$ are i.i.d. with distribution $P$.

If an adversary sees $\mathbf{Z}$ and, knowing $P$ and $Q$, tests

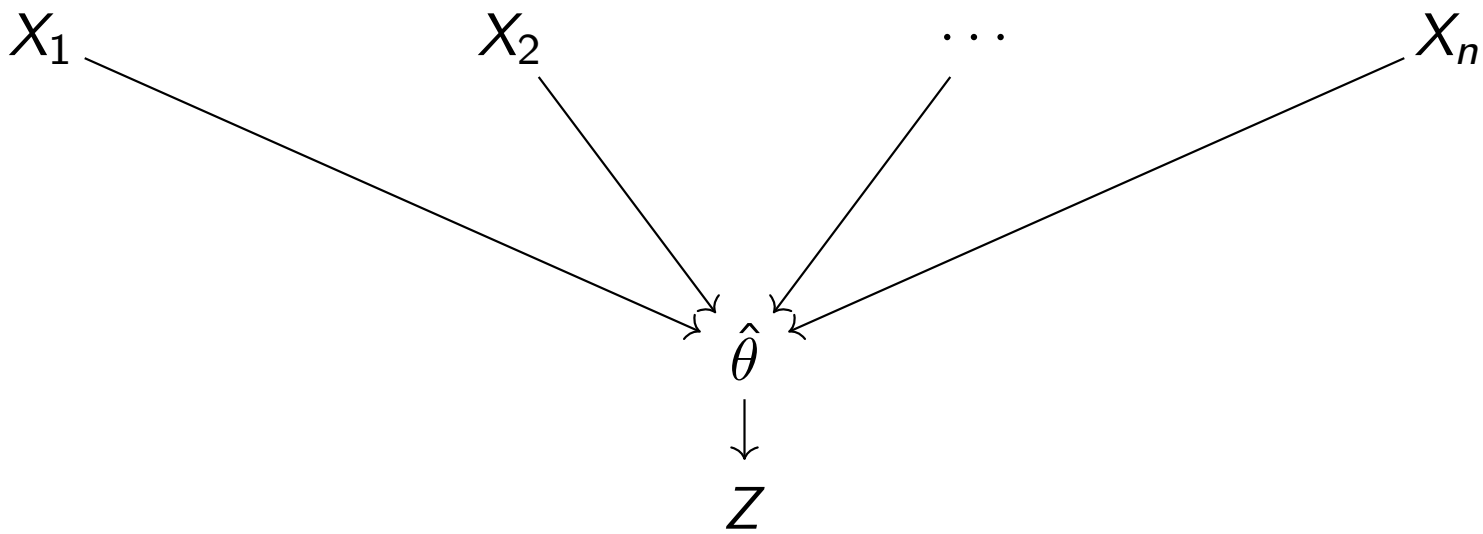
$$H_0 : X_1 = x_1 \quad \text{vs.} \quad H_1 : X_1 = x_1'$$

at significance level $\gamma$, the power of the test is bounded above by $\gamma e^{\alpha}$ (Wasserman & Zhou, 2010).

We cannot reliably work out whose data is in the database.

# Central model

The least restrictive model is the *central model* (Dwork & Lei, 2009; Wasserman & Zhou, 2010; Cai, Wang & Zhang, 2021).




The statistician has access to the raw data, but can only publish randomised statistics. E.g. the census.

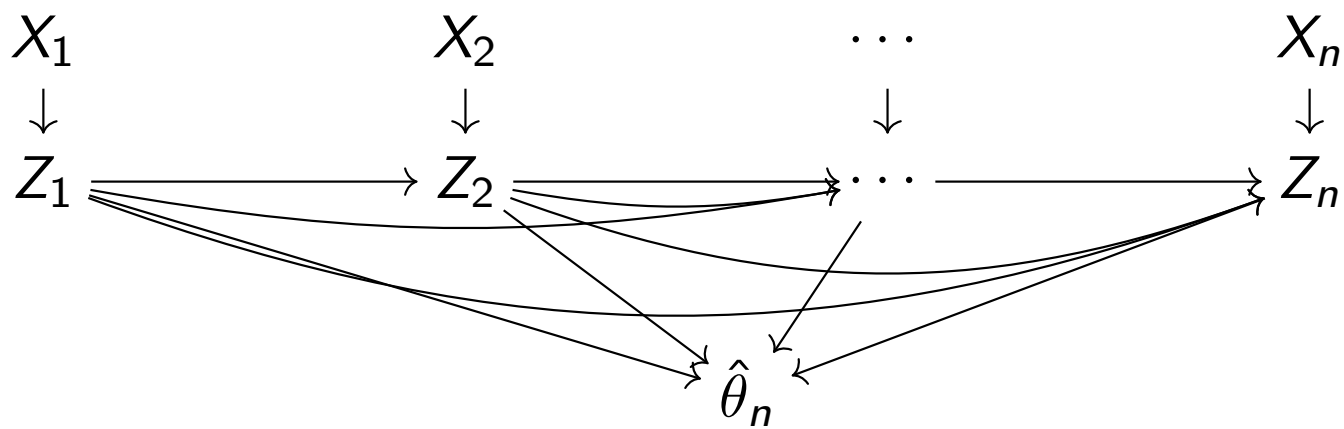
E.g., $Z = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n\alpha} W$.

# Local differential privacy

We consider the *local model* (Kairouz, Oh & Viswanath, 2014; Duchi, Jordan & Wainwright, 2018; Rohde & Steinberger, 2020; Acharya et al., 2024), where data are randomised before collection.

$$\sup_A \sup_{x_i, x_i', z^{(1:i-1)}} \frac{\mathbb{P}(Z_i \in A | X_i = x_i)}{\mathbb{P}(Z_i \in A | X_i = x_i')} \leq e^\alpha \text{ for all } i.$$

No trusted third party: analyse $Z = (Z_1, \ldots, Z_n)$ with




E.g., $Z_i = X_i + \frac{1}{\alpha} W_i$ and $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{n} \sum_{i=1}^n X_i + \frac{1}{n\alpha} \sum_{i=1}^n W_i$.

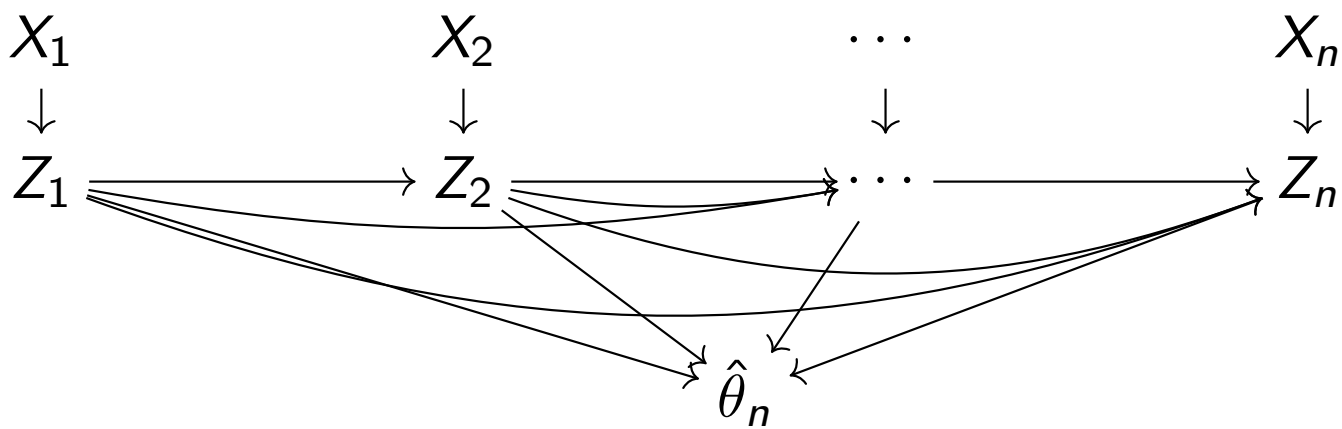
For given $\alpha > 0$, we look for a procedure with minimal statistical error.

# Local differential privacy

We consider the *local model* (Kairouz, Oh & Viswanath, 2014; Duchi, Jordan & Wainwright, 2018; Rohde & Steinberger, 2020; Acharya et al., 2024), where data are randomised before collection.

$$\sup_A \sup_{x_i, x_i', z^{(1:i-1)}} \frac{\mathbb{P}(Z_i \in A | X_i = x_i)}{\mathbb{P}(Z_i \in A | X_i = x_i')} \leq e^\alpha \text{ for all } i.$$

No trusted third party: analyse $Z = (Z_1, \ldots, Z_n)$ with




E.g., $Z_i = X_i + \frac{1}{\alpha} W_i$ and $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{n} \sum_{i=1}^n X_i + \frac{1}{n\alpha} \sum_{i=1}^n W_i$.

For given $\alpha > 0$, we look for a procedure with minimal statistical error.

# Simple hypothesis testing – Scheffé test

Let's reconsider the simple hypothesis testing problem

$$H_0 : P = P_0 \quad \text{vs.} \quad H_1 : P = P_0$$

for fixed distributions $P_0, P_1$, given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$.

The classical LR statistic $\prod_{i=1}^{n} \frac{dP_1}{dP_0}(X_i)$ is difficult to privatise, but we can use ideas from *robust statistics* (e.g. Devroye & Lugosi, 2001; Chen, Gao & Ren, 2016).

In the non-private setting the *Scheffé test* rejects $H_0$ if and only if

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}} > \frac{1}{2}\{P_0(A) + P_1(A)\},$$

where $A$ is such that $P_1(A) - P_0(A) = \sup_S \{P_1(S) - P_0(S)\}$. We can see that each data point now has a limited effect on the test.

# Simple hypothesis testing – Scheffé test

Let's reconsider the simple hypothesis testing problem

$$H_0 : P = P_0 \quad \text{vs.} \quad H_1 : P = P_0$$

for fixed distributions $P_0, P_1$, given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$.

The classical LR statistic $\prod_{i=1}^n \frac{dP_1}{dP_0}(X_i)$ is difficult to privatise, but we can use ideas from *robust statistics* (e.g. Devroye & Lugosi, 2001; Chen, Gao & Ren, 2016).

In the non-private setting the *Scheffé test* rejects $H_0$ if and only if

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \in A\}} > \frac{1}{2}\{P_0(A) + P_1(A)\},$$

where $A$ is such that $P_1(A) - P_0(A) = \sup_S\{P_1(S) - P_0(S)\}$. We can see that each data point now has a limited effect on the test.

# Simple hypothesis testing

This can be applied to the output of the randomised response mechanism (Warner, 1965; Gopi et al., 2020)

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in A\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in A\}}, & \text{otherwise.} \end{cases}$$

This is a valid $\alpha$-LDP mechanism since

$$\frac{\mathbb{P}(Z=1|X \in A)}{\mathbb{P}(Z=1|X \in A^c)} = \frac{\mathbb{P}(Z=0|X \in A^c)}{\mathbb{P}(Z=0|X \in A)} = e^\alpha.$$

Reject $H_0$ if and only if

$$\frac{e^\alpha+1}{n(e^\alpha-1)} \sum_{i=1}^{n} \left( Z_i - \frac{1}{e^\alpha+1} \right) > \frac{1}{2}\{P_0(A) + P_1(A)\}.$$

# Simple hypothesis testing

This can be applied to the output of the randomised response mechanism (Warner, 1965; Gopi et al., 2020)

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in A\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in A\}}, & \text{otherwise.} \end{cases}$$

This is a valid $\alpha$-LDP mechanism since

$$\frac{\mathbb{P}(Z=1|X \in A)}{\mathbb{P}(Z=1|X \in A^c)} = \frac{\mathbb{P}(Z=0|X \in A^c)}{\mathbb{P}(Z=0|X \in A)} = e^\alpha.$$

Reject $H_0$ if and only if

$$\frac{e^\alpha+1}{n(e^\alpha-1)} \sum_{i=1}^n \left( Z_i - \frac{1}{e^\alpha+1} \right) > \frac{1}{2}\{P_0(A) + P_1(A)\}.$$

# Simple hypothesis testing

This can be applied to the output of the randomised response mechanism (Warner, 1965; Gopi et al., 2020)

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in A\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in A\}}, & \text{otherwise.} \end{cases}$$

This is a valid $\alpha$-LDP mechanism since

$$\frac{\mathbb{P}(Z=1|X \in A)}{\mathbb{P}(Z=1|X \in A^c)} = \frac{\mathbb{P}(Z=0|X \in A^c)}{\mathbb{P}(Z=0|X \in A)} = e^\alpha.$$

Reject $H_0$ if and only if

$$\frac{e^\alpha+1}{n(e^\alpha-1)} \sum_{i=1}^n \left( Z_i - \frac{1}{e^\alpha+1} \right) > \frac{1}{2}\{P_0(A) + P_1(A)\}.$$

# Upper bound

By construction, we have

$$\mathbb{E}\left\{\frac{e^\alpha+1}{n(e^\alpha-1)}\sum_{i=1}^n\left(Z_i-\frac{1}{e^\alpha+1}\right)\right\}=\frac{e^\alpha+1}{e^\alpha-1}\left\{\mathbb{P}(Z=1)-\frac{1}{e^\alpha+1}\right\}$$
$$=\frac{e^\alpha+1}{e^\alpha-1}\left\{\frac{e^\alpha}{1+e^\alpha}P(A)+\frac{1}{1+e^\alpha}P(A^c)-\frac{1}{e^\alpha+1}\right\}=P(A).$$

We can then use Hoeffding's inequality:

$$\mathbb{E}_{P_0}(\phi)=\mathbb{P}_{P_0}\left(\frac{e^\alpha+1}{n(e^\alpha-1)}\sum_{i=1}^n(Z_i-\mathbb{E}Z_i)>\frac{1}{2}\{P_0(A)+P_1(A)\}-P_0(A)\right)$$
$$=\mathbb{P}_{P_0}\left(\frac{1}{n}\sum_{i=1}^n(Z_i-\mathbb{E}Z_i)>\frac{1}{2}\frac{e^\alpha-1}{e^\alpha+1}\mathrm{TV}(P_0,P_1)\right)$$
$$\leq\exp\left(-\frac{n}{2}\left(\frac{e^\alpha-1}{e^\alpha+1}\right)^2\mathrm{TV}(P_0,P_1)^2\right),$$

with an almost identical bound for the Type II error.

# Upper bound

By construction, we have

$$\mathbb{E}\left\{\frac{e^\alpha+1}{n(e^\alpha-1)}\sum_{i=1}^n\left(Z_i-\frac{1}{e^\alpha+1}\right)\right\}=\frac{e^\alpha+1}{e^\alpha-1}\left\{\mathbb{P}(Z=1)-\frac{1}{e^\alpha+1}\right\}$$
$$=\frac{e^\alpha+1}{e^\alpha-1}\left\{\frac{e^\alpha}{1+e^\alpha}P(A)+\frac{1}{1+e^\alpha}P(A^c)-\frac{1}{e^\alpha+1}\right\}=P(A).$$

We can then use Hoeffding's inequality:

$$\mathbb{E}_{P_0}(\phi)=\mathbb{P}_{P_0}\left(\frac{e^\alpha+1}{n(e^\alpha-1)}\sum_{i=1}^n(Z_i-\mathbb{E}Z_i)>\frac{1}{2}\{P_0(A)+P_1(A)\}-P_0(A)\right)$$
$$=\mathbb{P}_{P_0}\left(\frac{1}{n}\sum_{i=1}^n(Z_i-\mathbb{E}Z_i)>\frac{1}{2}\frac{e^\alpha-1}{e^\alpha+1}\mathrm{TV}(P_0,P_1)\right)$$
$$\leq\exp\left(-\frac{n}{2}\left(\frac{e^\alpha-1}{e^\alpha+1}\right)^2\mathrm{TV}(P_0,P_1)^2\right),$$

with an almost identical bound for the Type II error.

# Private information inequality

To prove a matching lower bound, we require the following result, thought of as a *data processing inequality*.

**Theorem (Duchi, Jordan & Wainwright 2018)**

*Let $Q$ be any $\alpha$-LDP privacy mechanism and $P_0, P_1$ any distributions. Then*

$$\mathrm{KL}\big(QP_0^{\otimes n}, QP_1^{\otimes n}\big) \leq 4n(e^\alpha - 1)^2 \mathrm{TV}(P_0, P_1)^2.$$

For small $\alpha$ the RHS scales like $O(\alpha^2)$, so that privacy mechanisms act as contractions obscuring information.

We can bound a private KL divergence by non-private TV distance, the opposite direction to Pinsker!

# Lower bound

Let $Q$ be any $\alpha$-LDP privacy mechanism and $\phi$ any test based on its output. Then

$$\begin{aligned}
\mathbb{E}_{P_0}\phi + \mathbb{E}_{P_1}(1-\phi) &\geq 1 - \mathrm{TV}(QP_0^{\otimes n}, QP_1^{\otimes n}) && \text{as before} \\
&\geq \frac{1}{2}\exp\bigl(-\mathrm{KL}\bigl(QP_0^{\otimes n}, QP_1^{\otimes n}\bigr)\bigr) && \text{Bretagnolle–Huber} \\
&\geq \frac{1}{2}\exp\bigl(-4n(e^\alpha - 1)^2 \mathrm{TV}(P_0, P_1)^2\bigr)
\end{aligned}$$

Duchi, Jordan & Wainwright (2018)

The minimax risk satisfies

$$\frac{1}{2}\exp\bigl(-4n(e^\alpha - 1)^2 \mathrm{TV}(P_0, P_1)^2\bigr) \leq \inf_Q \inf_\phi \bigl\{\mathbb{E}_{P_0}\phi + \mathbb{E}_{P_1}(1-\phi)\bigr\}$$

$$\leq 2\exp\left(-\frac{n}{2}\left(\frac{e^\alpha - 1}{e^\alpha + 1}\right)^2 \mathrm{TV}(P_0, P_1)^2\right).$$

Up to constants, this is tight in the *high-privacy* regime $\alpha \leq 1$.

# Lower bound

Let $Q$ be any $\alpha$-LDP privacy mechanism and $\phi$ any test based on its output. Then

$$\begin{aligned}
\mathbb{E}_{P_0}\phi + \mathbb{E}_{P_1}(1-\phi) &\geq 1 - \mathrm{TV}(QP_0^{\otimes n}, QP_1^{\otimes n}) && \text{as before} \\
&\geq \frac{1}{2}\exp\bigl(-\mathrm{KL}\bigl(QP_0^{\otimes n}, QP_1^{\otimes n}\bigr)\bigr) && \text{Bretagnolle–Huber} \\
&\geq \frac{1}{2}\exp\bigl(-4n(e^\alpha - 1)^2 \mathrm{TV}(P_0, P_1)^2\bigr)
\end{aligned}$$

Duchi, Jordan & Wainwright (2018)

The minimax risk satisfies

$$\begin{aligned}
\frac{1}{2}\exp\bigl(-4n(e^\alpha - 1)^2 \mathrm{TV}(P_0, P_1)^2\bigr) &\leq \inf_Q \inf_\phi \bigl\{\mathbb{E}_{P_0}\phi + \mathbb{E}_{P_1}(1-\phi)\bigr\} \\
&\leq 2\exp\left(-\frac{n}{2}\left(\frac{e^\alpha - 1}{e^\alpha + 1}\right)^2 \mathrm{TV}(P_0, P_1)^2\right).
\end{aligned}$$

Up to constants, this is tight in the *high-privacy* regime $\alpha \leq 1$.

# Cost of privacy for two-point testing

Let's compare this to classical results and *Huber's contamination model* $X \sim (1-\epsilon)P + \epsilon G$.

For combined error rate $\leq 0.1$ we require:

- Classical model: $\mathrm{H}(P_0, P_1) \gtrsim 1/\sqrt{n}$;
- $\alpha$-LDP: $\mathrm{TV}(P_0, P_1) \gtrsim 1/\sqrt{n\alpha^2}$ (e.g. Gopi et al., 2020);
- $\varepsilon$-Huber with $n = \infty$: $\mathrm{TV}(P_0, P_1) > \varepsilon/(1-\varepsilon)$ (e.g. Chen, Gao & Ren, 2016);
- $\alpha$-LDP and $\varepsilon$-Huber: $\mathrm{TV}(P_0, P_1) \gtrsim \varepsilon + 1/\sqrt{n\alpha^2}$ (Li et al., 2023)

when $\alpha \leq 1$.

There are deep connections between robust statistics and (local) differential privacy (e.g. Dwork & Lei, 2009; Avella-Medina, 2021; Li et al., 2023).

Extends to estimation of functionals (Donoho & Liu, 1991; Rohde & Steinberger, 2020).

# Cost of privacy for two-point testing

Let's compare this to classical results and *Huber's contamination model* $X \sim (1-\epsilon)P + \epsilon G$.

For combined error rate $\leq 0.1$ we require:

- Classical model: $\mathrm{H}(P_0, P_1) \gtrsim 1/\sqrt{n}$;
- $\alpha$-LDP: $\mathrm{TV}(P_0, P_1) \gtrsim 1/\sqrt{n\alpha^2}$ (e.g. Gopi et al., 2020);
- $\varepsilon$-Huber with $n = \infty$: $\mathrm{TV}(P_0, P_1) > \varepsilon/(1-\varepsilon)$ (e.g. Chen, Gao & Ren, 2016);
- $\alpha$-LDP and $\varepsilon$-Huber: $\mathrm{TV}(P_0, P_1) \gtrsim \varepsilon + 1/\sqrt{n\alpha^2}$ (Li et al., 2023)

when $\alpha \leq 1$.

There are deep connections between robust statistics and (local) differential privacy (e.g. Dwork & Lei, 2009; Avella-Medina, 2021; Li et al., 2023).

Extends to estimation of functionals (Donoho & Liu, 1991; Rohde & Steinberger, 2020).

# Uniformity testing

Recall that when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ on $[d]$ we want to test

$$H_0 : P = \mathrm{Unif}([d]) \quad \text{vs.} \quad H_1(\rho) : \mathrm{TV}(P, \mathrm{Unif}([d])) \geq \rho.$$

For simple hypothesis testing a *non-interactive* method was optimal.



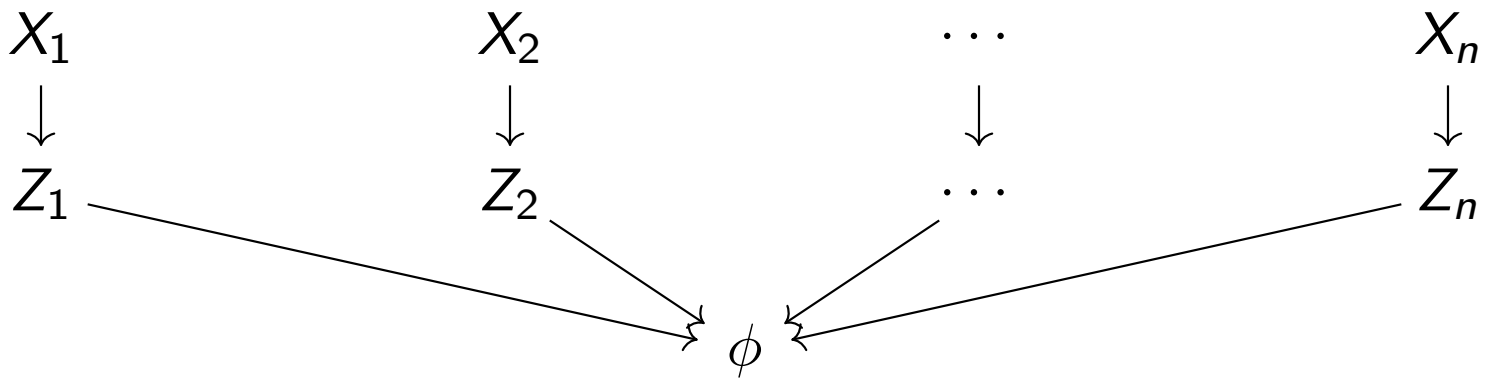

In GoF we see that *sequentially interactive* methods can do better.

# Uniformity testing

Recall that when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ on $[d]$ we want to test

$$H_0 : P = \mathrm{Unif}([d]) \quad \text{vs.} \quad H_1(\rho) : \mathrm{TV}(P, \mathrm{Unif}([d])) \geq \rho.$$

For simple hypothesis testing a *non-interactive* method was optimal.



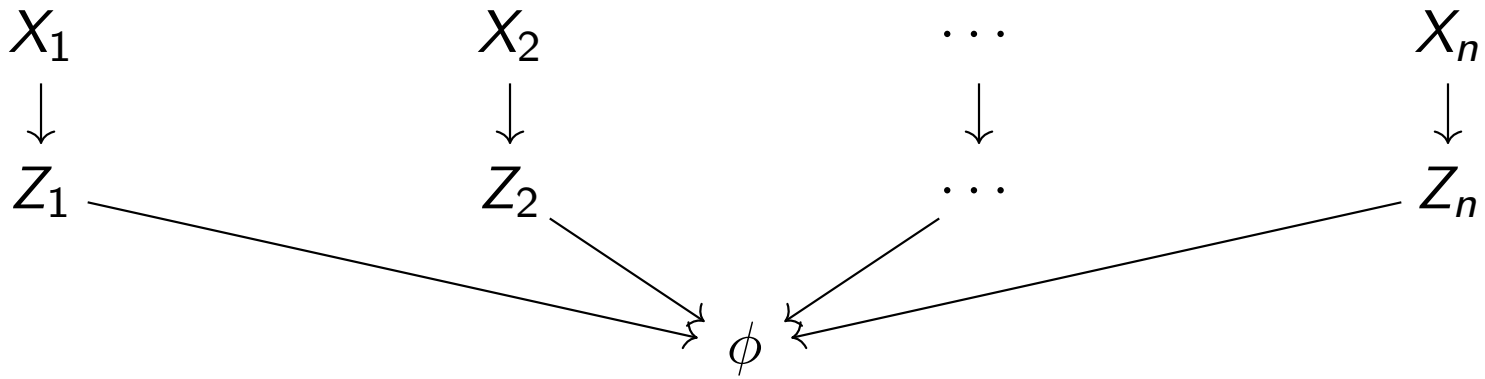

In GoF we see that *sequentially interactive* methods can do better.

# Uniformity testing

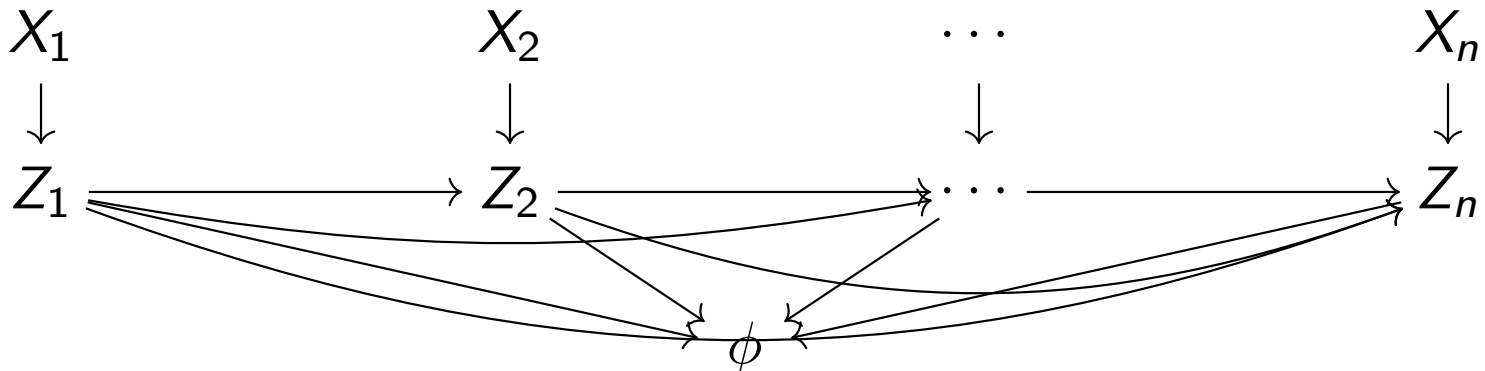
Recall that when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ on $[d]$ we want to test

$$H_0 : P = \mathrm{Unif}([d]) \quad \text{vs.} \quad H_1(\rho) : \mathrm{TV}(P, \mathrm{Unif}([d])) \geq \rho.$$

For simple hypothesis testing a *non-interactive* method was optimal.




In GoF we see that *sequentially interactive* methods can do better.



# Non-interactive rates


**Theorem (Acharya et al. 2019)**

*Among non-interactive $\alpha$-LDP ($\alpha \leq 1$) procedures, we have*

$$\rho_{n,\alpha}^{\mathrm{NI},*} \asymp \frac{d^{3/4}}{\sqrt{n\alpha^2}}.$$

We can compare this to the classical rate of $d^{1/4}/\sqrt{n}$:

- Effective sample size $n \mapsto n\alpha^2$.
- Effect of dimension worsened $d \mapsto d^3$.

# Upper bound methodology

We can use the RAPPOR mechanism (Erlingsson, Pihur & Korolova, 2014) to adapt (roughly speaking) the previous non-private test (Acharya et al., 2019).

Independently across individuals and coordinates, calculate

$$Z_{ij} = \begin{cases} \mathbb{1}_{\{X_i=j\}}, & \text{w.pr. } e^{\alpha/2}/(1+e^{\alpha/2}), \\ 1 - \mathbb{1}_{\{X_i=j\}}, & \text{otherwise,} \end{cases}$$

i.e. one-hot encoding then independent randomised responses.

This is $\alpha$-LDP: for $z \in \{0,1\}^d$ and $x \neq x' \in [d]$ we have

$$\frac{\mathbb{P}(Z_1 = z | X_1 = x)}{\mathbb{P}(Z_1 = z | X_1 = x')} = \frac{\mathbb{P}(Z_{1x} = z_x | X_1 = x)}{\mathbb{P}(Z_{1x} = z_x | X_1 = x')} \frac{\mathbb{P}(Z_{1x'} = z_{x'} | X_1 = x)}{\mathbb{P}(Z_{1x'} = z_{x'} | X_1 = x')}$$
$$\leq e^{\alpha/2} \times e^{\alpha/2} = e^{\alpha}.$$

# Upper bound methodology

We can use the RAPPOR mechanism (Erlingsson, Pihur & Korolova, 2014) to adapt (roughly speaking) the previous non-private test (Acharya et al., 2019).

Independently across individuals and coordinates, calculate

$$Z_{ij} = \begin{cases} \mathbb{1}_{\{X_i=j\}}, & \text{w.pr. } e^{\alpha/2}/(1+e^{\alpha/2}), \\ 1-\mathbb{1}_{\{X_i=j\}}, & \text{otherwise,} \end{cases}$$

i.e. one-hot encoding then independent randomised responses.

This is $\alpha$-LDP: for $z \in \{0,1\}^d$ and $x \neq x' \in [d]$ we have

$$\begin{aligned} \frac{\mathbb{P}(Z_1 = z | X_1 = x)}{\mathbb{P}(Z_1 = z | X_1 = x')} &= \frac{\mathbb{P}(Z_{1x} = z_x | X_1 = x)}{\mathbb{P}(Z_{1x} = z_x | X_1 = x')} \frac{\mathbb{P}(Z_{1x'} = z_{x'} | X_1 = x)}{\mathbb{P}(Z_{1x'} = z_{x'} | X_1 = x')} \\ &\leq e^{\alpha/2} \times e^{\alpha/2} = e^{\alpha}. \end{aligned}$$

# Upper bound

Easy to check that

$$\mathbb{E}(Z_{ij}) = \frac{1}{1+e^{\alpha/2}} + \frac{e^{\alpha/2}-1}{e^{\alpha/2}-1} p(j).$$

We therefore consider the $U$-statistic

$$T = \frac{1}{\binom{n}{2}} \sum_{i_1<i_2} \sum_{j=1}^{d} \left( Z_{i_1 j} - \frac{1}{1+e^{\alpha/2}} \right) \left( Z_{i_2 j} - \frac{1}{1+e^{\alpha/2}} \right) - \left( \frac{e^{\alpha/2}-1}{e^{\alpha/2}-1} \right)^2 \frac{1}{d},$$

which has $\mathbb{E}_P(T) = (\frac{e^{\alpha/2}-1}{e^{\alpha/2}-1}) \|p - p_0\|_2^2$. We also have

$$\mathrm{Var}_{P_0}(T) \lesssim \frac{d}{n^2}.$$

Calculations similar to those in the first lecture reveal that we have a powerful test when $\|p - p_0\|_2^2 \gtrsim \sqrt{d}/(n\alpha^2)$, or $\|p - p_0\|_1 \gtrsim d^{3/4}/\sqrt{n\alpha^2}$.

# Upper bound

Easy to check that

$$\mathbb{E}(Z_{ij}) = \frac{1}{1+e^{\alpha/2}} + \frac{e^{\alpha/2}-1}{e^{\alpha/2}-1} p(j).$$

We therefore consider the $U$-statistic

$$T = \frac{1}{\binom{n}{2}} \sum_{i_1<i_2} \sum_{j=1}^{d} \left( Z_{i_1 j} - \frac{1}{1+e^{\alpha/2}} \right) \left( Z_{i_2 j} - \frac{1}{1+e^{\alpha/2}} \right) - \left( \frac{e^{\alpha/2}-1}{e^{\alpha/2}-1} \right)^2 \frac{1}{d},$$

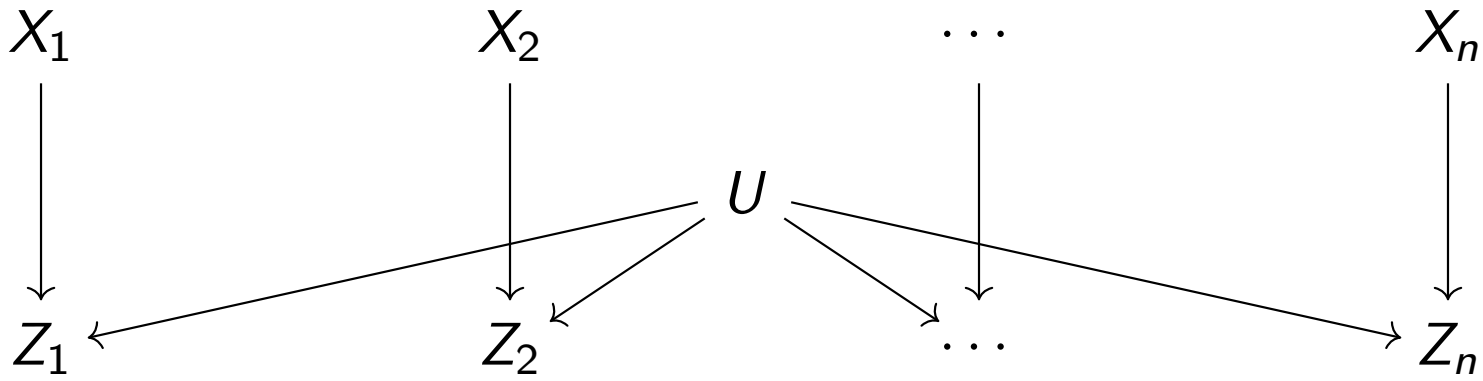
which has $\mathbb{E}_P(T) = (\frac{e^{\alpha/2}-1}{e^{\alpha/2}-1}) \|p - p_0\|_2^2$. We also have

$$\mathrm{Var}_{P_0}(T) \lesssim \frac{d}{n^2}.$$

Calculations similar to those in the first lecture reveal that we have a powerful test when $\|p - p_0\|_2^2 \gtrsim \sqrt{d}/(n\alpha^2)$, or $\|p - p_0\|_1 \gtrsim d^{3/4}/\sqrt{n\alpha^2}$.

# Public-coin mechanism

Surprisingly, we can do much better with very limited interaction.




**Theorem (Acharya et al. 2019)**

*Among public-coin $\alpha$-LDP ($\alpha \le 1$) procedures, we have*

$$\rho^{\mathrm{PC},*}_{n,\alpha} \asymp \frac{\sqrt{d}}{\sqrt{n\alpha^2}}.$$

This rate cannot be improved by allowing more general sequentially interactive mechanisms (Amin et al., 2020).

# Public-coin method

Uses the RAPTOR mechanism (Acharya et al., 2019): generate $S \subset [d]$ with $|S| = d/2$ uniformly at random, then use randomised response

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in S\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in S\}}, & \text{otherwise.} \end{cases}$$

Clearly $\mathbb{E}(Z|S) = \frac{1}{1+e^\alpha} + \frac{e^\alpha - 1}{e^\alpha + 1} p(S)$, so we can reject $H_0$ if

$$\left| \bar{Z}_n - \frac{1}{1+e^\alpha} - \frac{1}{2}\frac{e^\alpha - 1}{e^\alpha + 1} \right| \geq \frac{C}{\sqrt{n}}$$

some suitable constant $C$. This has power if $|p(S) - 1/2| \gtrsim (n\alpha^2)^{-1/2}$.

**Lemma (Acharya et al. 2019)**

*We have*

$$\mathbb{P}\left( \left| p(S) - \frac{1}{2} \right| > \frac{\mathrm{TV}(P, P_0)}{\sqrt{5d}} \right) > \frac{1}{477}.$$

If $\mathrm{TV}(P, P_0) \gtrsim \frac{\sqrt{d}}{\sqrt{n\alpha^2}}$ there is a constant probability of having power. By boosting, this is enough to get the rate.

# Public-coin method

Uses the RAPTOR mechanism (Acharya et al., 2019): generate $S \subset [d]$ with $|S| = d/2$ uniformly at random, then use randomised response

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in S\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in S\}}, & \text{otherwise.} \end{cases}$$

Clearly $\mathbb{E}(Z|S) = \frac{1}{1+e^\alpha} + \frac{e^\alpha - 1}{e^\alpha + 1} p(S)$, so we can reject $H_0$ if

$$\left| \bar{Z}_n - \frac{1}{1+e^\alpha} - \frac{1}{2}\frac{e^\alpha - 1}{e^\alpha + 1} \right| \geq \frac{C}{\sqrt{n}}$$

some suitable constant $C$. This has power if $|p(S) - 1/2| \gtrsim (n\alpha^2)^{-1/2}$.

**Lemma (Acharya et al. 2019)**

*We have*

$$\mathbb{P}\left( \left| p(S) - \frac{1}{2} \right| > \frac{\mathrm{TV}(P, P_0)}{\sqrt{5d}} \right) > \frac{1}{477}.$$

If $\mathrm{TV}(P, P_0) \gtrsim \frac{\sqrt{d}}{\sqrt{n\alpha^2}}$ there is a constant probability of having power. By boosting, this is enough to get the rate.

# Public-coin method

Uses the RAPTOR mechanism (Acharya et al., 2019): generate $S \subset [d]$ with $|S| = d/2$ uniformly at random, then use randomised response

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in S\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in S\}}, & \text{otherwise.} \end{cases}$$

Clearly $\mathbb{E}(Z|S) = \frac{1}{1+e^\alpha} + \frac{e^\alpha - 1}{e^\alpha + 1} p(S)$, so we can reject $H_0$ if

$$\left| \bar{Z}_n - \frac{1}{1+e^\alpha} - \frac{1}{2}\frac{e^\alpha - 1}{e^\alpha + 1} \right| \geq \frac{C}{\sqrt{n}}$$

some suitable constant $C$. This has power if $|p(S) - 1/2| \gtrsim (n\alpha^2)^{-1/2}$.

**Lemma (Acharya et al. 2019)**

*We have*

$$\mathbb{P}\left( \left| p(S) - \frac{1}{2} \right| > \frac{\mathrm{TV}(P, P_0)}{\sqrt{5d}} \right) > \frac{1}{477}.$$

If $\mathrm{TV}(P, P_0) \gtrsim \frac{\sqrt{d}}{\sqrt{n\alpha^2}}$ there is a constant probability of having power. By boosting, this is enough to get the rate.

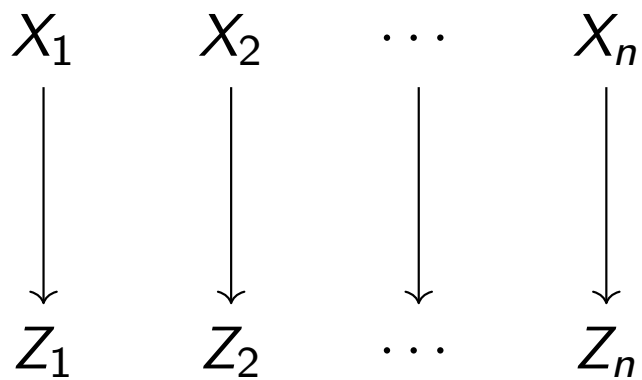
# Summary of uniformity testing

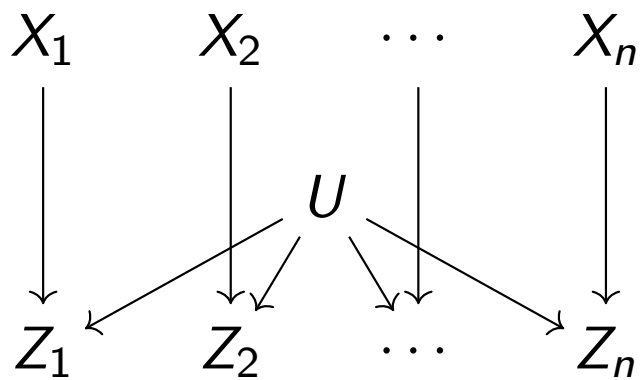
Recall that when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ on $[d]$ we have been testing

$$H_0 : P = \mathrm{Unif}([d]) \quad \text{vs.} \quad H_1(\rho) : \mathrm{TV}(P, \mathrm{Unif}([d])) \geq \rho.$$

We see costs of privacy and lack of interaction. Optimal separation rates are

- $\frac{d^{1/4}}{\sqrt{n}}$ without privacy constraints
- $\frac{\sqrt{d}}{\sqrt{n\alpha^2}}$ with public-coin $\alpha$-LDP
- $\frac{d^{3/4}}{\sqrt{n\alpha^2}}$ with non-interactive $\alpha$-LDP.



# Instance optimality – non-interactive rates

We get similar phenomena with non-uniform nulls. Recall that in the classical model the optimal separation rate is like $\sqrt{\|p_0\|_{2/3}/n}$.


**Theorem (B. & Butucea 2020)**

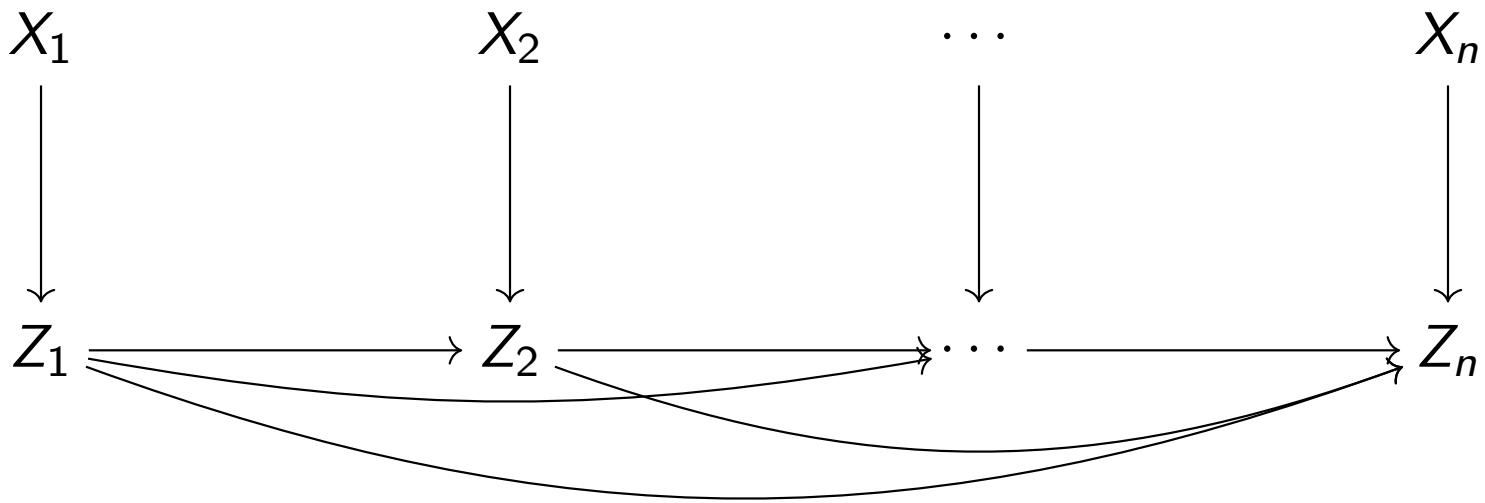
$$\rho^{\mathrm{NI},*}_{n,\alpha}(p_0) \lesssim j_*^{3/4}/\sqrt{n\alpha^2},$$

with nearly matching lower bound. Here $j_*$ is an 'effective support size'

$$j_* = j_*(n\alpha^2, p_0, ) := \min\left\{ j \in \mathbb{N} : \frac{j^{3/4}}{(n\alpha^2)^{1/2}} \geq \sum_{j'=j+1}^{\infty} p_0(j') \right\}.$$

# Interactive rates

Going back to the interactive setting:




**Theorem (B. & Butucea 2020)**

$$\rho^{\mathrm{I},*}_{n,\alpha}(p_0) \lesssim \tilde{j}^{1/4}/\sqrt{n\alpha^2},$$

with (nearly) matching lower bound, where $\tilde{j}$ is another 'effective support size'.

# Examples

Table: Separation rates (up to log factors) for testing discrete distributions on $\mathbb{N}$.

| $p_0$ | Non-interactive | Interactive | Non-private |
|---|---|---|---|
| Unif$[d]$ | $\frac{d^{3/4}}{\sqrt{n\alpha^2}}$ | $\frac{d^{1/2}}{\sqrt{n\alpha^2}}$ | $\frac{d^{1/4}}{\sqrt{n}}$ |
| $\propto j^{-1-\beta}$ | $(n\alpha^2)^{-\frac{2\beta}{4\beta+3}}$ | $(n\alpha^2)^{-\frac{2\beta}{4\beta+2}}$ | $n^{-1/2} \vee n^{-\frac{2\beta}{2\beta+1}}$ |

We can also extend to continuous data (Dubois et al., 2022).

# Conclusion

With minimax bounds we can quantify the *cost of privacy* in terms of rates of convergence.

We see a reduction in the effective sample size and typically worse dimension dependence.

By considering simple hypothesis testing we see links between robust statistics and LDP (also in mean/median estimation, density estimation...)

With more complex problems there can be a gap between non-interactive and sequentially interactive rates.

# References


Acharya, J., Canonne, C. L., Freitag, C. & Tyagi, H. (2019a) Test without trust: Optimal locally private distribution testing. *AISTATS*.

Acharya, J., Canonne, C. L., Sun, Z. & Tyagi, H. (2024) Unified lower bounds for interactive high-dimensional estimation under information constraints. *NeurIPS*.

Albert, M., Laurent, B., Marrel, A. & Meynaoui, A. (2022) Adaptive test of independence based on HSIC measures. *Ann. Statist.*, **50**, 858–879.

Amin, K., Joseph, M., & Mao, J. (2020) Pan-private uniformity testing. *COLT*.

Arias-Castro, E., Pelletier, B. & Saligrama, V. (2018) Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *J. Nonparametric Stat.*, **30**, 448–471.

Avella-Medina, M. (2021) Privacy-preserving parametric inference: a case for robust statistics. *J. Amer. Statist. Assoc.*, **116**, 969–983.

Balakrishnan, S. & Wasserman, L. (2019) Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.*, **47**, 1893–1927.

Berrett, B. & Butucea, C. (2020) Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *NeurIPS*.

Berrett, B. & Samworth, R. J. (2019) Nonparametric independence testing via mutual information. *Biometrika*, **106**, 547–566.

Berrett, T. B., Kontoyiannis, I. & Samworth, R. J. (2021) Optimal rates for independence testing via $U$-statistic permutation tests. *Ann. Statist.*, **49**, 2457–2490.

# References


Berrett, T. B., Wang, Y., Barber, R. F. & Samworth, R. J. (2020) The conditional permutation test for independence while controlling for confounders. *J. Roy. Statist. Soc. Ser. B*, **82**, 175–197.

Besag, J. & Clifford, P. (1989) Generalized Monte Carlo Significance Tests. *Biometrika*, **76**, 633–642.

Bordino, A. & B. (2025) Nonparametric inference for ratios of densities via uniformly valid and powerful permutation tests. *Available at* `arXiv:2505.24529`.

Cai, T. T., Wang, Y., & Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Ann. Statist.*, **49**(5), 2825–2850.

Candès, E., Fan, Y., Janson, L. & Lv, J. (2018) Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc. B*, **80**, 551–577.

Canonne, C. L. (2020) A survey on distribution testing: Your data is big. But is it blue? *Theory of Computing*, 1–100.

Chatterjee, S. (2021) A new coefficient of correlation. *J. Amer. Statist. Assoc.*, **116**, 2009–2022.

Chen, M., Gao, C. & Ren, Z. (2016) A general decision theory for Huber's $\epsilon$-contamination model. *Electron. J. Stat.*, **10**, 3752–3774.

Chhor, J. & Carpentier, A. (2022) Sharp local minimax rates for goodness-of-fit testing in multivariate binomial and Poisson families and in multinomials. *Mathematical Statistics & Learning*, **5**, 1–54.

Cohen, A. & Nassim, K. (2020) Towards formalizing the GDPR's notion of singling out. *PNAS*, **117**, 8344–8352.

# References


Cox, D. R. (1969) Some sampling problems in technology. *New Developments in Survey Sampling*, **1**, 506–527.

Dawid, P. A. (1979) Conditional independence in statistical theory. *J. R. Statist. Soc. B.*, **41**, 1–15.

Deb, N. & Sen, B. (2023) Multivariate rank-based distribution-free nonparametric testing using measure transportation. *J. Amer. Statist. Assoc.*, **118**, 192–207.

Devroye, L. & Lugosi, G. (2001) *Combinatorial Methods in Density Estimation*. Springer Science & Business Media.

Diakonikolas, I. & Kane, D. M. (2016) A new approach for testing properties of discrete distributions. *IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*

Ding, B., Kulkarni, J., & Yekhanin S. (2017) Collecting telemetry data privately. *NeurIPS*.

Donoho, D. L. & Liu, R. C. (1991) Geometrizing rates of convergence, III. *Ann. Statist.*, 668–701.

Dubois, A., B. & Butucea, C. (2022) Goodness-of-fit testing for Hölder continuous densities under local differential privacy. *Foundations of Modern Statistics – Festschrift in Honor of Vladimir Spokoiny*

Duchi, J. C., Jordan, M. I. & Wainwright, M. J. (2018) Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.*, **113**, 182–201.

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 265–284.

# References


Dwork, C. & Lei, J. (2009) Differential privacy and robust statistics. *Annual ACM Symposium on Theory of Computing*, 371–380.

Dwork, C. (2019) Differential privacy and the US census. *PODS*.

Efromovich, S. (2004) Density estimation for biased data. *Ann. Statist.*, **32**, 1137–1161.

Erlingsson, U., Pihur, V. & Korolova, A. (2014) Rappor: Randomized aggregatable privacy-preserving ordinal response. *Proc. 2014 ACM SIGSAC conference on computer and communications security*, 1054–1067.

Fisher, R. A. (1935) *The Design of Experiments* (1st Ed.). Oliver and Boyd, Edinburgh.

Gopi, S., Kamath, G., Kulkarni, J., Nikolov, A., Wu, Z. S. & Zhang, H. (2020) Locally private hypothesis selection. *COLT*.

Gretton A., Bousquet O., Smola A. & Schölkopf B. (2005) Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Algorithmic Learning Theory*, 63–77.

Hoeffding, W. (1948) A non-parametric test of independence. *Ann. Math. Statist.*, **19**, 546–57.

Hoeffding, W. (1952) The Large-Sample Power of Tests Based on Permutations of Observations. *Ann. Math. Statist.*, **23**, 169 – 192.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F. & Craig, D. W. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, **4**, e1000167.

Ingster, Y. I. (1989) Asymptotic minimax testing of independence hypothesis. *J. Sov. Math.*, **44**, 466–476.

# References


Janssen, A. (2000) Global power functions of goodness of fit tests. *Ann. Statist.*, **28**, 239–253.

Kairouz, P., Oh, S. & Viswanath, P. (2014). Extremal mechanisms for local differential privacy. *NeurIPS*.

Kendall, M. G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.

Kim, I., Balakrishnan, S. & Wasserman, L. (2022) Minimax optimality of permutation tests. *Ann. Statist.*, **50**, 225–251.

Lehmann, E. L. & Romano, J. P. (2005) *Testing Statistical Hypotheses*. New York, NY: Springer New York.

Li, M., B. & Yu, Y. (2023) On robustness and local differential privacy. *Ann. Statist.*, **51**, 717–737.

Near, J. (2018) Differential privacy at scale: Uber and Berkeley collaboration. *Enigma 2018*.

Paninski, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, **54**(10), 4750–4755.

Pearson, K. (1920) Notes on the history of correlation. *Biometrika*, **13**, 25–45.

Pensia, A., Jog, V. & Loh, P. L. (2024) The sample complexity of simple binary hypothesis testing. *COLT*

Pitman, E. J. G. (1938) Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, **29**, 322–335.

Pitsch, W. & Emrich, E. (2012) The frequency of doping in elite sports: Results of a replication study. *Int. Rev. Sport Sociol.*, **47**, 559–580.

Qin, J. (1993) Empirical likelihood in biased sample problems. *Ann. Statist.*, **21**, 1182–1196.

# References


Ramdas, A., Barber, R. F., Candès, E. J. & Tibshirani, R. J. (2022) Permutation Tests Using Arbitrary Permutation Distributions. *Sankhya A*, **85**, 1156 – 1177.

Rohde, A. & Steinberger, L. (2020). Geometrizing rates of convergence under local differential privacy constraints. *Ann. Statist.*, **48**, 2646–2670.

Serfling, R. J. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

Shah, R. D. & Peters, J. (2019) The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.*, **48**, 1514–1538.

Sweeney, L. (2002) k-anonymity: A model for protecting privacy. *Fuzziness and Knowledge-Based Systems*, **10**, 557–570.

Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

Tang J., Korolova, A., Bai, X., Wang, X. & Wang X. (2017) Privacy loss in Apple's implementation of differential privacy on macOS 10.12. *Available at* `arXiv:1709.02753`.

Tibshirani, R. J., Barber, R. F., Candès, E. & Ramdas, A. (2019). Conformal prediction under covariate shift. *NeurIPS*.

Valiant, G. & Valiant, P. (2017) An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, **46**, 429–455.

Warner, S. L. (1965) Randomized sesponse: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, **60**, 63–69.

Wasserman, L. & Zhou, S. (2010) A statistical framework for differential privacy. *J. Amer. Statist. Assoc.*, **105**, 375–389.