The Permutation Test

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Hypothesis testing

1. Choose a conjecture
2. Choose null hypothesis
3. Choose test statistic
4. Decide significance level α
5. Gather data and compute test statistic
6. Compute p value
7. Reject the null hypothesis if p value $\leq \alpha$

P value

Probability of observing larger or equal test statistic under null hypothesis

We need to know the distribution!

What if we don't?

Price of burgers

Conjecture: Burgers in NY are more expensive than in Madrid

Null hypothesis: Same distribution in both cities

Test statistic: Average in NY - Average in Madrid

Data

New York	New York	Madrid	Madrid
16	18	13	13

$$
\begin{aligned}
t_{\text{data}} &= m(\text{NY}) - m(\text{Madrid}) \\
&= \frac{16+18}{2} - \frac{13+13}{2} = 4
\end{aligned}
$$

Is this sufficient evidence against null hypothesis?

We need a p value!

Goal: Compute p value without parametric model for the test statistic

Key idea

If price distribution is the same, label is meaningless

New York	New York	Madrid	Madrid
16	18	13	13

Any permutation would be equally likely

New York	New York	Madrid	Madrid
13	18	13	16

Permutations

NY	NY	M	M	t
13	13	16	18	-4
13	13	18	16	-4
13	16	13	18	-1
13	16	18	13	-1
13	18	13	16	1
13	18	16	13	1
13	13	16	18	-4
13	13	18	16	-4
13	16	13	18	-1
13	16	18	13	-1
13	18	13	16	1
13	18	16	13	1

NY	NY	M	M	t
16	13	13	18	-1
16	13	18	13	-1
16	13	13	18	-1
16	13	18	13	-1
16	18	13	13	4
16	18	13	13	4
18	13	16	13	1
18	13	13	16	1
18	16	13	13	4
18	16	13	13	4
18	13	13	16	1
18	13	16	13	1

How many are larger or equal to $t_{\text{data}} = 4$?

Permutations

NY	NY	M	M	t
13	13	16	18	-4
13	13	18	16	-4
13	16	13	18	-1
13	16	18	13	-1
13	18	13	16	1
13	18	16	13	1
13	13	16	18	-4
13	13	18	16	-4
13	16	13	18	-1
13	16	18	13	-1
13	18	13	16	1
13	18	16	13	1

NY	NY	M	M	t
16	13	13	18	-1
16	13	18	13	-1
16	13	13	18	-1
16	13	18	13	-1
16	18	13	13	4
16	18	13	13	4
18	13	16	13	1
18	13	13	16	1
18	16	13	13	4
18	16	13	13	4
18	13	13	16	1
18	13	16	13	1

How many are larger or equal to $t_{\text{data}} = 4$? $4/24 = 16.7\%$

What have we computed?

Conditional *probability of observing larger or equal test statistic under null hypothesis* given that data are permutation of observed data

Sounds like a p value!

Multiset of permutations

For any $x \in \mathbb{R}^n$ Π_x is multiset of $n!$ permutations

$$x := \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$\Pi_x = \left\{ \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \begin{bmatrix} a \\ c \\ b \end{bmatrix}, \begin{bmatrix} b \\ a \\ c \end{bmatrix}, \begin{bmatrix} b \\ c \\ a \end{bmatrix}, \begin{bmatrix} c \\ a \\ b \end{bmatrix}, \begin{bmatrix} c \\ b \\ a \end{bmatrix} \right\}$$

P-value function of a permutation test

Observed data: $x_{\text{data}} \in \mathbb{R}^n$

Observed test statistic: $t_{\text{data}} := T(x_{\text{data}})$

Model for data under null hypothesis: random vector $\tilde{x}_{\text{null}}$

Test statistic under null hypothesis: $\tilde{t}_{\text{null}} := T(\tilde{x}_{\text{null}})$

P-value function

$$\text{pv}(t) := \mathrm{P}\left(\tilde{t}_{\text{null}} \geq t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)$$

Exchangeability

The entries of a $\tilde{x}$ are exchangeable if permuting them does not change the distribution of $\tilde{x}$

Π_x: multiset of permutations of x

The entries of a discrete random vector $\tilde{x}$ are exchangeable if

$$p_{\tilde{x}}(x) = p_{\tilde{x}}(v) \qquad \text{for all } v \in \Pi_x$$

The entries of a continuous random vector $\tilde{x}$ are exchangeable if

$$f_{\tilde{x}}(x) = f_{\tilde{x}}(v) \qquad \text{for all } v \in \Pi_x$$

I.i.d. random variables

If $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_d$ are i.i.d.

$$f_{\tilde{x}}(x) = \prod_{i=1}^{d} f_{\tilde{x}_i}(x_i) = \prod_{i=1}^{d} f_{\text{marg}}(x_i)$$

then they are exchangeable

For any $v \in \Pi_x$

$$f_{\tilde{x}}(v) = \prod_{i=1}^{d} f_{\text{marg}}(v_i) = \prod_{i=1}^{d} f_{\text{marg}}(x_i) = f_{\tilde{x}}(x)$$

Consequence of exchangeability

If entries of $\tilde{x}_{\text{null}}$ are exchangeable

$$\mathrm{P}\left(\tilde{x}_{\text{null}} = v \,\middle|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) = \frac{1}{n!} \quad \text{for any } v \in \Pi_x$$

For any $v_1, v_2 \in \Pi_{x_{\text{data}}}$

$$p_{\tilde{x}_{\text{null}}}(v_1) = p_{\tilde{x}_{\text{null}}}(v_2)$$

$$\begin{aligned}
\mathrm{P}\left(\tilde{x}_{\text{null}} = v_1 \,\middle|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) &= \frac{\mathrm{P}\left(\tilde{x}_{\text{null}} = v_1, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)}{\mathrm{P}\left(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)} \\
&= \frac{p_{\tilde{x}_{\text{null}}}(v_1)}{\mathrm{P}\left(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)} \\
&= \frac{p_{\tilde{x}_{\text{null}}}(v_2)}{\mathrm{P}\left(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)} \\
&= P\left(\tilde{x}_{\text{null}} = v_2 \,\middle|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)
\end{aligned}$$

Consequence of exchangeability

$P(\tilde{x}_{\text{null}} = v \mid \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}})$ is the same for all $v \in \Pi_{x_{\text{data}}}$

$$\begin{aligned}\sum_{v \in \Pi_x} \mathrm{P}(\tilde{x}_{\text{null}} = v \mid \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) \\ &= \mathrm{P}(\cup_{v \in \Pi_x} \tilde{x}_{\text{null}} = v \mid \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) \\ &= \mathrm{P}(\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}} \mid \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) \\ &= 1\end{aligned}$$

$$\mathrm{P}(\tilde{x}_{\text{null}} = v \mid \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) = \frac{1}{n!}$$

Nonparametric p-value function

$$
\begin{aligned}
\mathrm{pv}(t) &:= \mathrm{P}\left(T\left(\tilde{x}_{\text{null}}\right) \geq t \,\middle|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) \\
&= \mathrm{P}\left(\cup_{\left\{v \in \Pi_{x_{\text{data}}} : T(v) \geq t\right\}} \left\{\tilde{x}_{\text{null}} = v\right\} \,\middle|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) \\
&= \sum_{\left\{v \in \Pi_{x_{\text{data}}} : T(v) \geq t\right\}} \mathrm{P}\left(\tilde{x}_{\text{null}} = v \,\middle|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) \\
&= \frac{\sum_{v \in \Pi_{x_{\text{data}}}} 1\left(T(v) \geq t\right)}{n!}
\end{aligned}
$$

where $1\left(T(v) \geq t\right)$ equals 1 if $T(v) \geq t$ and 0 otherwise

Burgers

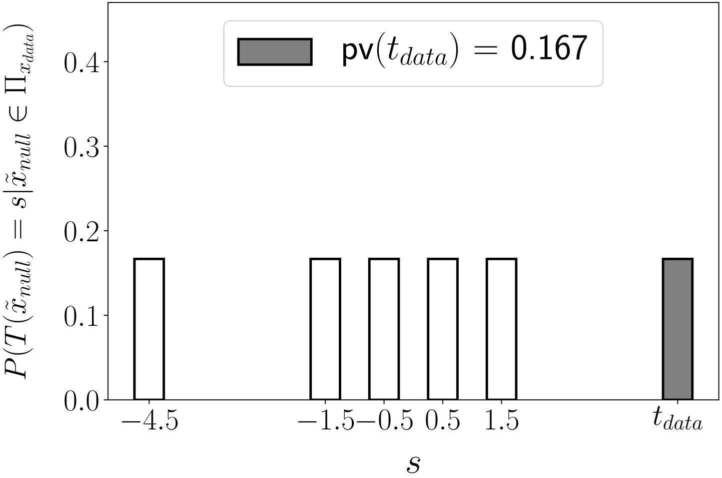
NY	NY	M	M	t
13	13	16	18	-4
13	13	18	16	-4
13	16	13	18	-1
13	16	18	13	-1
13	18	13	16	1
13	18	16	13	1
13	13	16	18	-4
13	13	18	16	-4
13	16	13	18	-1
13	16	18	13	-1
13	18	13	16	1
13	18	16	13	1

NY	NY	M	M	t
16	13	13	18	-1
16	13	18	13	-1
16	13	13	18	-1
16	13	18	13	-1
16	18	13	13	4
16	18	13	13	4
18	13	16	13	1
18	13	13	16	1
18	16	13	13	4
18	16	13	13	4
18	13	13	16	1
18	13	16	13	1

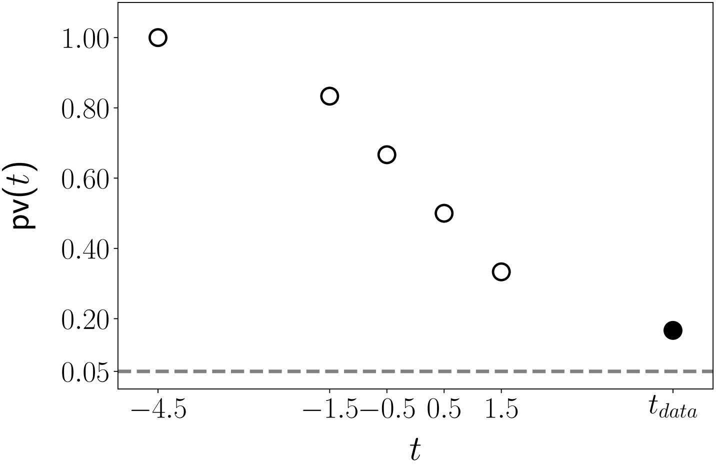
$$\mathrm{pv}(t) = \frac{\sum_{v \in \Pi_{x_{\mathrm{data}}}} 1\left(T(v) \geq t\right)}{n!}$$

$$= \frac{4}{24} = 0.167$$

Conditional pmf of test statistic



P-value function



$$\mathrm{P}\left(\text{False positive}\right) \leq \alpha$$

$$\tilde{t}_{\text{null}} := T(\tilde{x}_{\text{null}})$$

$$\begin{aligned} \mathrm{P}\left(\text{False positive} \,|\, \tilde{x}_{\text{null}} \in \Pi_x\right) &= \mathrm{P}\left(\text{pv}(\tilde{t}_{\text{null}}) \leq \alpha \,|\, \tilde{x}_{\text{null}} \in \Pi_x\right) \\ &= F_{\tilde{u}}(\alpha \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) \leq \alpha \end{aligned}$$

$$\text{pv}(t) := \mathrm{P}\left(\tilde{t}_{\text{null}} \geq t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right)$$

$$\begin{aligned} \tilde{u} &:= \text{pv}(\tilde{t}_{\text{null}}) \\ &= 1 - F_{\tilde{t}_{\text{null}}}\left(\tilde{t}_{\text{null}} \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) \end{aligned}$$

$$F_{\tilde{u}}(u \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}) \leq u$$

Antetokounmpo's free throws

Conjecture: Free throw percentage is higher at home than away

Null hypothesis: Percentage is the same

Test statistic:

$$\frac{\text{Made at home}}{\text{Attempted at home}} - \frac{\text{Made away}}{\text{Attempted away}}$$

Under null hypothesis, data are i.i.d. and hence exchangeable

Permutation test

Free throws: 44 at home and 41 away

$$x_{\text{data}} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 1 \end{bmatrix}$$

$$T(v) = \frac{1}{44} \sum_{i=1}^{44} v[i] - \frac{1}{41} \sum_{i=45}^{85} v[i]$$

$$\text{pv}(t) = \frac{\sum_{v \in \Pi_{x_{\text{data}}}} 1\left(T(v) \geq t\right)}{n!}$$

Problem

$85! > 10^{128}$

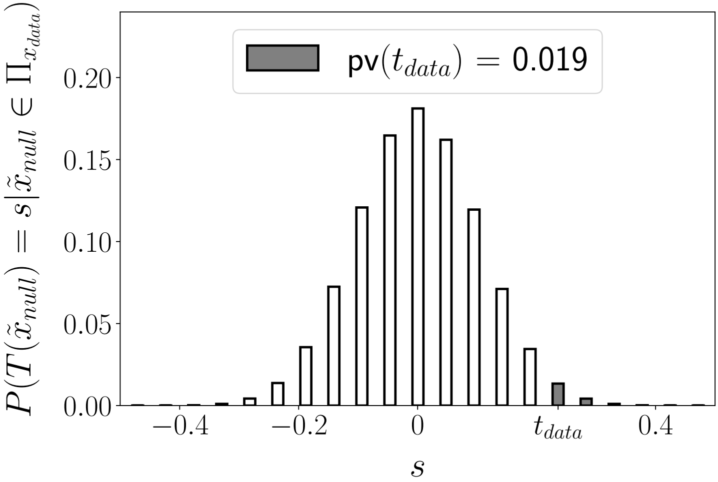
Solution: Monte Carlo estimation

Generate k independent permutations $v_1, \ldots, v_k \in \Pi_{x_{\text{data}}}$

Compute test statistics, $t_i = T(v_i)$, $1 \leq i \leq k$

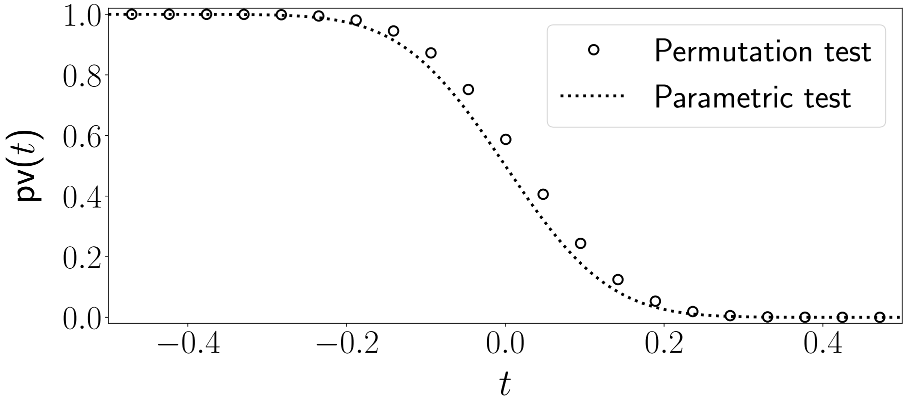
$$\text{pv}(t_{\text{data}}) \approx \frac{\sum_{i=1}^{k} 1\left(T(v_i) \geq t_{\text{data}}\right)}{k}$$

P-value



$\mathsf{pv}(t_{data}) = 0.019$
$P(T(\tilde{x}_{null}) = s|\tilde{x}_{null} \in \Pi_{x_{data}})$
0.20
0.15
0.10
0.05
0.00
-0.4
-0.2
0
t_{data}
0.4
s

P-value function



Grades

Conjecture: Grades from two schools have different distributions

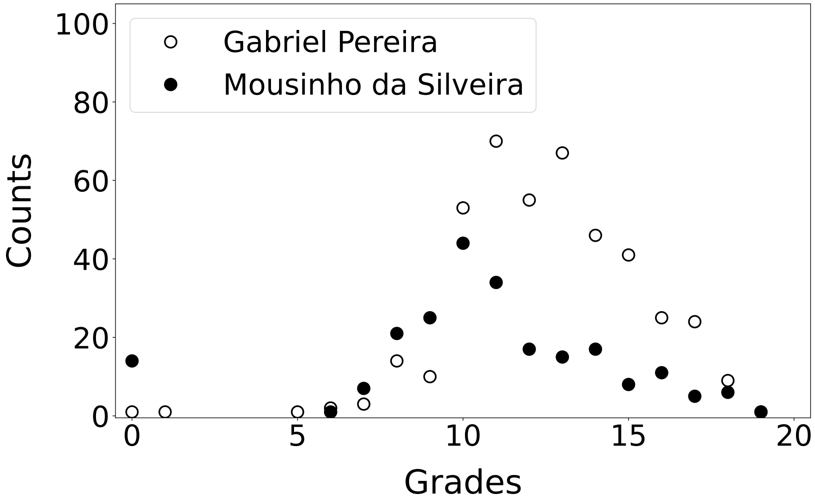
Null hypothesis: Distributions are the same

Test statistic: Difference of medians

Distribution of test statistic under null hypothesis?

¯_(ツ)_/¯

Grades from two schools in Portugal



Permutation test

Data $x = [x^A \; x^B]$

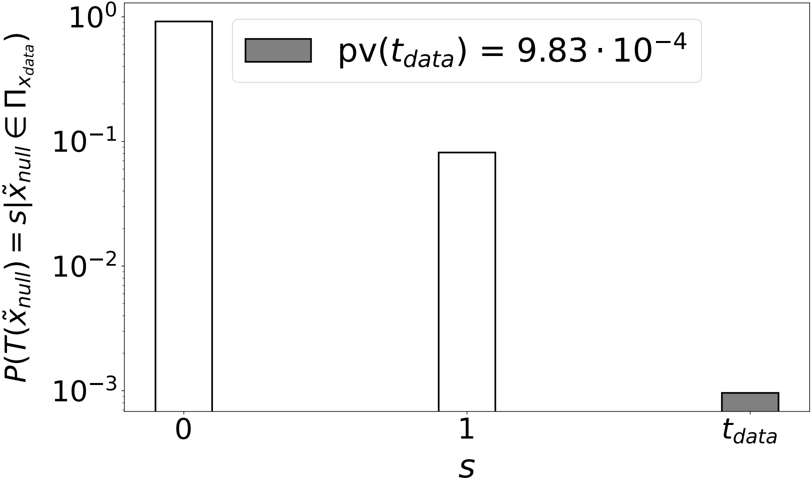
$$t_{\text{data}} = \left| \text{median}\left(x^A\right) - \text{median}\left(x^B\right) \right|$$

We generate $k := 10^6$ permutations $v_1, \ldots, v_k \in \Pi_{x_{\text{data}}}$

$$T(v_i) = \left| \text{median}\left(v_i^A\right) - \text{median}\left(v_i^B\right) \right|$$

$$\text{pv}(t_{\text{data}}) \approx \frac{\sum_{i=1}^{k} 1\left(T(v_i) \geq t_{\text{data}}\right)}{k}$$

One million permutations



What have we learned

The permutation test

P values can be computed without a parametric model