Overview of Hypothesis Testing

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

Goal: Determine whether data supports a conjecture

Play devil's advocate: *Maybe it's just chance*

If this is very unlikely $\implies$ Data supports conjecture

Plan

- Hypothesis-testing framework
- Statistical significance
- Multiple testing
- Hypothesis testing and causal inference
- Practical significance
- The power
- Nonparametric testing: The permutation test

Conjecture

Giannis Antetokounmpo's free throw percentage is different at home and away

Null hypothesis

Contradicts conjecture

Free throw percentage is the same at home and away

Original conjecture is the alternative hypothesis

Test statistic

Function of the data

Large value is evidence against null hypothesis

$$t_{\text{data}} := \frac{\text{Made at home}}{\text{Attempted at home}} - \frac{\text{Made away}}{\text{Attempted away}}$$

2021 NBA finals

$$t_{\text{data}} = \frac{34}{44} - \frac{22}{41} = 0.236$$

Evidence against null hypothesis?

P value

Probability of observing larger or equal test statistic under null hypothesis

Parametric testing

Distribution depends on a small number of parameters θ

Simple null hypothesis: Parameters equal single value $\theta = \theta_{\text{null}}$

Composite null hypothesis: Parameters belong to a set $\theta \in \Theta_{\text{null}}$

Two-sample z test

Data: $\tilde{x}_1, \ldots, \tilde{x}_n$

Two groups: A and B

One-tailed test statistic

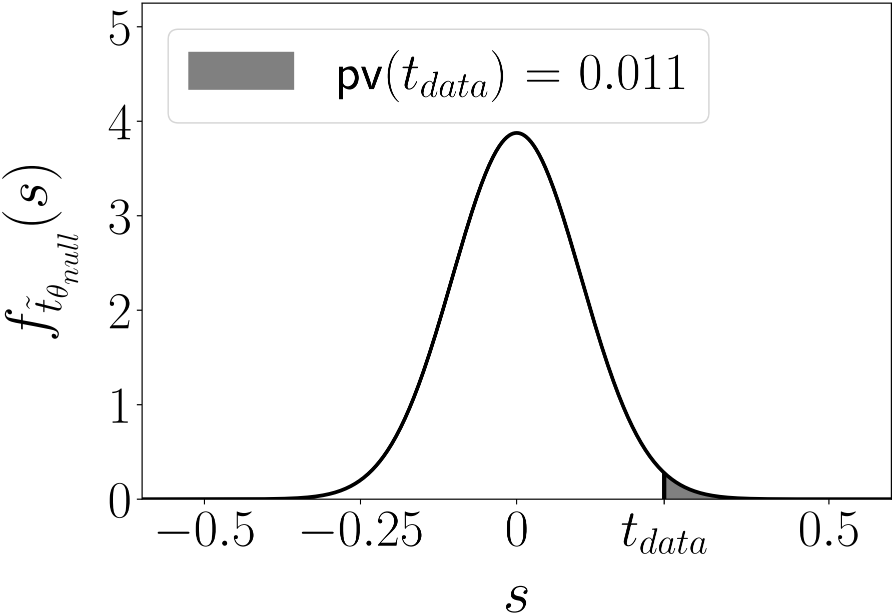
$$\tilde{t}_{\text{1-tail}} = \frac{1}{n_A} \sum_{i \in \mathcal{A}} \tilde{x}_i - \frac{1}{n_B} \sum_{i \in \mathcal{B}} \tilde{x}_i$$

Null hypothesis: All data are i.i.d. Bernoulli with parameter θ_{null}

$\approx$ Gaussian with mean 0 and variance

$$\sigma^2_{\text{null}} := \theta_{\text{null}}(1 - \theta_{\text{null}}) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)$$

One-tailed test

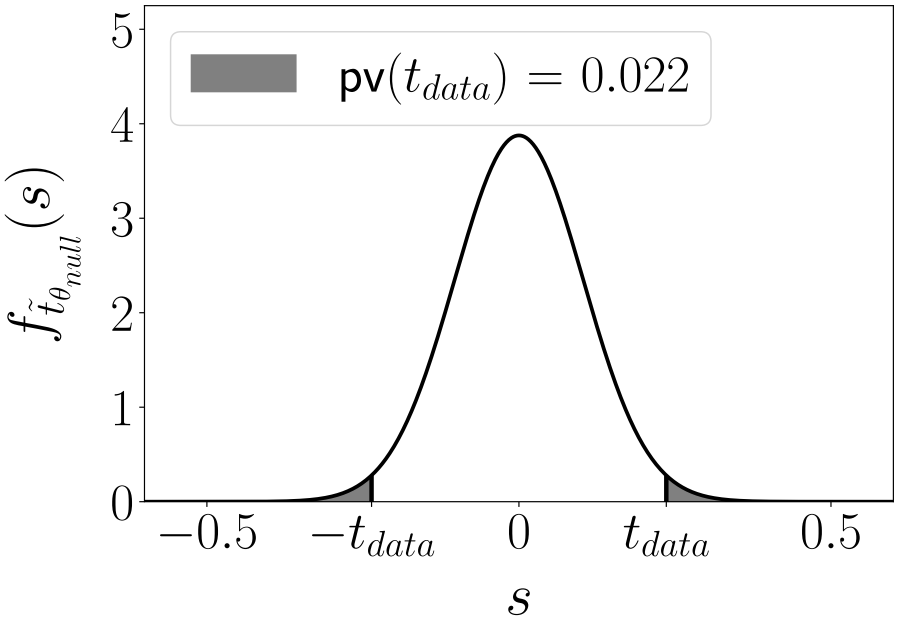


Two-tailed test

$$\tilde{t}_{\text{1-tail}} = \frac{1}{n_A} \sum_{i \in \mathcal{A}} \tilde{x}_i - \frac{1}{n_B} \sum_{i \in \mathcal{B}} \tilde{x}_i$$

$$\tilde{t}_{\text{2-tails}} = \left| \tilde{t}_{\text{1-tail}} \right|$$

Two-tailed test



Statistical significance

How do we decide whether p value is evidence against null hypothesis?

Fix significance level α beforehand

Reject null hypothesis if p value $\leq \alpha$

What can go wrong?

Type 1 error: False positive

Null hypothesis holds, but we reject it

Type 2 error: False negative

Null hypothesis does not hold, but we do not reject it

False positive

A false positive happens if

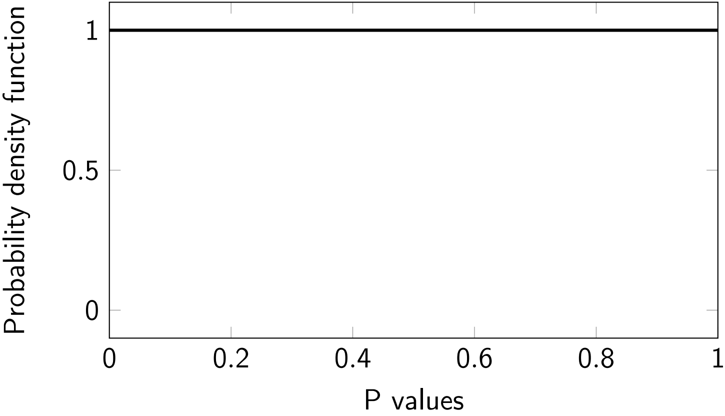
1. Null hypothesis holds
2. P value $\leq \alpha$

P value under null hypothesis? Uniformly distributed in $[0, 1]$!

$$\begin{aligned} \mathrm{P}\left(\text{False positive}\right) &= \mathrm{P}\left(\text{P value} \leq \alpha \text{ under null hypothesis}\right) \\ &= \mathrm{P}\left(\tilde{u} \leq \alpha\right) = \alpha \end{aligned}$$

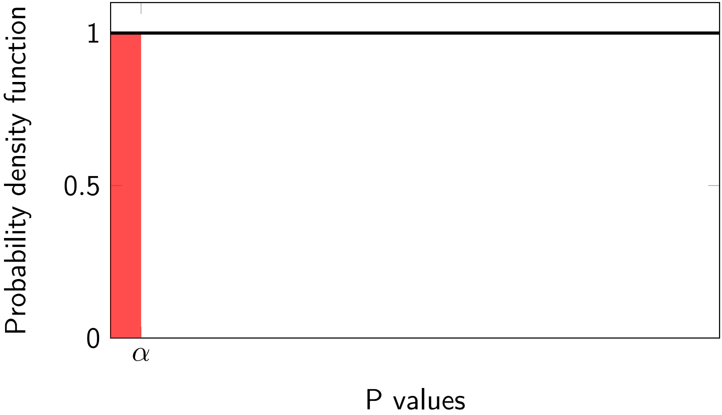
Small detour

Distribution of p value under simple null hypothesis for continuous test statistics



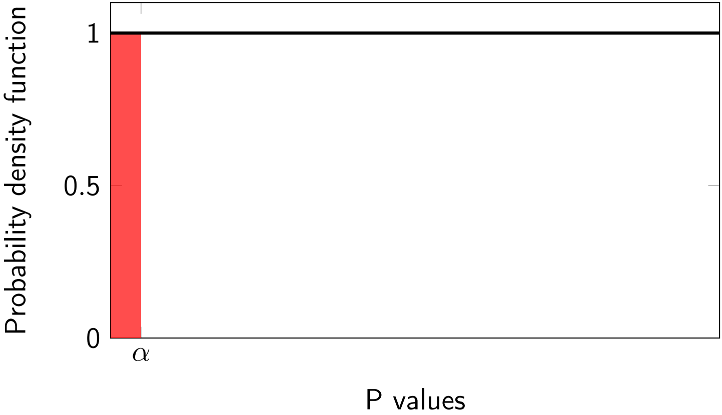
P-value distribution

Probability of a single false positive? α



P-value distribution

Under null hypothesis, fraction of false positives among many tests? α!



Pizza and COVID-19

100 studies to determine whether pizza cures COVID-19

$\approx$ 95 true negatives

$\approx$ 5 false positives

If all results are published no problem

Unfortunately, much easier to publish if result is statistically significant!

Publication bias: You only hear about the false positives!

Multiple testing

k independent hypothesis tests with significance level α

Probability of false positive in each test $= \alpha$

$$\begin{aligned} \mathrm{P}(\geq 1 \text{ false positive}) &= 1 - \mathrm{P}(\text{No false positives}) \\ &= 1 - (1-\alpha)^k \end{aligned}$$

For $\alpha := 0.05$ and $k := 100$, the probability is 0.99!

Solution? Decrease α

Bonferroni's correction

We reject null hypothesis if p value $\leq \tau := \alpha / k$

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

But increases false negatives!

More sophisticated approaches order by p value and accept a certain fraction of false positives

Back to the free throws

$\alpha := 0.05 \geq 0.011$ (or 0.022)

We reject the null hypothesis!

Does this mean taunts cause worse percentage? No!

Could be due to confounding factors

Causal inference

To identify causal effect, outcome and treatment must be independent

How can we achieve this? Randomizing the treatment

COVID-19 vaccine

43,448 patients randomly divided into

- Treatment group of 21,720 patients: 8 cases (0.037%)
- Control group of 21,728 patients: 162 (0.746%)

P value $< 10^{-23}$

Causal inference vs hypothesis testing

Causal inference and hypothesis testing have complementary roles

Is there a difference between control and treatment groups?

Yes, it is statistically significant by the hypothesis test

Is the difference due to a causal effect?

Yes, because the trial is randomized

Statistical vs practical significance

In large-scale trials, tiny differences can be statistically significant

Fictitious vaccine trial

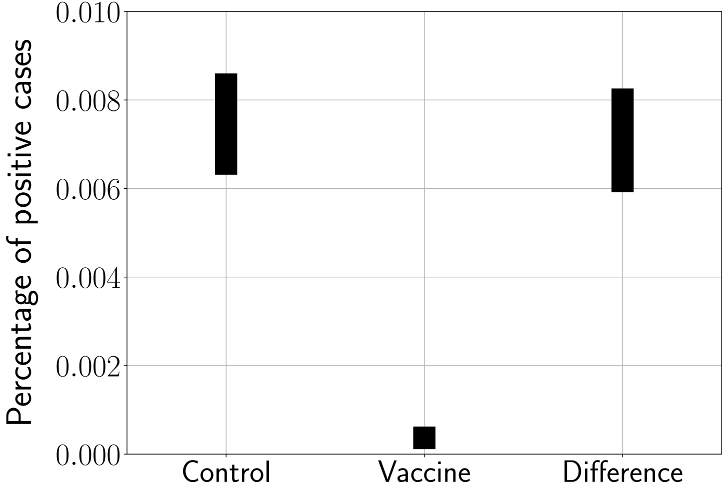
43,448 patients randomly divided into

- Treatment group of 21,720 patients: 120 cases (0.552%)
- Control group of 21,728 patients: 162 (0.746%)

$$\mathrm{pv}(t_{\mathrm{data}}) = 0.006$$

Ratio of positive cases is 3/4, not practically significant! (Real data: 1/20)

Actual vaccine trial



Is it enough to control false positives?

No, we also want to find true positives!

The power is the probability of a true positive

Parametric testing

Distribution of test statistic depends on parameters θ

Power function:

$$\text{pow}(\theta) := \mathrm{P}\left(\text{Rejecting the null hypothesis}\right)$$

Power function

Null hypothesis: $\theta \in \Theta_{\text{null}}$

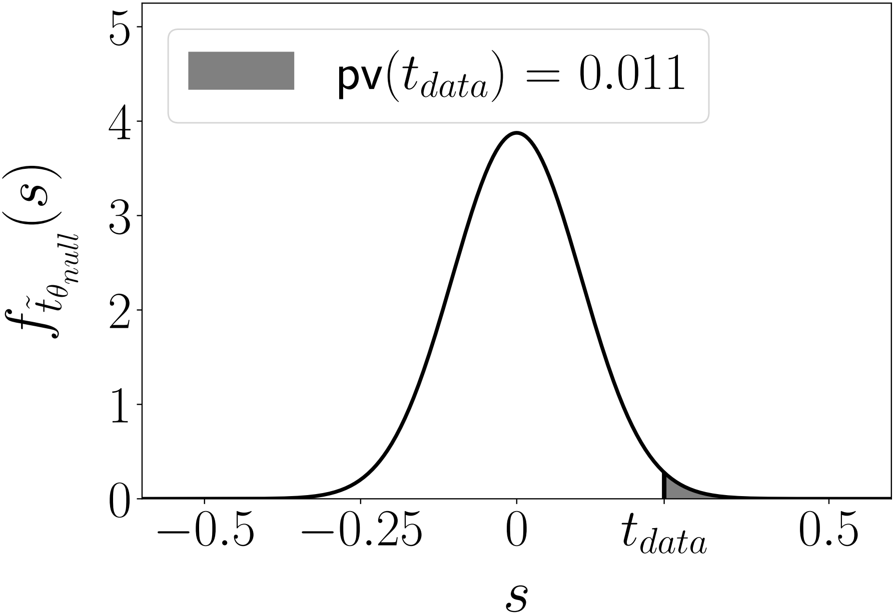
$$\text{pow}(\theta) = \mathrm{P}\left(\text{False positive}\right) \leq \alpha$$

Power function

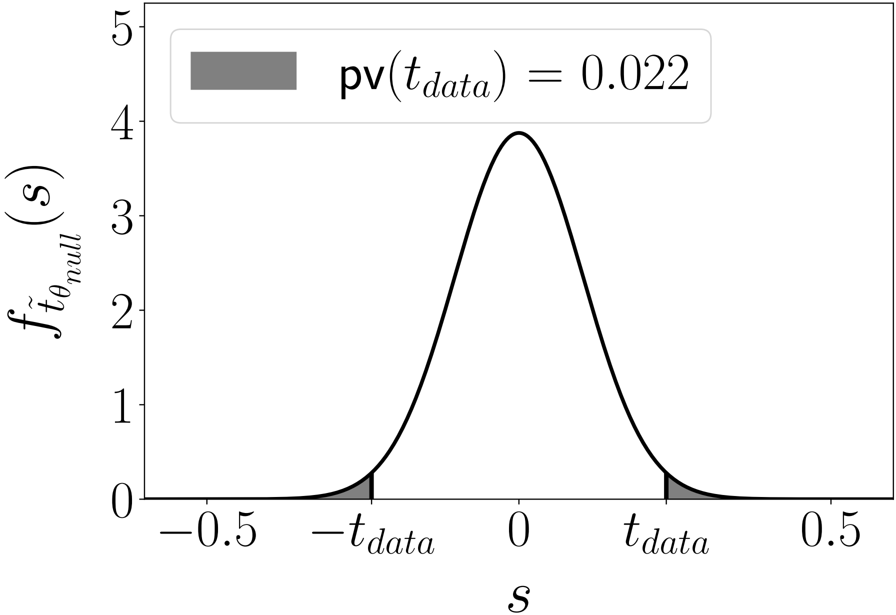
Alternative hypothesis: $\theta \in \Theta_{\text{alt}}$

$$\text{pow}(\theta) = \mathrm{P}\left(\text{True positive}\right)$$

One-tailed test



Two-tailed test



Power function

Parametric model:

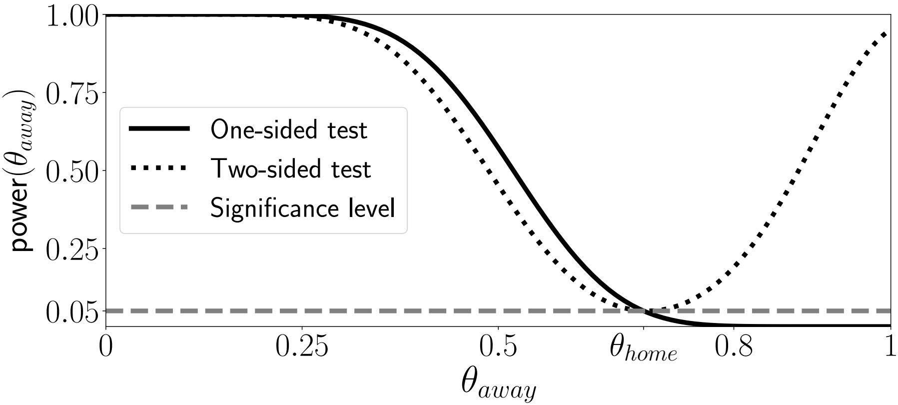
Home free-throw %: θ_{home}

Away free-throw %: θ_{away}

We fix $\theta_{\text{home}} := 0.685$ (season %)

Power function as a function of θ_{away}

Power function for fixed θ_{home}



Nonparametric testing

What if we don't have a model for the test statistic under the nulll hypothesis?

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

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

Key idea

If 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 $t_{\text{data}} = 4$?

This is a p value!

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

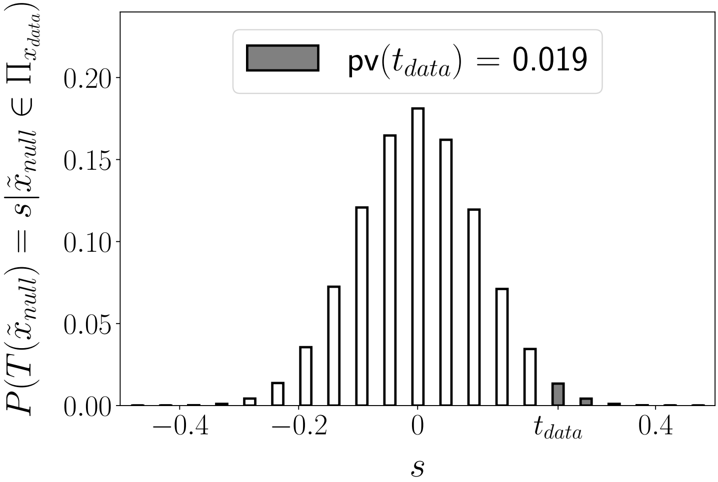
How many are larger or equal to $t_{\text{data}} = 4$? $4/24 = 16.7\%$

Problem

For Antetokounmpo's free throws, we have $n! = 85! > 10^{128}$

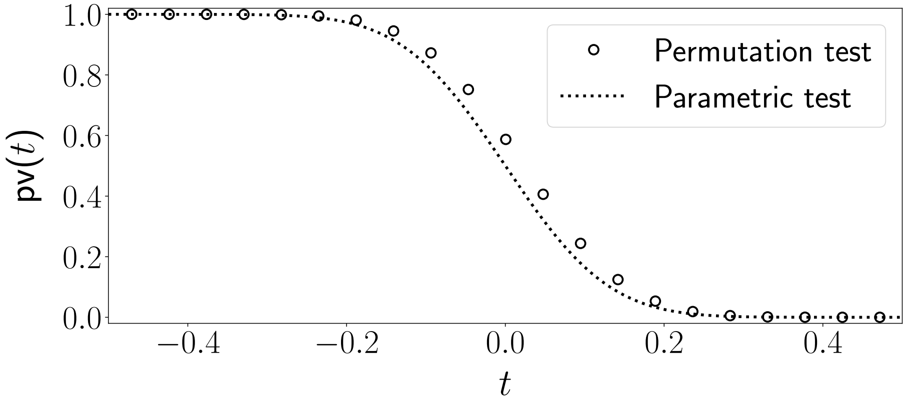
Solution: Sample many permutations (Monte Carlo estimation)

P-value



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

P-value function



What have we learned?

- Hypothesis-testing framework
- Statistical significance
- Multiple testing
- Hypothesis testing and causal inference
- Practical significance
- The power
- Nonparametric testing: The permutation test