Do UPPER CASE YouTube videos get more (or less) views?

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


Probability and Statistics
for Data Science
PROBABI AND
STATIS FOR
DATA SCIE CE

Plan

Data acquisition

Selection of statistical metric

Hypothesis testing

Data acquisition

Selection of statistical metric

Hypothesis testing

Causal inference

Key question: Do upper case titles cause more (or less) views

If upper case titles have more (or less) views, does that imply a causal effect?

No!

There could be confounders

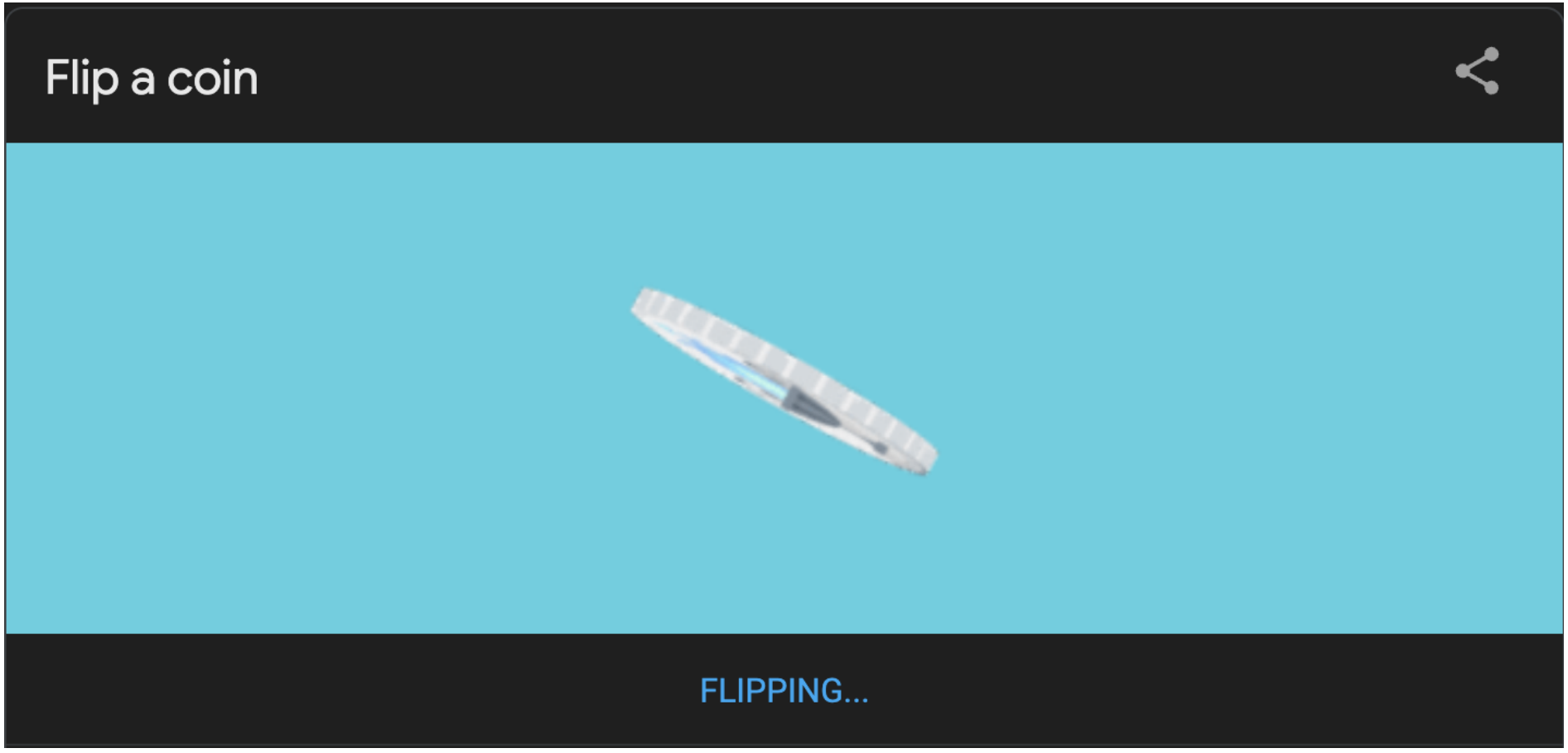
Upper case titles could correspond to:

- Older videos
- More exciting topics
- Prettier thumbnails

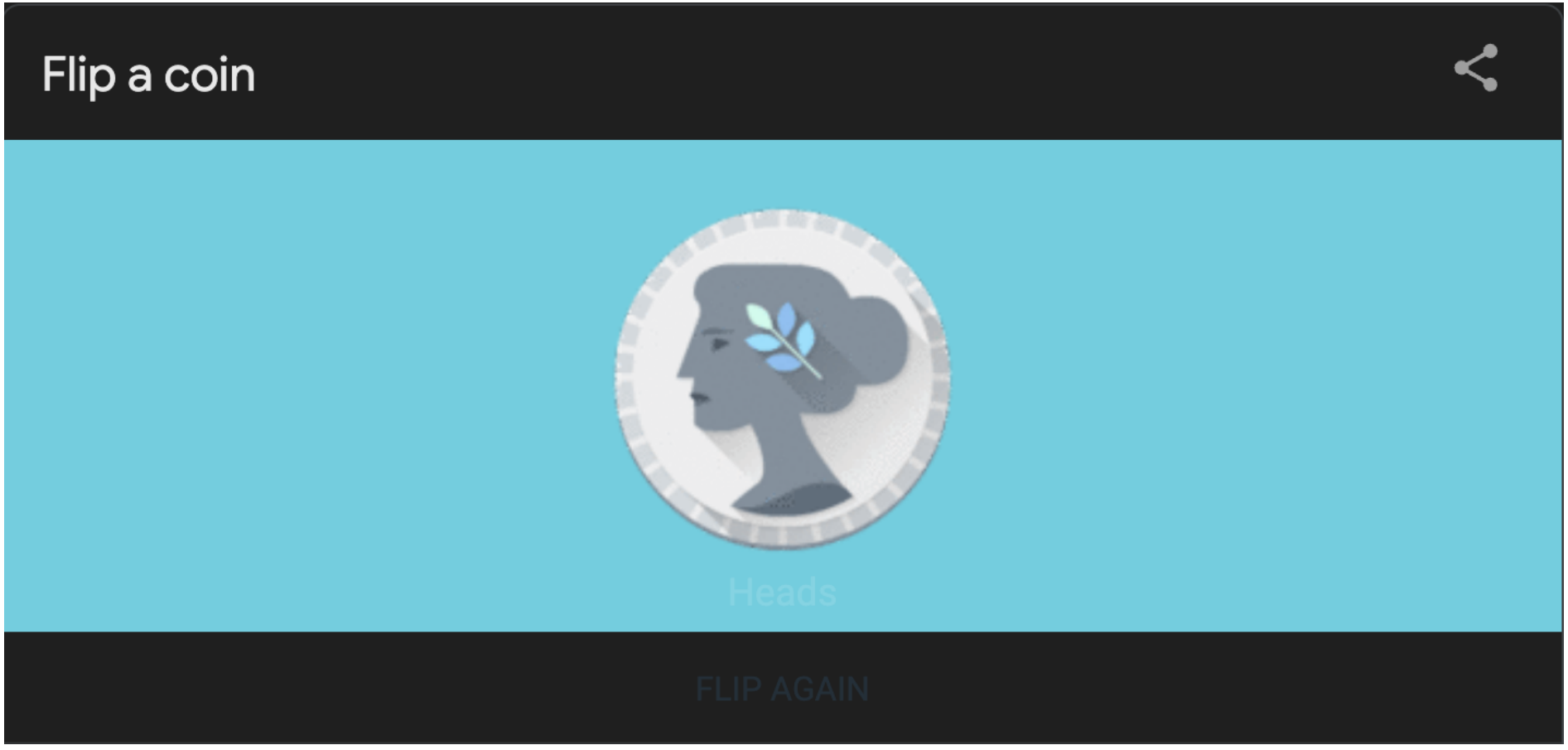
How can we avoid this?

Randomizing capitalization neutralizes confounders *even if they are* ***unknown*** *to us*!

Randomization



Randomization



Data acquisition

Selection of statistical metric

Hypothesis testing

What metric should we use?

Average treatment effect = Difference of means

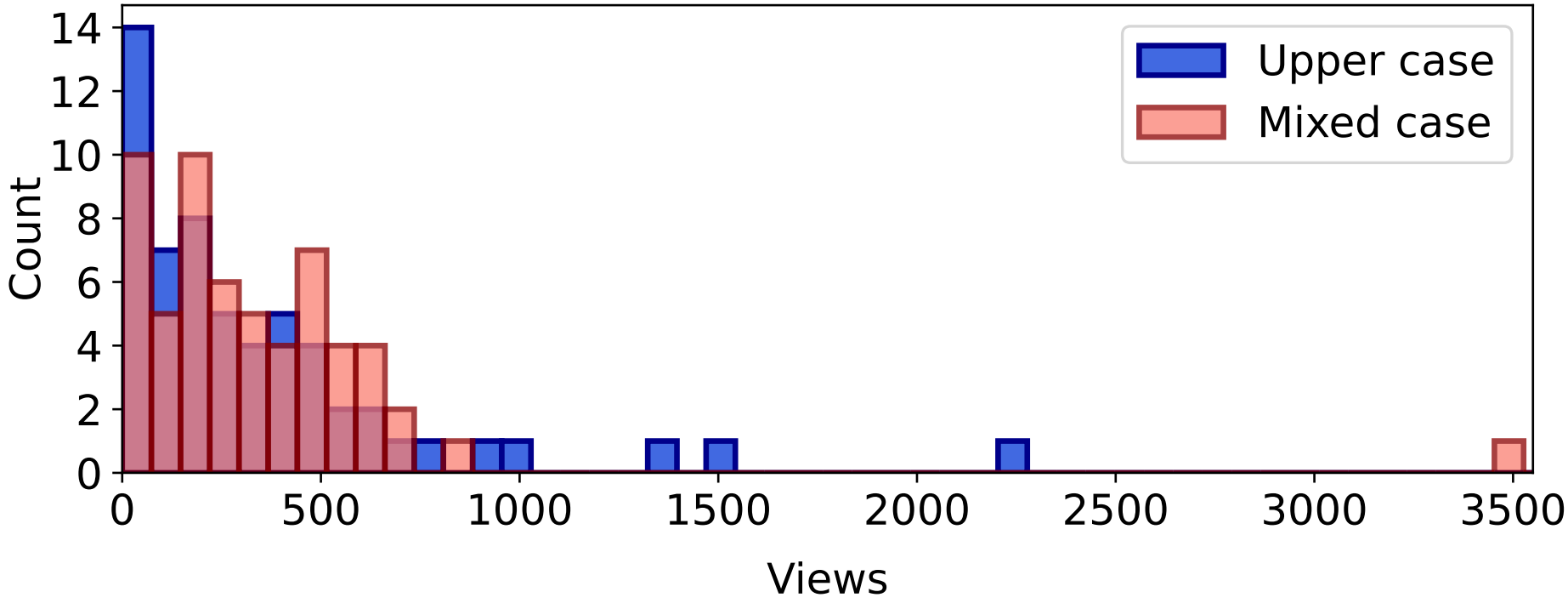
Possible concern? Mean is distorted by outliers

Some videos have a lot of views

Difference of medians is more robust to outliers

Views

	Mean	Mean w/o max	Median	Median w/o max
Upper case	349.5	316.3	229	214
Mixed case	359.8	305.3	281	268.5
Difference	**-10.3**	**11**	**-52**	**-54.5**



So mixed-case titles cause more views?

Not necessarily!

Maybe this is just random noise

Data acquisition

Selection of statistical metric

Hypothesis testing

Hypothesis testing

1. Choose a conjecture: *Capitalization makes a difference*
2. Choose null hypothesis: *Capitalization makes no difference*
3. Choose test statistic: **Absolute difference of medians**
4. Decide significance level α: **0.05**
5. Gather data and compute test statistic: **52**
6. Compute p value
7. Reject the null hypothesis if p value $\leq \alpha := 0.05$

Computing the p value

Parametric testing requires modeling test statistic under null hypothesis

Distribution of absolute difference of medians?

¯_(ツ)_/¯

Solution: Nonparametric testing

Permutation test

Assumption: Under null hypothesis, treatment is **irrelevant** to the views

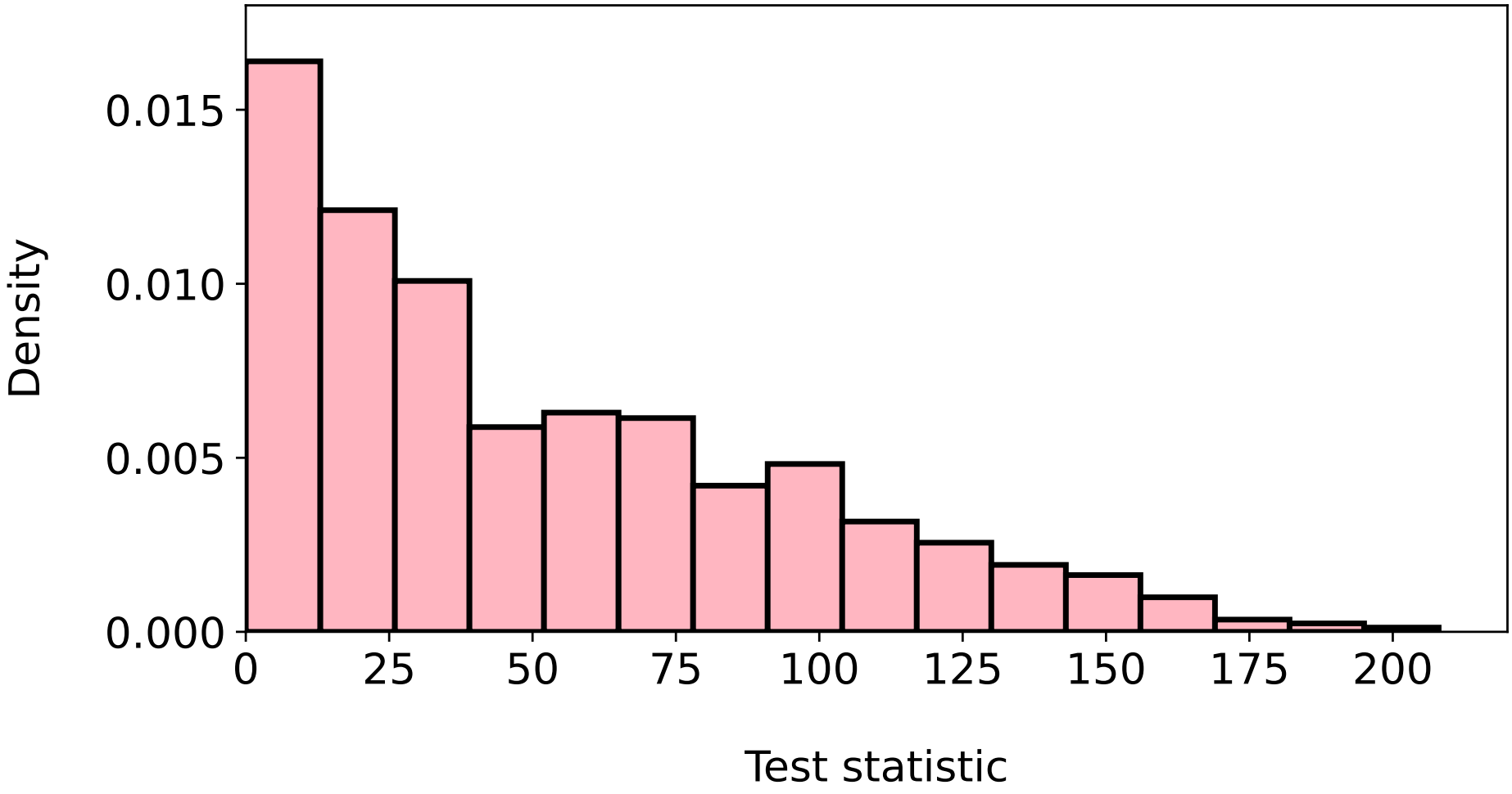
Switching treatment labels should make **no** difference

We permute the treatment labels and recompute test statistic many times

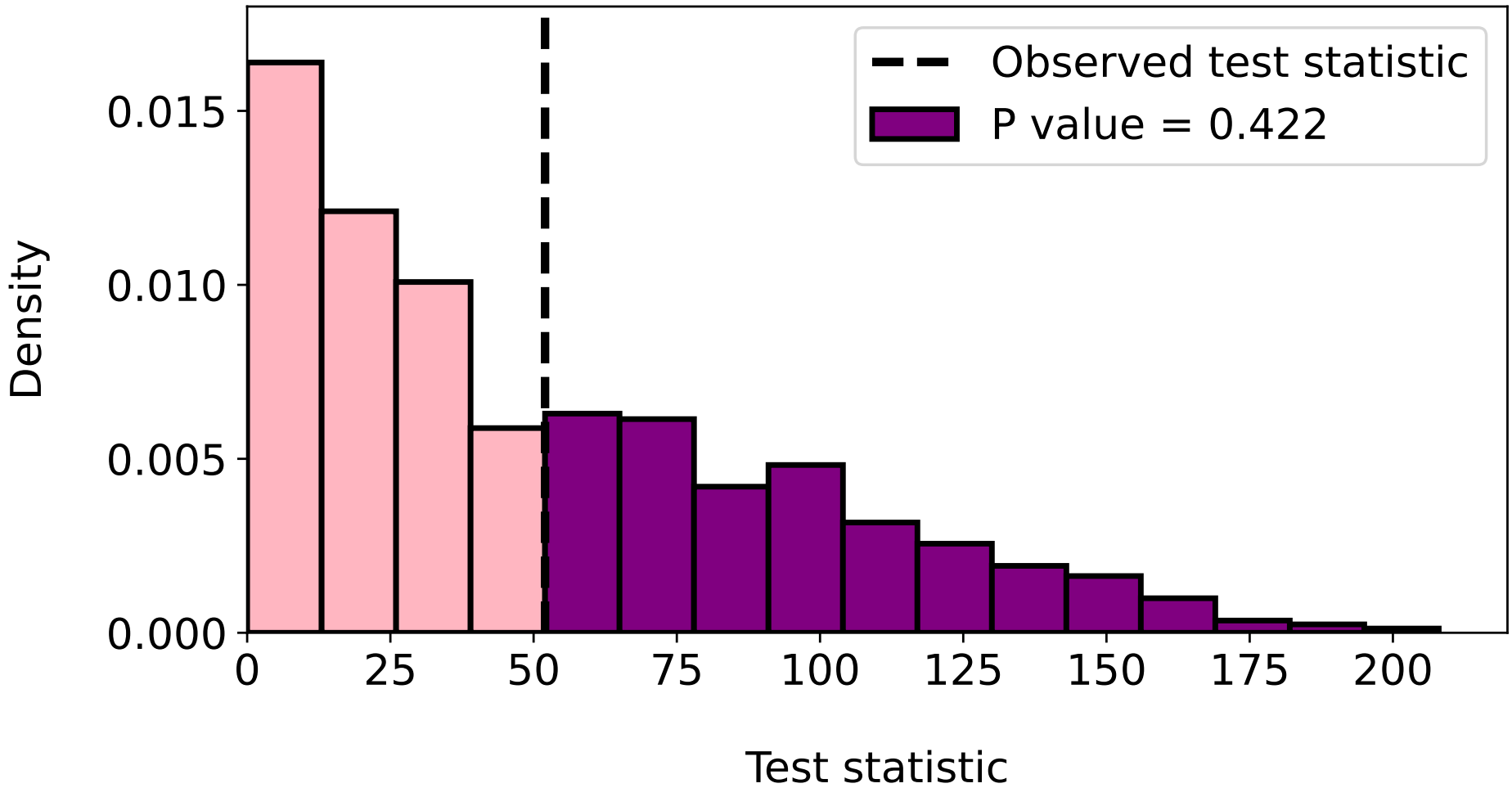
P value is fraction of permutations for which

$$\text{test statistic} \geq \text{observed test statistic}$$

One million permutations



P value



$0.422 \geq 0.05$: Not statistically significant

Conclusion

Data are not inconsistent with null hypothesis

We don't have enough evidence that capitalization has an effect

Recap

Treatment randomization enables **causal inference**

Statistic based on the median ensures **robustness to outliers**

Permutation testing determines **statistical significance**