Applied Natural Language Processing

Info 256

Lecture 10: Nonparametric tests (Sept 27, 2023)

David Bamman, UC Berkeley

September 27, 2023

BIDS’ Center for Cultural Analytics Lecture with Professor David Blei

4:30 - 6 p.m.
Sutardja Dai Hall Auditorium, UC Berkeley

Sponsor(s): Berkeley Institute for Data Science (BIDS), Center for Cultural Analytics

→ Join us in person for a lecture and reception with David Blei, Professor of Statistics and Computer Science at Columbia University.

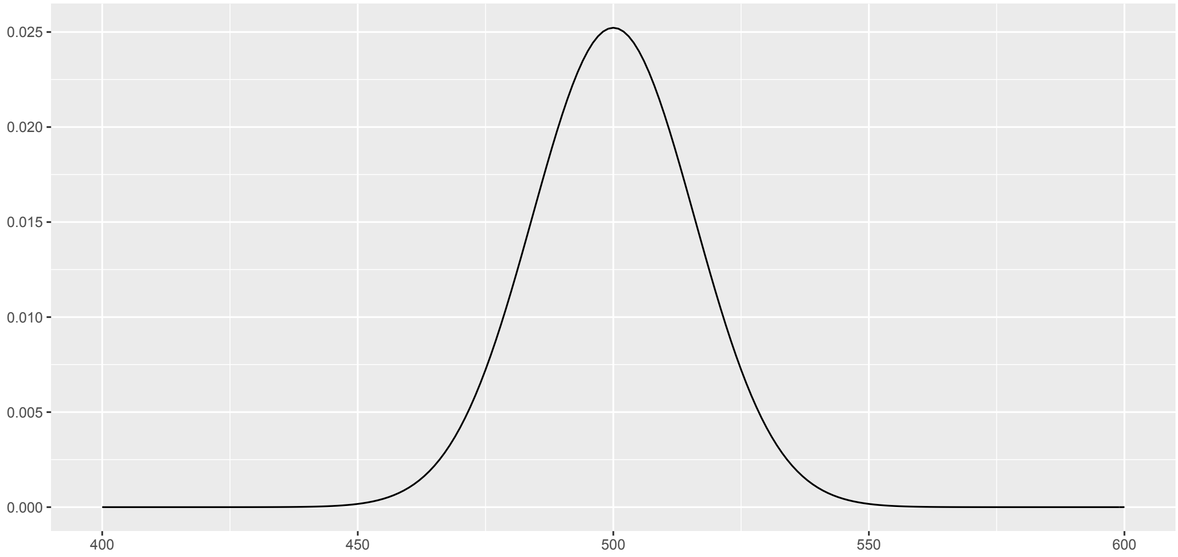
Beyond Roll Call: Inferring Politics from Text

“The ideal point model is a staple of quantitative political science. It is a probabilistic model of roll call data—how a group of lawmakers vote on a collection of bills—that can be used to quantify the lawmakers’ political

Hypothesis tests

- At what point is the sample statistic so unusual that we can reject the null hypothesis as being too unlikely to have generated it?

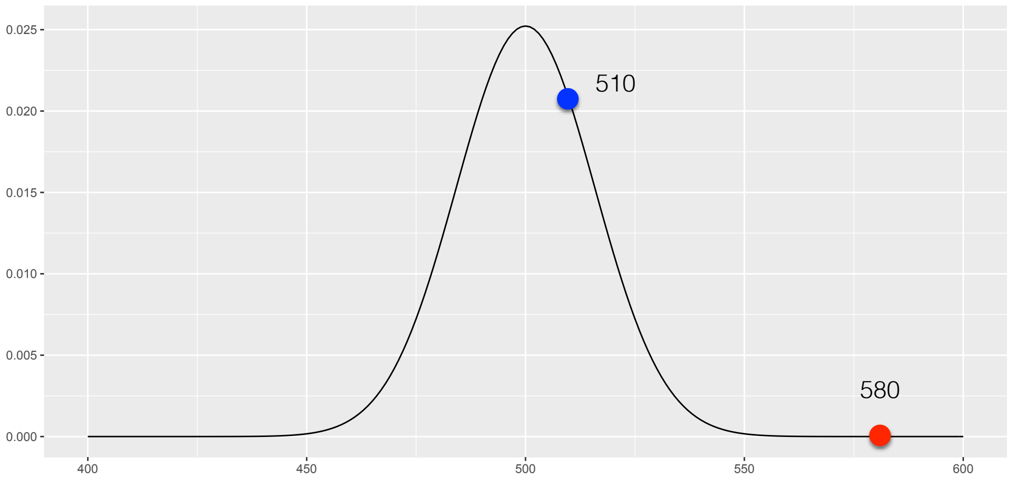
Example



Binomial probability distribution for number of correct predictions in n=1000 with p = 0.5

Example

At what point is a sample statistic unusual enough to reject the null hypothesis?

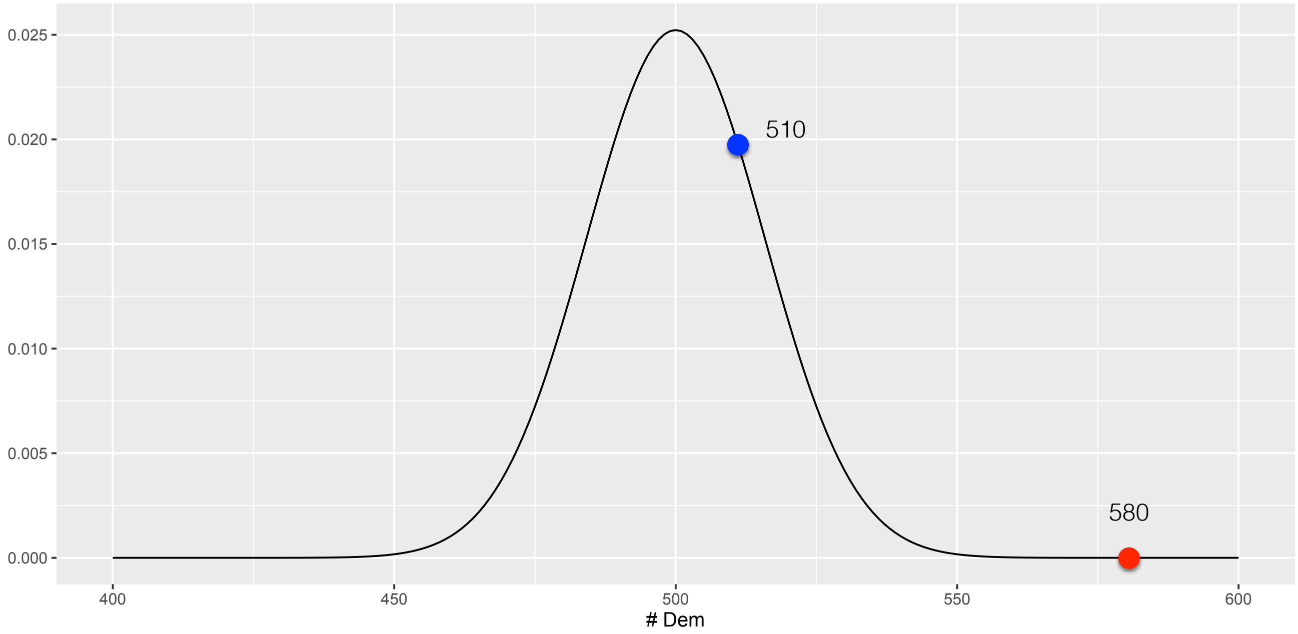


Hypothesis tests

- How do we define what “too unusual” means?
- Parametric tests state that the null hypothesis follows a probability distribution with a fixed set of parameters:
    - Binomial (parameterized by the success rate p and number of trials n)
    - Normal (parametrized by mean μ and standard deviation σ)

Hypothesis tests

- How do we define what "too unusual" means?
- Parametric tests state that the null hypothesis follows a probability distribution with a fixed set of parameters
- In these tests, we can calculate the probability of the statistic by just looking it up
    - e.g., $P(x=580 \mid p=0.50, n=1000)$ in Binomial distribution.


0.025
0.020
0.015
0.010
0.005
0.000
510
580
400
450
500
550
600
Dem

Parametric tests

- Parametric tests often rely on a normal approximation for large sample sizes, using the central limit theorem (CLT)
- CLT: the average of independent random variables tends toward a normal distribution, even if the original variables themselves are not normally distributed.

Accuracy

$$\frac{1}{N}\sum_{i=1}^{N} I[\hat{y}_i = y_i]$$

$$I[x] \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Accuracy: proportion of all data points that are correctly predicted.

Predicted (ŷ)

True (y)	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	30	40	70

Metrics

Metric	Simple averaging?
Accuracy	✔
Precision	
Recall	
F1	

Precision

Precision(Dem) =

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{Dem})}{\sum_{i=1}^{N} I(\hat{y}_i = \text{Dem})}$$

Precision: proportion of predicted class that are actually that class.

Predicted (ŷ)

True (y)	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	30	40	70

Metrics

Metric	Simple averaging?
Accuracy	✔
Precision	✔
Recall	
F1	

Recall

Recall(Dem) =

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{Dem})}{\sum_{i=1}^{N} I(y_i = \text{Dem})}$$

Recall: proportion of true class that are predicted to be that class.

True (y) \ Predicted (ŷ)	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	30	40	70

Metrics

Metric	Simple averaging?
Accuracy	✔
Precision	✔
Recall	✔
F1	

F score

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

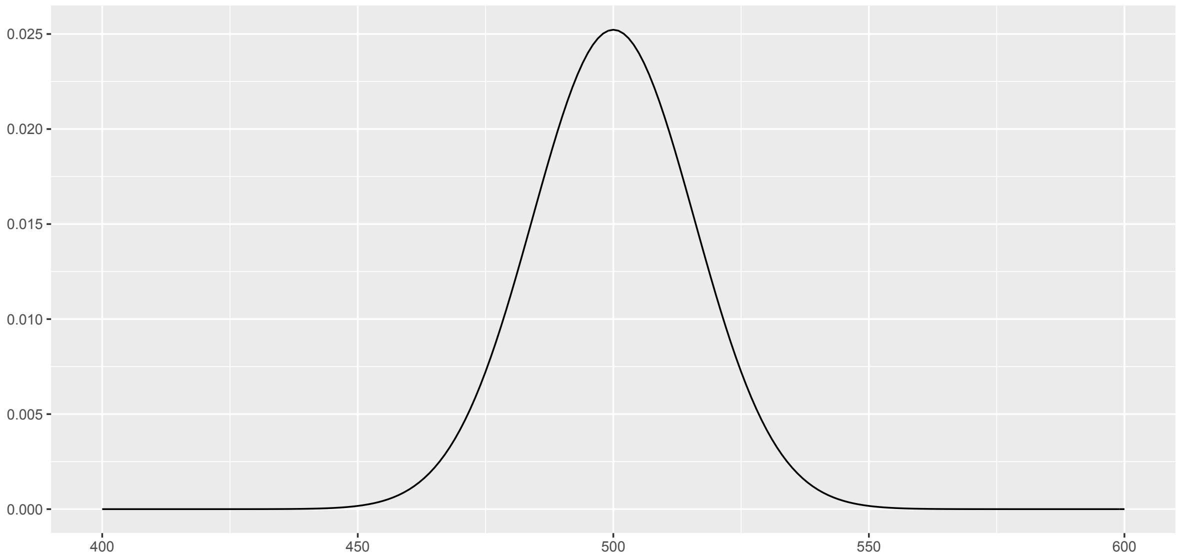
Metrics

Metric	Simple averaging?
Accuracy	✔
Precision	✔
Recall	✔
F1	✖

Nonparametric tests

- The big question: if we can't make a parametric assumption (e.g., that accuracy follows a normal distribution), how can we say how unlikely a given test statistic is?

- How do we construct a null distribution?


0.025
0.020
0.015
0.010
0.005
0.000
400
450
500
550
600

Nonparametric tests

- Many hypothesis tests rely on parametric assumptions (e.g., normality)
- Alternatives that don't rely on those assumptions:
    - permutation test
    - the bootstrap

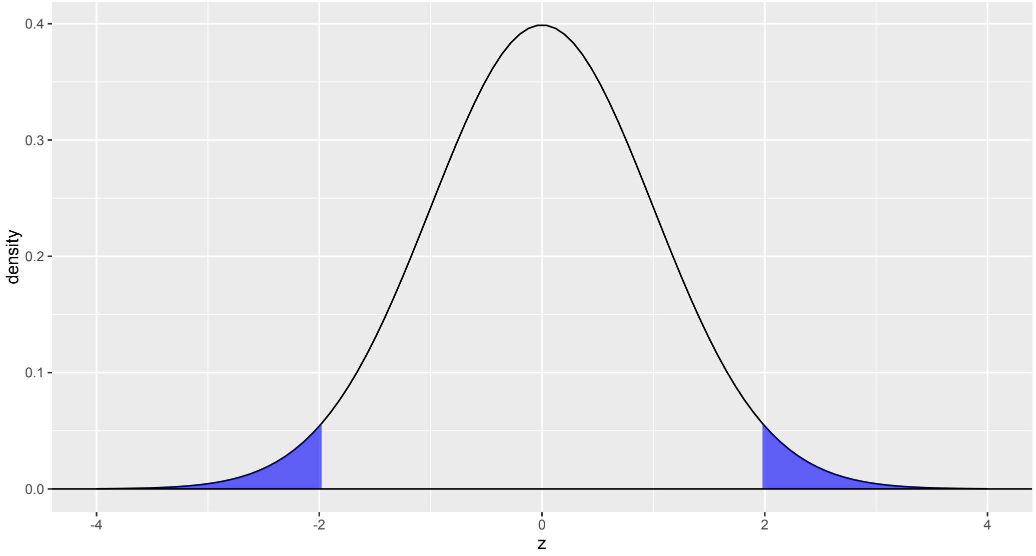
Back to logistic regression

β	change in odds	feature name
2.17	8.76	Eddie Murphy
1.98	7.24	Tom Cruise
1.70	5.47	Tyler Perry
1.70	5.47	Michael Douglas
1.66	5.26	Robert Redford
…	…	…
-0.94	0.39	Kevin Conway
-1.00	0.37	Fisher Stevens
-1.05	0.35	B-movie
-1.14	0.32	Black-and-white
-1.23	0.29	Indie

Significance of coefficients

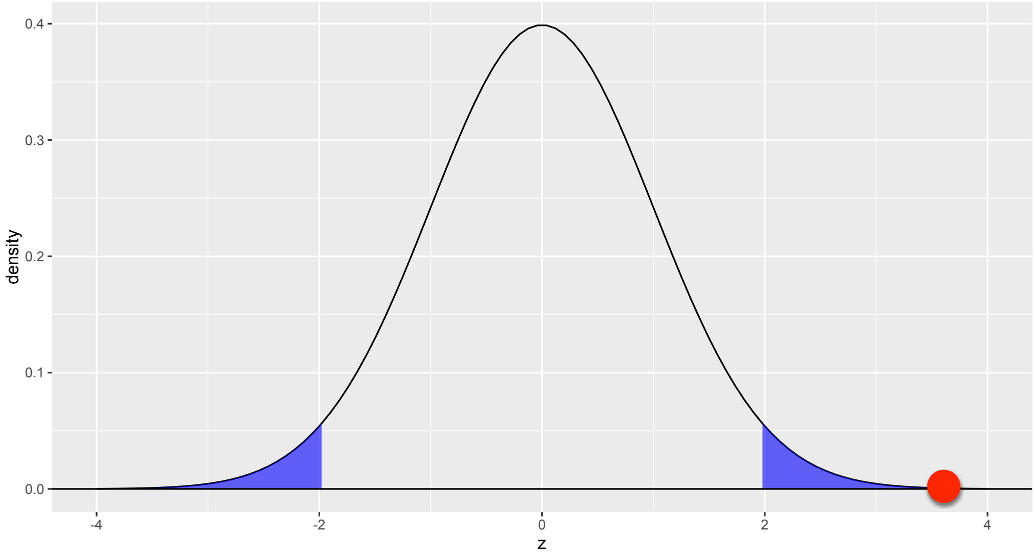
- A β_i value of 0 means that feature x_i has no effect on the prediction of y
- How great does a β_i value have to be for us to say that its effect probably doesn't arise by chance?
- People often use parametric tests (coefficients are drawn from a normal distribution) to assess this for logistic regression, but we can use it to illustrate another more robust test.

Hypothesis tests



Hypothesis tests measure how (un)likely an observed statistic is under the null hypothesis

Hypothesis tests



Permutation test

- Non-parametric way of creating a null distribution (parametric = normal etc.) for testing the difference in two populations A and B
- For example, the respect shown by OPD to drivers who are Black (=A) vs. White (=B)
- We shuffle the labels of the data under the null assumption that the labels don't matter (the null is that A = B)

Permutation test

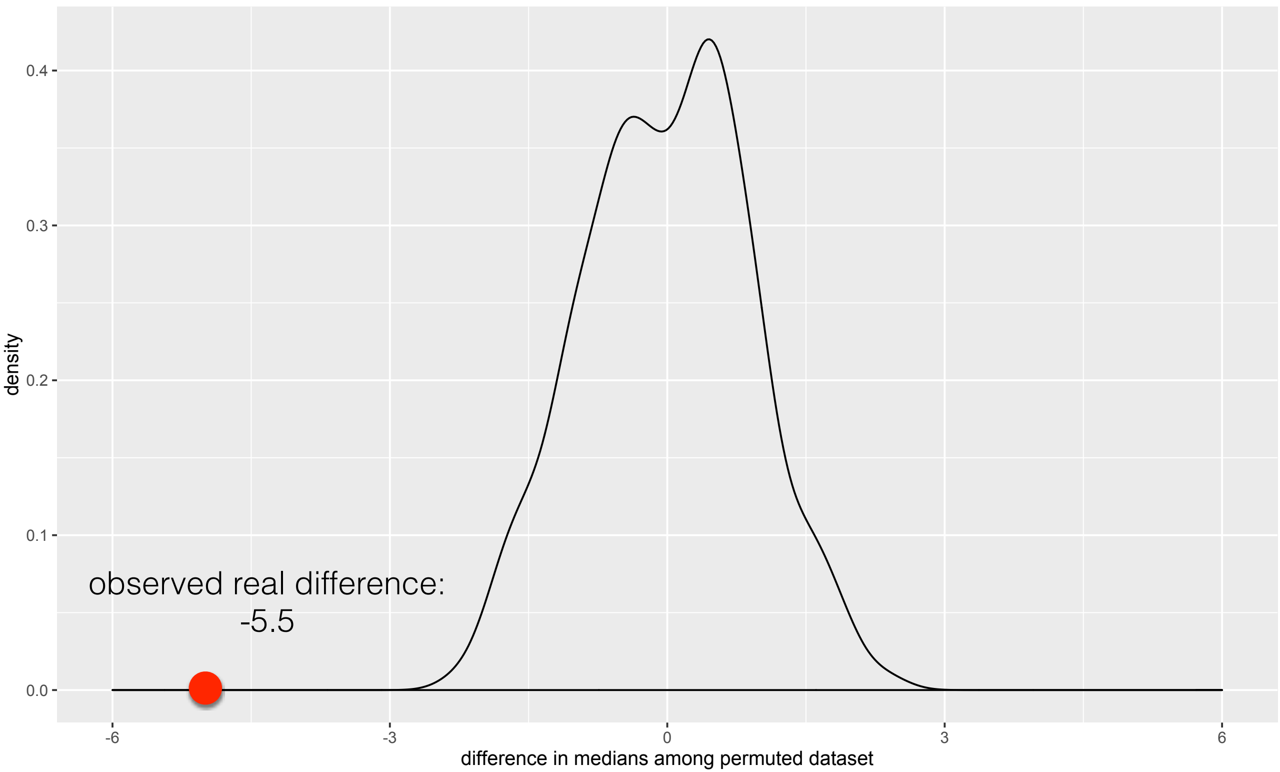
- Core idea: if the null hypothesis were true and there's no difference between groups, then it doesn't matter which label each data point has.

	respect	true labels	perm 1	perm 2	perm 3	perm 4	perm 5
x1	62.8	black	white	white	black	white	white
x2	66.2	black	white	white	white	black	black
x3	65.1	black	white	white	black	white	white
x4	68.0	black	white	black	white	black	black
x5	61.0	black	black	white	white	white	white
x6	73.1	white	black	black	white	black	black
x7	67.0	white	white	black	white	black	white
x8	71.2	white	black	black	black	white	white
x9	68.4	white	black	white	black	white	black
x10	70.9	white	black	black	black	black	black

observed true difference in medians: -5.5

	respect	true label	perm 1	perm 2	perm 3	perm 4	perm 5
x1	62.8	black	white	white	black	white	white
x2	66.2	black	white	white	white	black	black
…	…	…	…	…	…	…	…
x9	68.4	white	black	white	black	white	black
x10	70.9	white	black	black	black	black	black
difference in medians:		-5.5	-0.8	0.3	1.4	1.2	-2.0

How many times is the difference in medians between the permuted groups greater than the observed difference?



A=100 samples from Norm(70,4)   B=100 samples from Norm(65, 3.5)

p values

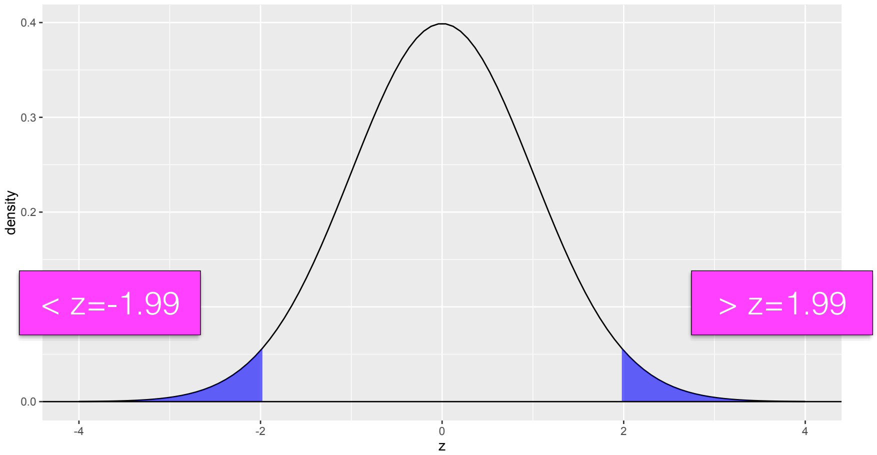
A p value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.

- Two-tailed test $\quad \text{p-value}(z) = 2 \times P(Z \leq -|z|)$
- Lower-tailed test $\quad \text{p-value}(z) = P(Z \leq z)$
- Upper-tailed test $\quad \text{p-value}(z) = 1 - P(Z \leq z)$

P-value

- If our test statistic is 1.99, then the two-tailed p-value is the sum of the shaded probability mass in the extremes.

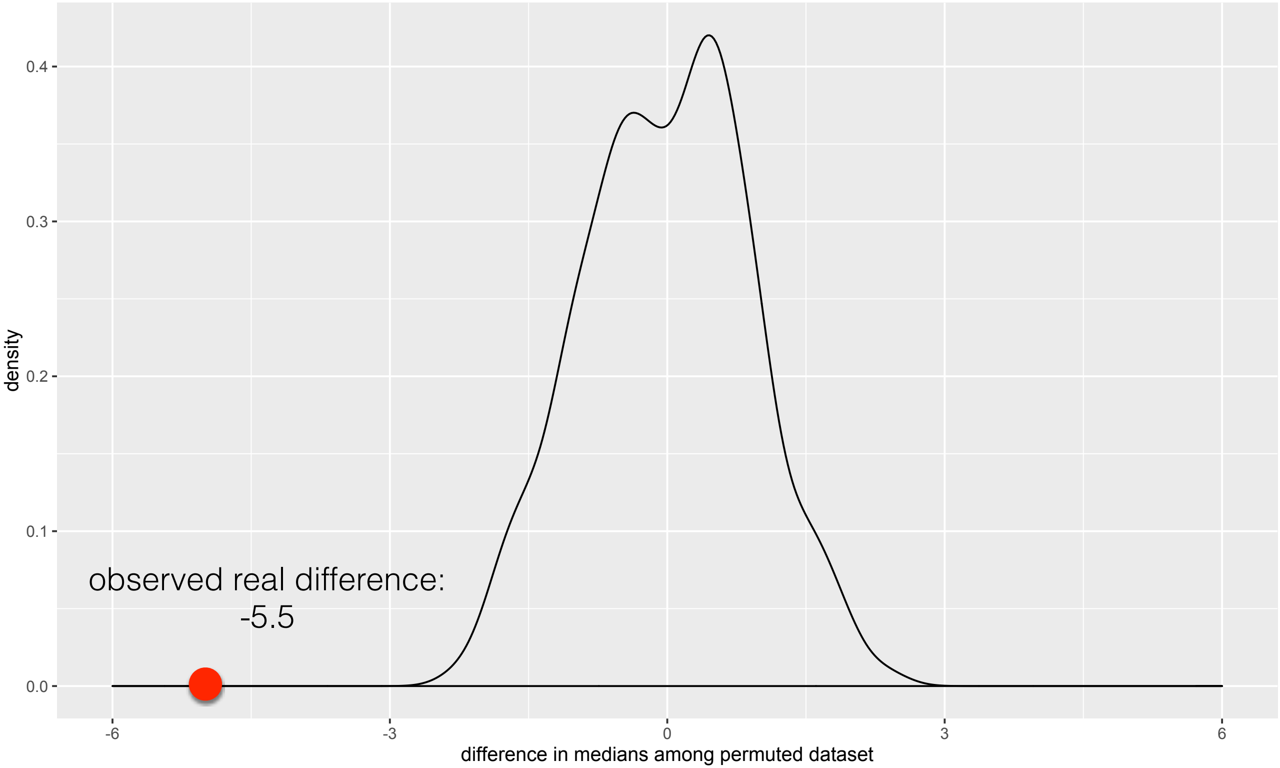
- In parametric tests, we can calculate this using the CDF $P(X < x)$ of the null distribution.



Permutation test

The p-value is the number of times the permuted test statistic t_p is more extreme than the observed test statistic t:

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} I[abs(t) < abs(t_p)]$$



A=100 samples from Norm(70,4)

B=100 samples from Norm(65, 3.5)

Permutation test

- The permutation test is a robust test that can be used for many different kinds of test statistics, including coefficients in logistic regression.

- How?

    - A = members of class 1
    - B = members of class 0
    - β are calculated as the (e.g.) the values that maximize the conditional probability of the class labels we observe; its value is determined by the data points that belong to A or B

Permutation test

- To test whether the coefficients have a statistically significant effect (i.e., they're not 0), we can conduct a permutation test where, for B trials, we:

    1. shuffle the class labels in the training data

    2. train logistic regression on the new permuted dataset

    3. tally whether the absolute value of β learned on permuted data is greater than the absolute value of β learned on the true data

Permutation test

The p-value is the number of times the permuted β_p is more extreme than the observed β_t:

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} I[abs(\beta_t) < abs(\beta_p)]$$

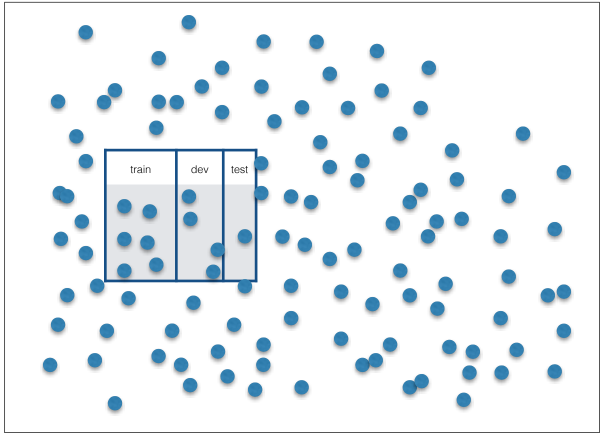
Bootstrap

- The permutation test assesses significance conditioned on the test data you have (we rearrange the labels to form the null distribution, but the data itself doesn't change).
- To also model the variability in the data we have, we can use the statistical bootstrap (Efron 1979).

Bootstrap

- Core idea: the data we happen to have is a sample from all data that could exist; let's sample from our sample to estimate the variability.

- Our estimate of the point value of the metric itself won't change, but we can infer something about the variability of the population from the variable in the resamples.



Bootstrap

- Start with test data x of size n
- Draw b bootstrap samples x(i) of size n by sampling with replacement from x
- For each x(i)
  - Let m(i) = the metric of interest calculated from x(i)

				accuracy
I love this movie	I hate this movie	I don't love this movie	Not the worst ever!	0.50

				m(i)
I love this movie	I don't love this movie	I don't love this movie	Not the worst ever!	0.25
I love this movie	I love this movie	I hate this movie	Not the worst ever!	0.75
I hate this movie	I don't love this movie	I don't love this movie	I love this movie	0.50
I love this movie	I hate this movie	I don't love this movie	I hate this movie	0.75
I don't love this movie	I don't love this movie	I don't love this movie	Not the worst ever!	0.00

Bootstrap percentile interval

- At the end of the process, you end up with a vector of values m = [m(1), …, m(b)] (for b bootstrap samples) — e.g. [0.25, 0.75, 0.50, 0.75, 0] for the example before.

- We can define a 95% confidence interval as the middle 95% of m

- e.g., $\alpha = 0.05$ (95% confidence intervals) = [2.5, 97.5] percentile

- Accurate for larger sample sizes

Activity

`7.tests/PermutationTest.ipynb`

- Explore using the permutation test to analyze the significance of logistic regression coefficients on your data

`7.tests/Bootstrap.ipynb`

- Using the bootstrap to calculate confidence intervals for any metric — try on your own model.