Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 10: Validity

Feb 16, 2017

Hypotheses

hypothesis
The average income in two sub-populations is different
Web design A leads to higher CTR than web design B
Self-reported location on Twitter is predictive of political preference
Male and female literary characters become more similar over time

Hypotheses

The first step is formalizing a question into a testable hypothesis.

hypothesis “area”
Voters in big cities prefer Hillary Clinton
Email marketing language A is better than language B
Slapstick comedies do not win Oscars
Joyce’s *Ulysses* changed the form of the novel after 1922

Null hypothesis

- A claim, assumed to be true, that we'd like to test (because we think it's wrong)

hypothesis	H_0
The average income in two sub-populations is different	The incomes are the same
Web design A leads to higher CTR than web design B	The CTR are the same
Self-reported location on Twitter is predictive of political preference	Location has no relationship with political preference
Male and female literary characters become more similar over time	There is no difference in M/F characters over time

Hypothesis testing

- If the null hypothesis were true, how likely is it that you'd see the data you see?

Example

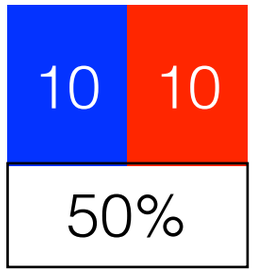
- Hypothesis: Berkeley residents tend to be politically liberal
- H_0: Among all *N* registered {Democrat, Republican} primary voters, there are an equal number of Democrats and Republicans in Berkeley.

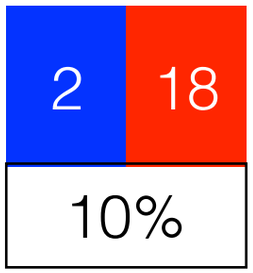
$$\frac{\#dem}{N} = \frac{\#rep}{N} = 0.5$$

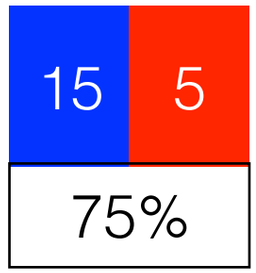
Example

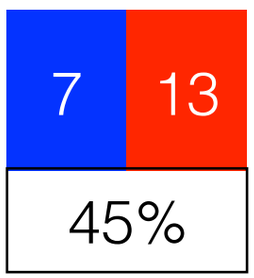
- If we had access to the party registrations (and knew the population), we would have our answer.

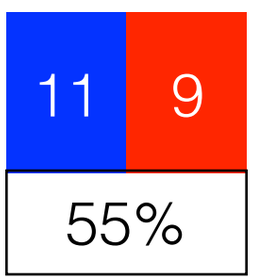
Example

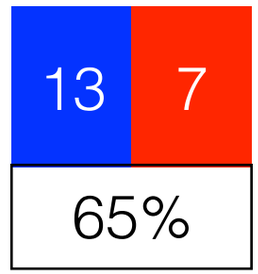








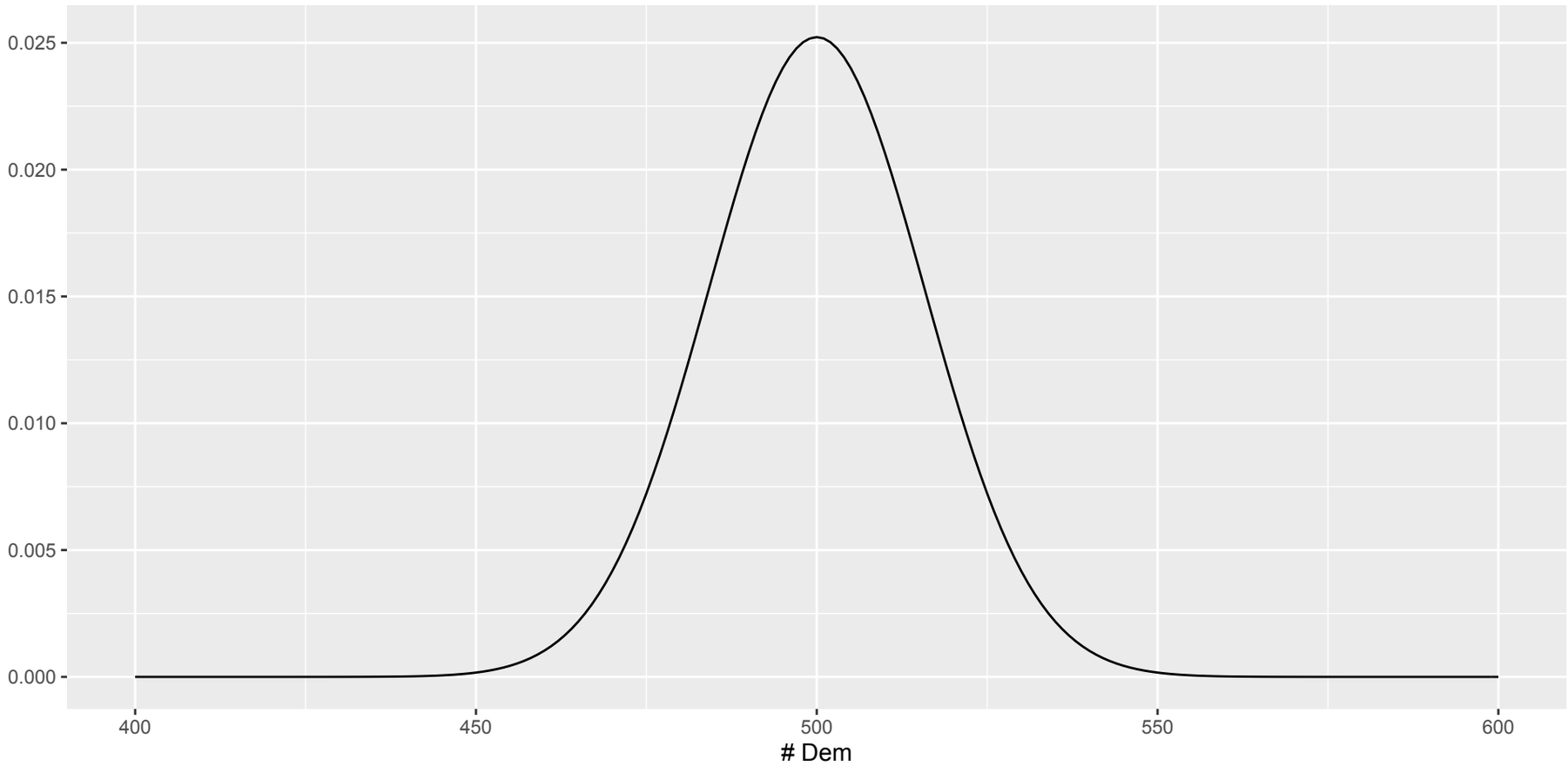




Hypothesis testing

- Hypothesis testing measures our confidence in what we can say about a null from a sample.

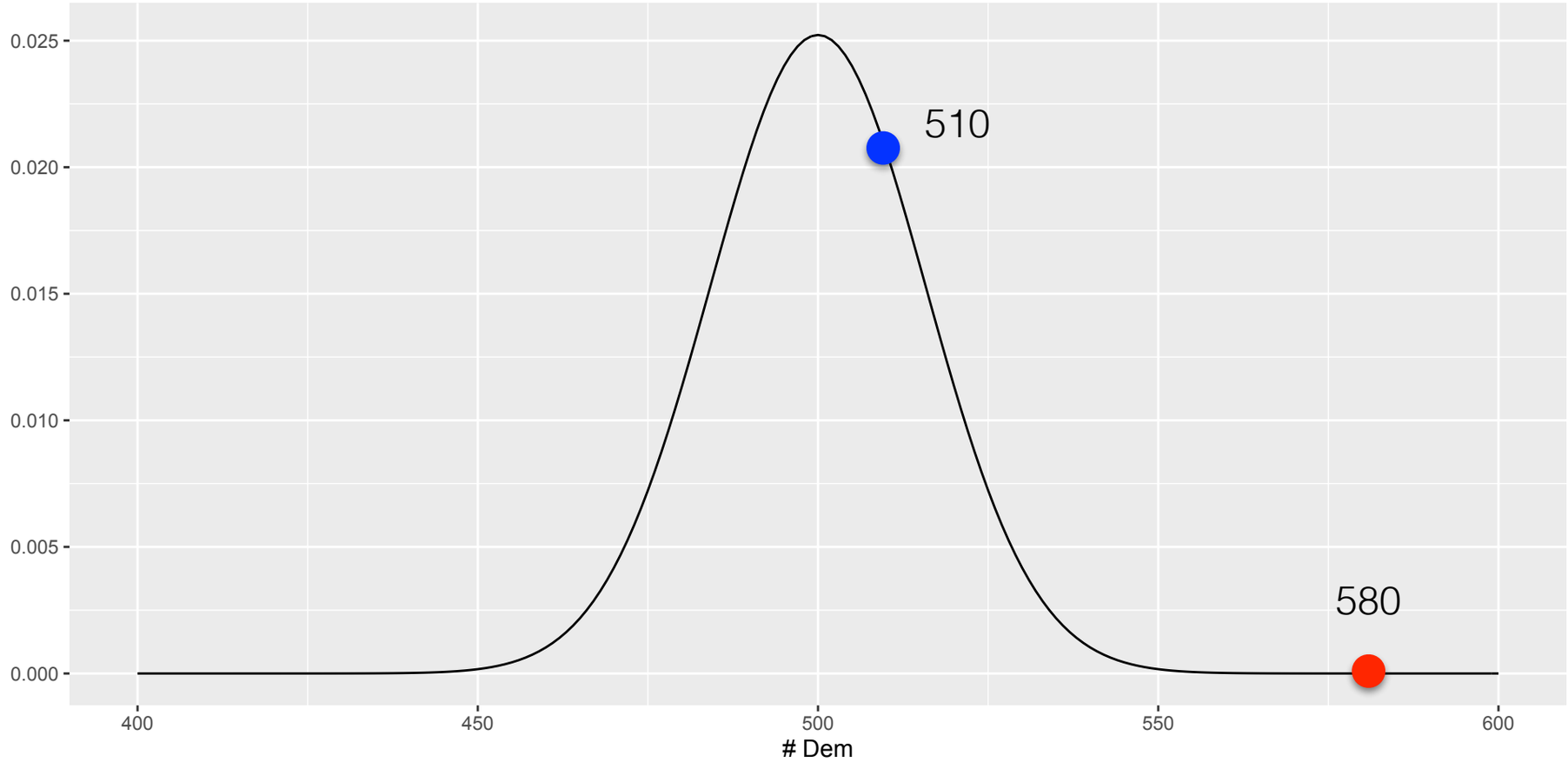
Example



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

Example

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



Example

- The form we assume for the null hypothesis lets us quantify that level of surprise.

- We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size n; for large n, we can often make a normal approximation.

Z score

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}}$$

For Normal distributions, transform into standard normal (mean = 0, standard deviation =1)

$$Z = \frac{Y - np}{\sqrt{(np(1-p))}}$$

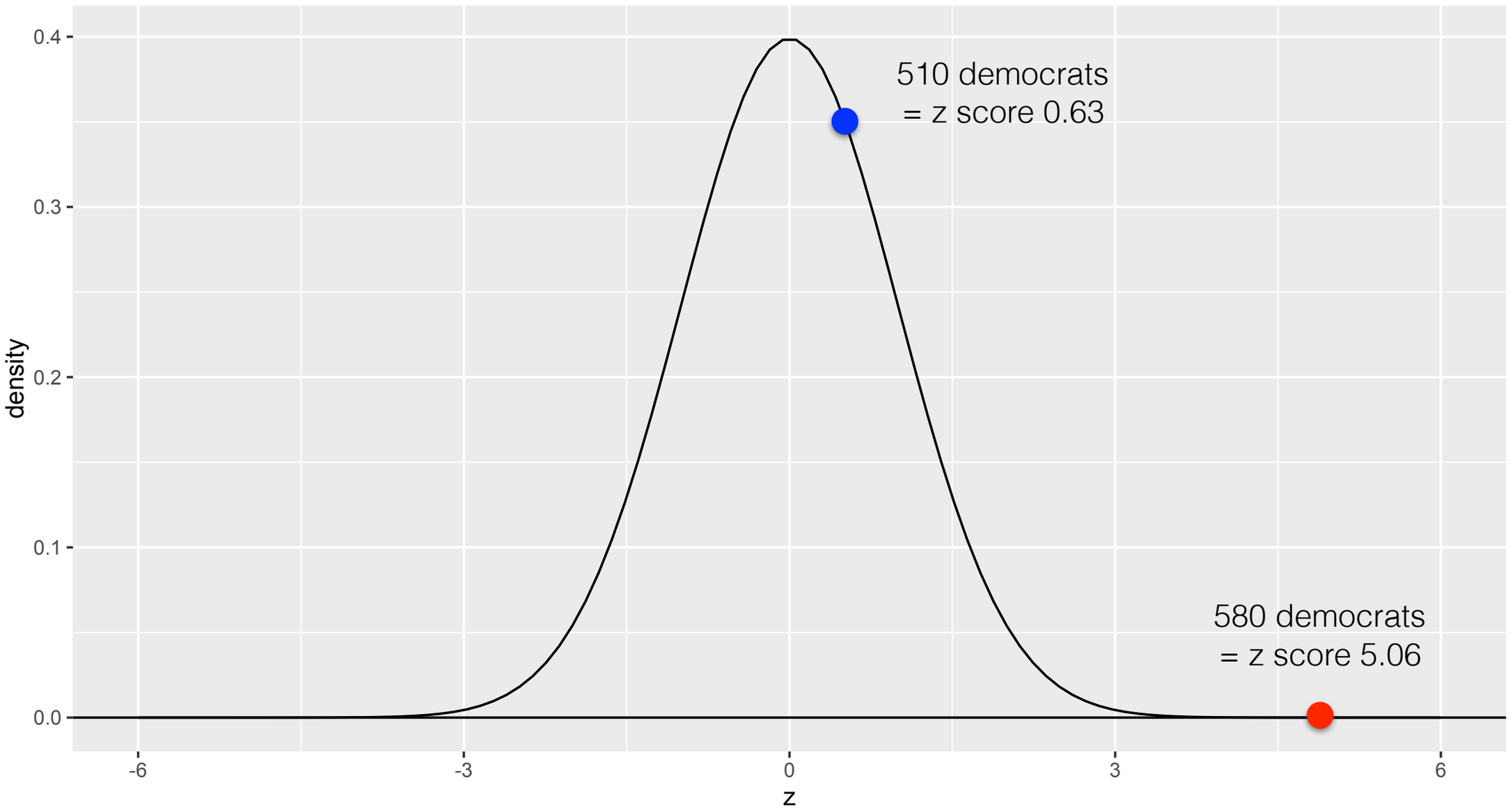
For Binomial distributions, normal approximation (for large n)

Y=580 (democrats in sample)

n=1000 (total sample size)

p = 0.5 (proportion we are testing)

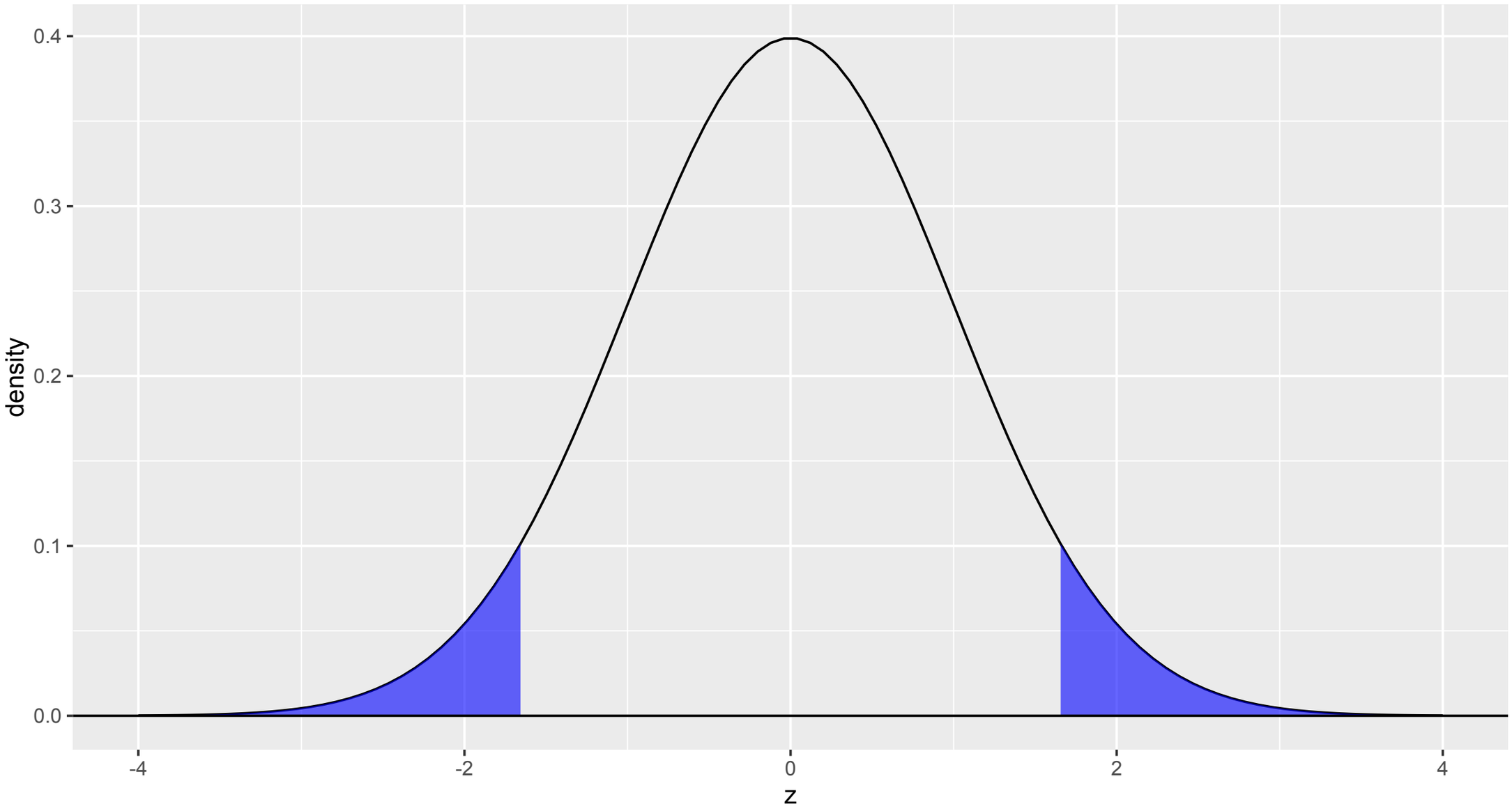
Z score


0.4
0.3
0.2
0.1
0.0
density
510 democrats
= z score 0.63
580 democrats
= z score 5.06
-6
-3
0
3
6
z

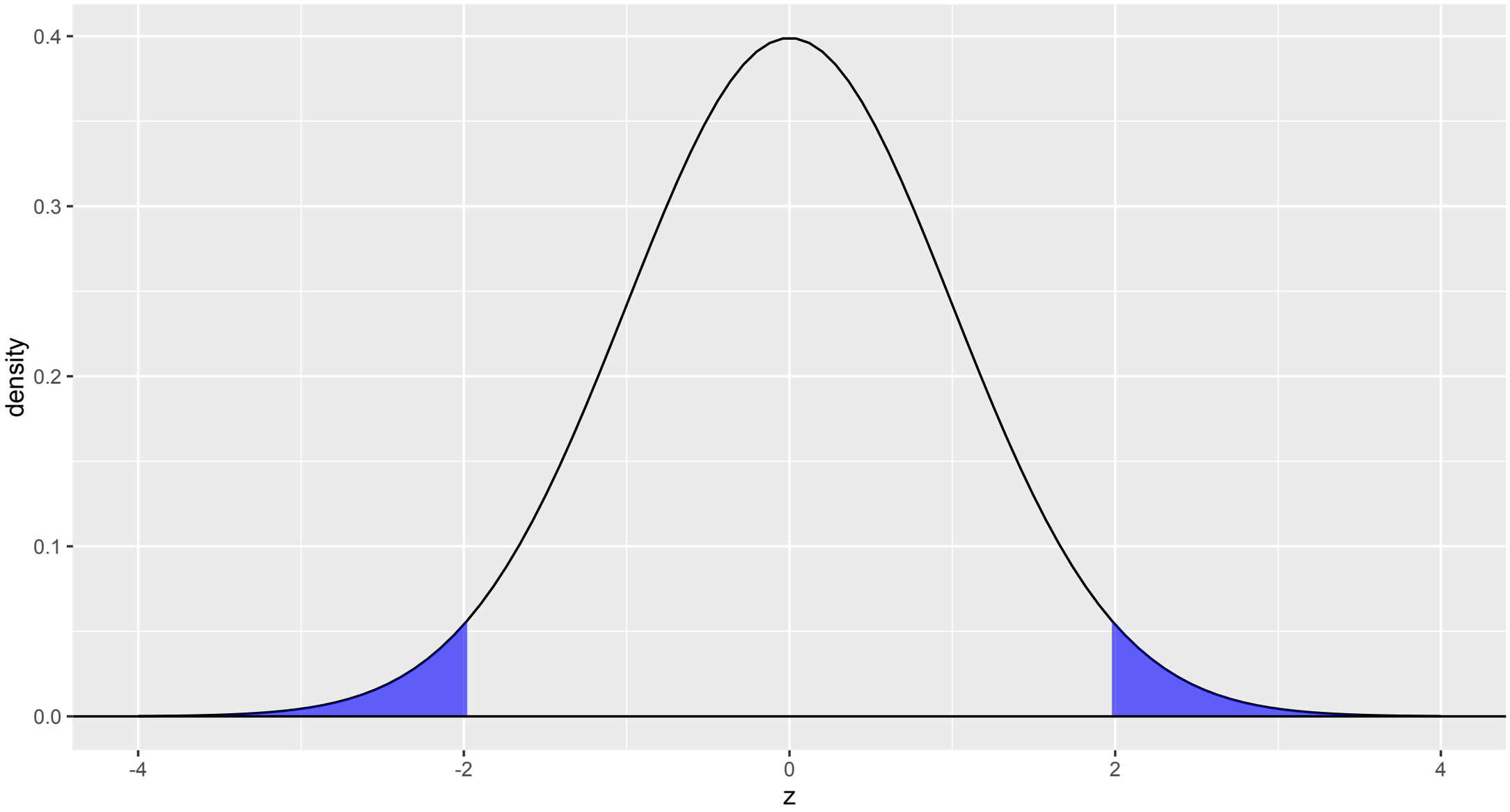
Tests

- We will define “unusual” to equal the most extreme areas in the tails

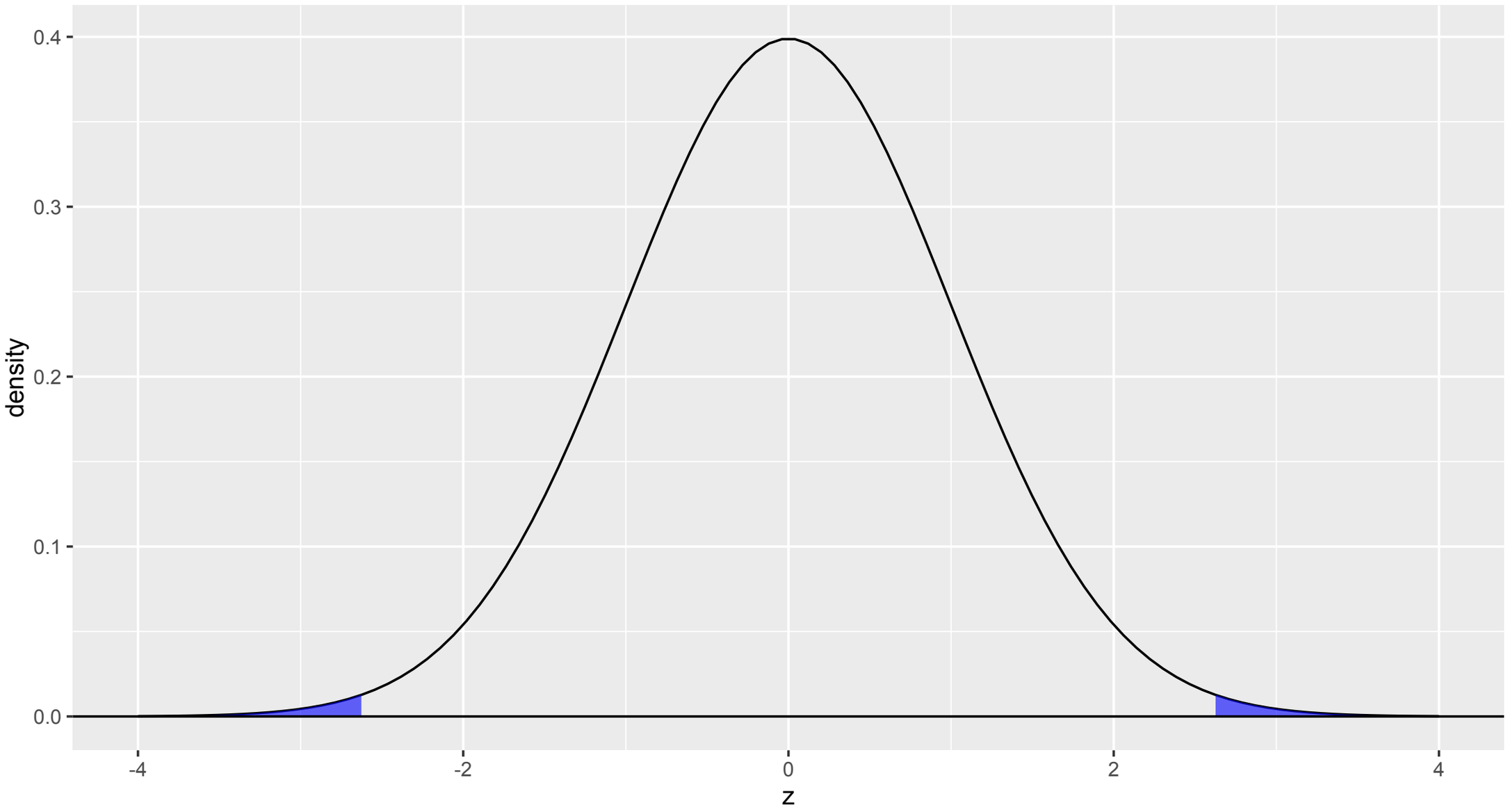
least likely 10%



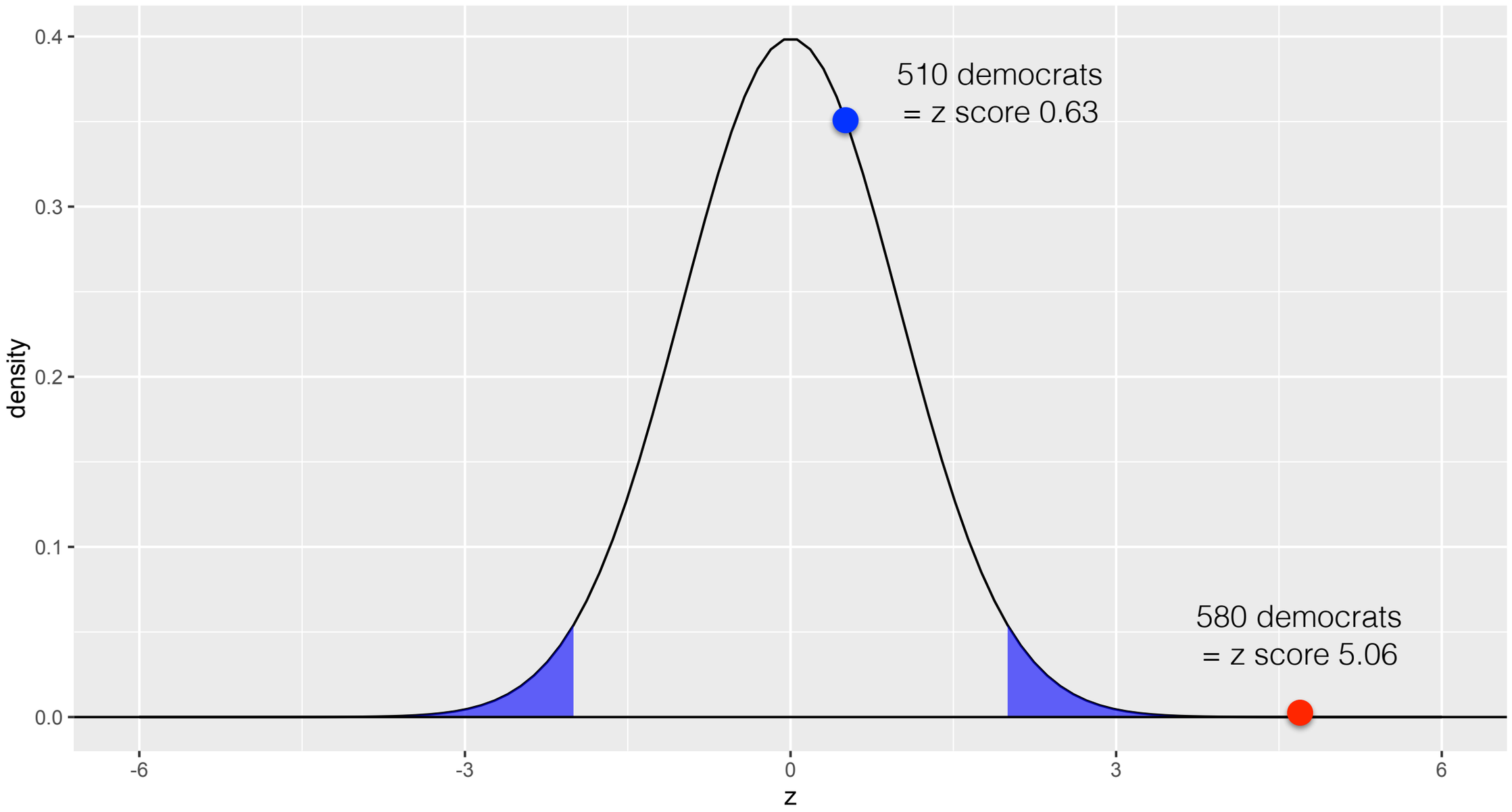
least likely 5%



least likely 1%



Tests


0.4
0.3
0.2
0.1
0.0
density
510 democrats
= z score 0.63
580 democrats
= z score 5.06
-6
-3
0
3
6
z

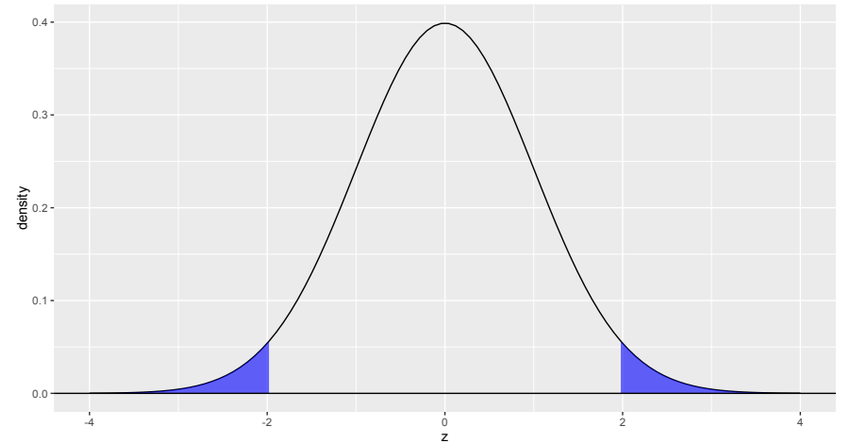
Tests

- Decide on the level of significance α. {0.05, 0.01}

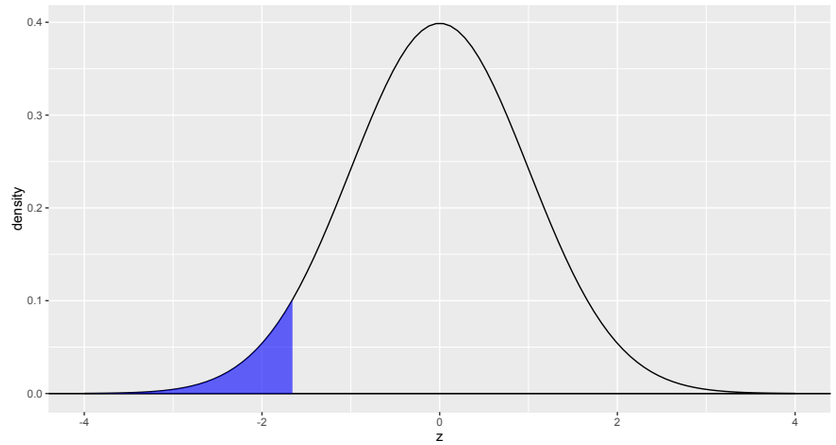
- Testing is evaluating whether the sample statistic falls in the rejection region defined by α

Tails

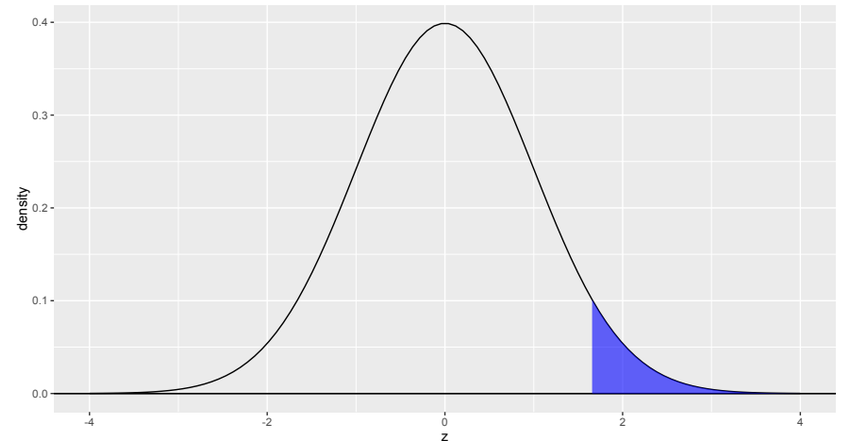
- Two-tailed tests measured whether the observed statistic is different (in either direction)
- One-tailed tests measure difference in a specific direction
- All differ in where the rejection region is located; $\alpha = 0.05$ for all.



two-tailed test



lower-tailed test



upper-tailed test

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

Errors

Truth \ Test results	keep null	reject null
keep null		Type I error α
reject null	Type II error β	Power

Errors

- Type I error: we reject the null hypothesis but we shouldn't have.

- Type II error: we don't reject the null, but we should have.

1	Berkeley residents tend to be politically liberal
2	San Francisco residents tend to be politically liberal
3	Albany residents tend to be politically liberal
4	El Cerrito residents tend to be politically liberal
5	San Jose residents tend to be politically liberal
6	Oakland residents tend to be politically liberal
7	Walnut Creek residents tend to be politically liberal
8	Sacramento residents tend to be politically liberal
9	Napa residents tend to be politically liberal
…	…
1,000	Atlanta residents tend to be politically liberal

Errors

- For any significance level α and n hypothesis tests, we can expect αxn type I errors.

- α=0.01, n=1000 = 10 “significant” results simply by chance

- When would this occur in practice?

Multiple hypothesis corrections

- Bonferroni correction: for family-wise significance level α_0 with n hypothesis tests:

$$\alpha \leftarrow \frac{\alpha_0}{n}$$

- [Very strict; controls the probability of at least one type I error.]

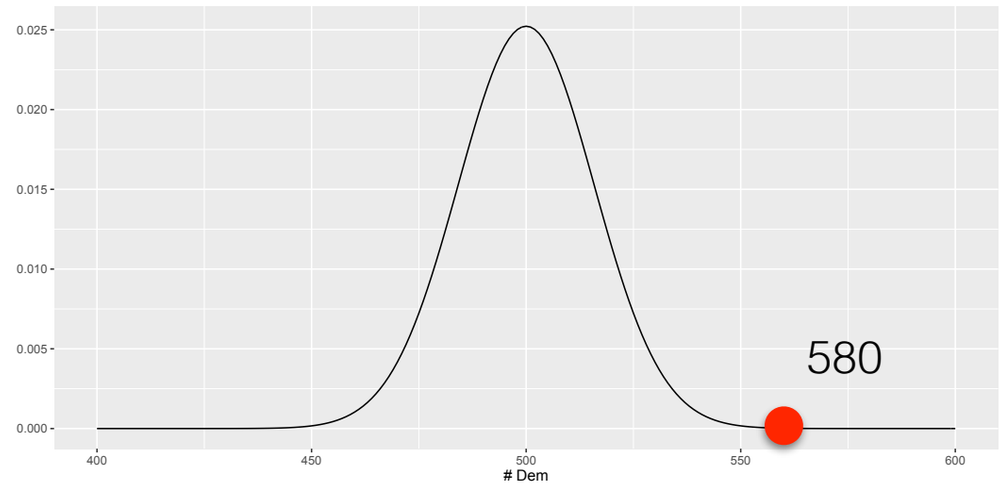
- False discovery rate

Effect size

- Hypothesis tests measure a binary decision (reject or do not reject a null). Many ways to attain significance; e.g.:
    - large true difference in effects
    - large n

Effect size

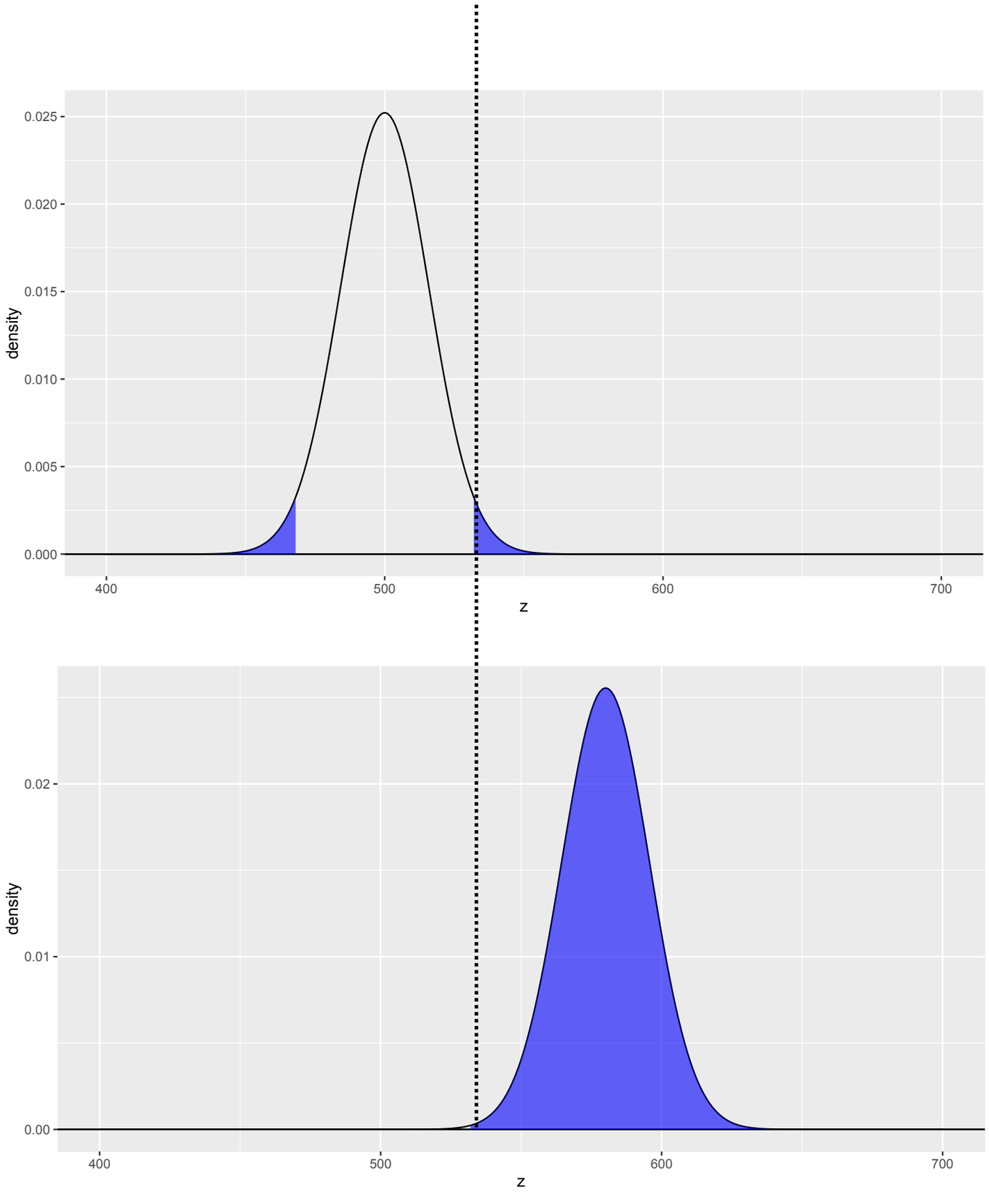
- Difference between the observed statistic and null hypothesis



null hypothesis	observed	effect size (%)	effect size (n)
0.50	0.58	8.0	80

Power

- The probability of a single sample to reject the null hypothesis when it should be rejected



For a fixed effect size, how much of alternative distribution is in the H_0 rejection region?

99.90% of samples from here will be in rejection region (if H_0 is false)

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

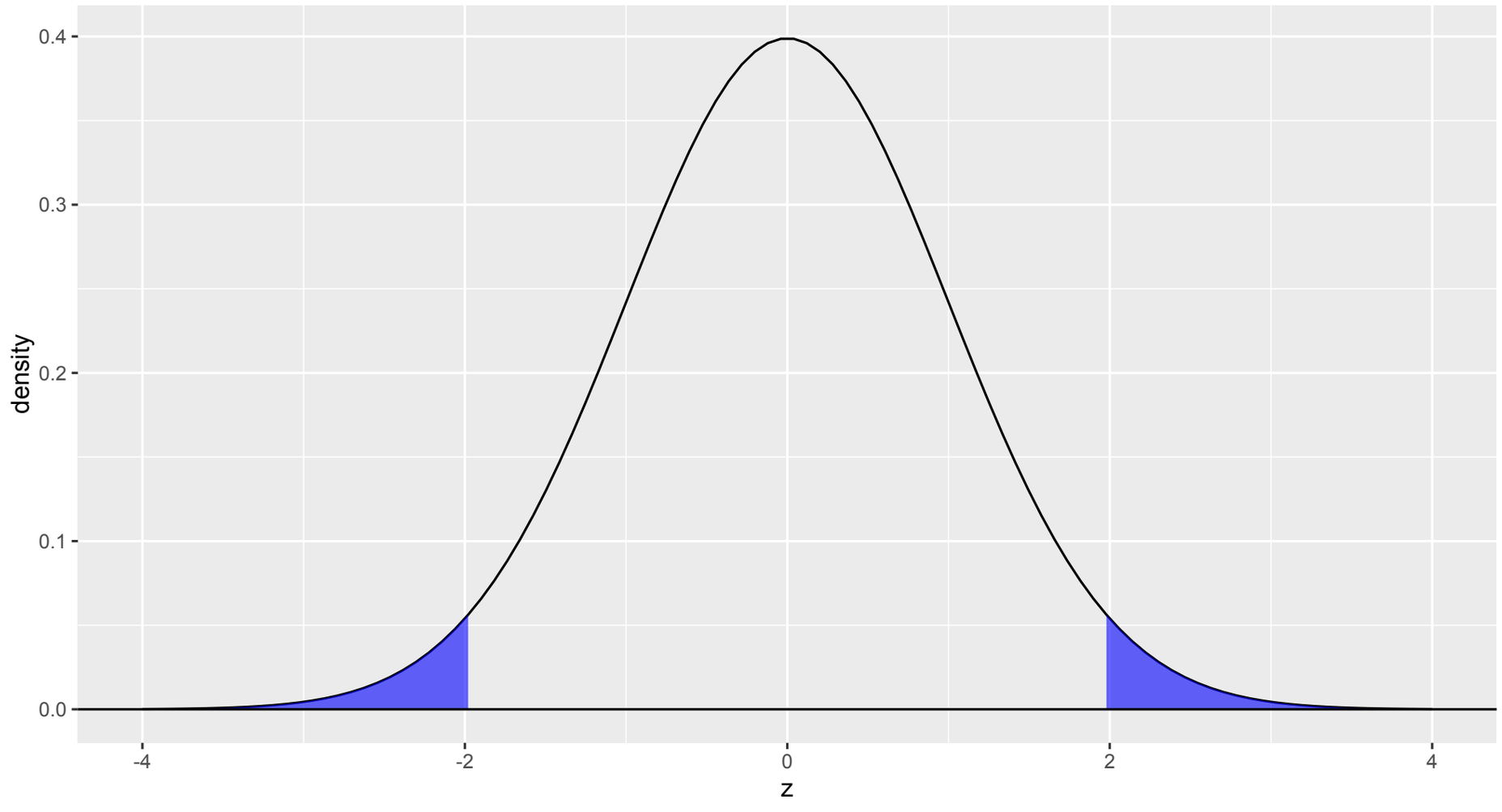
Significance of coefficients

- 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?

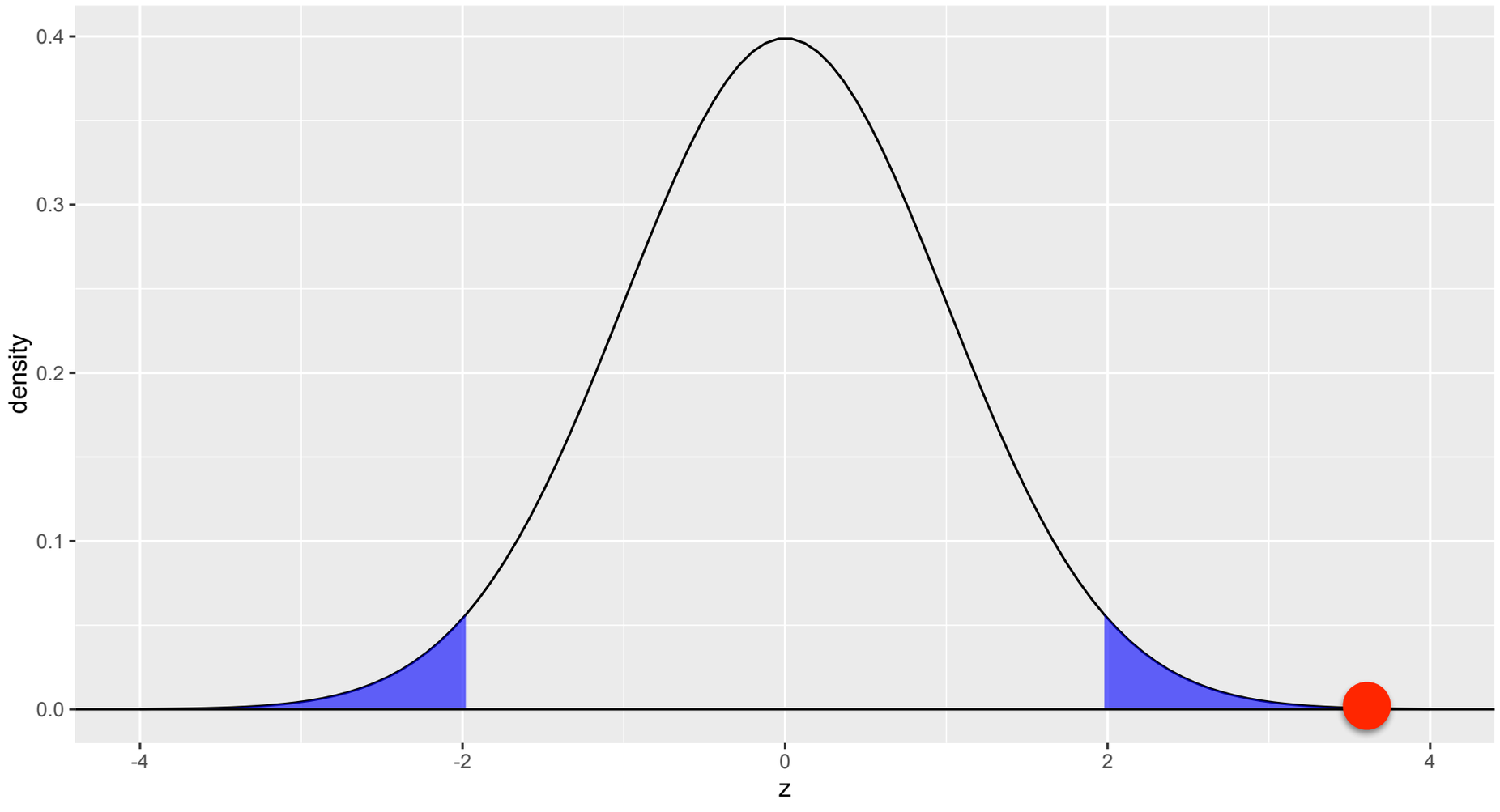
- People often use parametric tests (coefficients are drawn from a normal distribution) to assess this for logistic regression, but we can use it to illustrate another more robust test.

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 median height of men (=A) and women (=B)

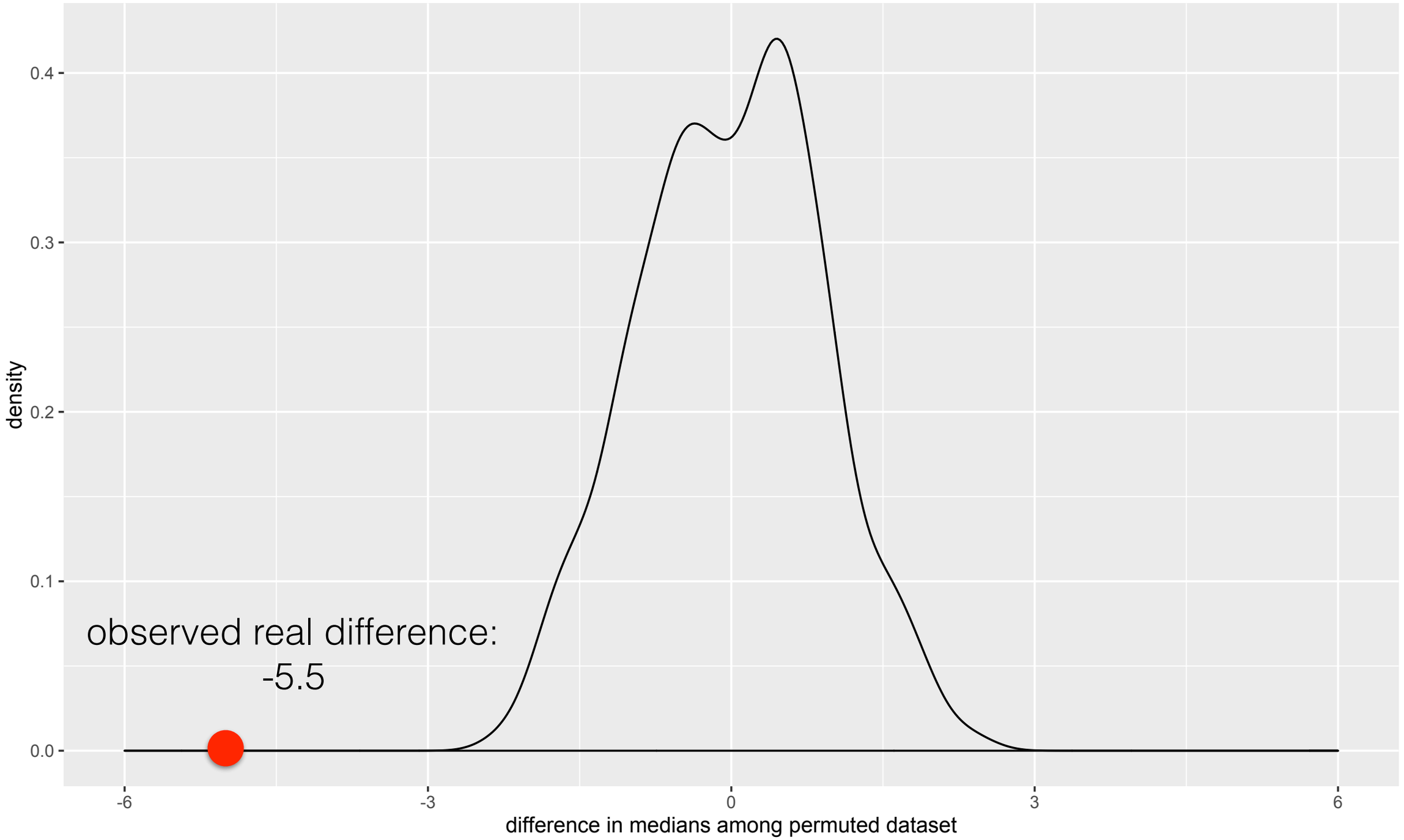
- We shuffle the labels of the data under the null assumption that the labels don’t matter (the null is that A = B)

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

observed true difference in medians: -5.5

		true	perm 1	perm 2	perm 3	perm 4	perm 5
x1	62.8	woman	man	man	woman	man	man
x2	66.2	woman	man	man	man	woman	woman
…	…	…	…	…	…	…	…
x9	68.4	man	woman	man	woman	man	woman
x10	70.9	man	woman	woman	woman	woman	woman
difference in medians:			4.7	5.8	1.4	2.9	3.3

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)

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:

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

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:

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

Observational data

- A survey of the political affiliation of Berkeley residents is observational data
    - the independent variable (living in Berkeley) is not under our control

- Tweets, books, surveys, the web, the census etc. — is all observational.

Observational data

- Hypothesis tests for observational data assess the relationship between variables but don't establish causality.

- Example: if we intervened and relocated someone to Berkeley, would they become liberal?

Experimental data

- Data that allows you to perform an intervention and determine the value of some variable
  - Clinical data: treatment vs. placebo
  - Web design: one of two homepage designs
  - Political email campaigns: one of two (differently worded) solicitations

Experimental data

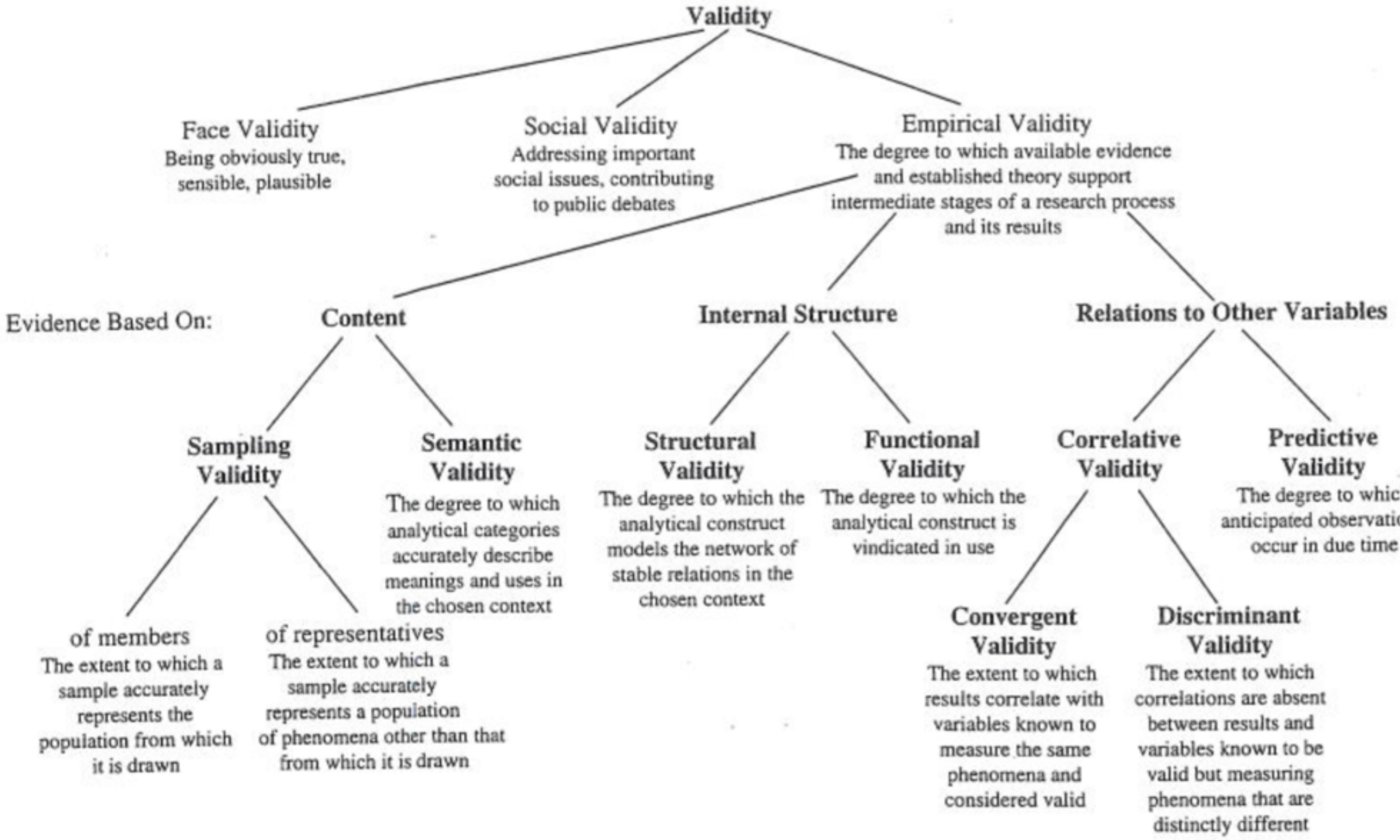
- A potential confound exists if any other variable is correlated with your intervention decision:

- e.g., users volunteering to receive a drug (and not the placebo)

Randomization experiments

- Users are randomly assigned an outcome (which web page), which allows us to better establish causality

- A/B testing = significance test in randomized experiment with two outcomes



Krippendorff (2004)

Face validity

- Does a finding “make sense” (in retrospect)?

- The “gatekeeper for all other kinds of validity”

Social validity

- Does a finding make a “contribution to the public discussion of important social concerns?”

Sampling validity

- Does a finding contain sample:
    - large enough to support its results?
    - not biased in the quantity of interest?

- e.g., Twitter

Semantic validity

- Does a finding ascribe meaning to its categories in a way that corresponds to how its subjects understand them?

- e.g., sentiment analysis, {democrat, republican}, libel

Structural validity

- Does a finding rely on methods that have internal coherence?

- e.g., fame from google books, historical argument

Functional validity

- Does a finding rely on a method that has a record of success?

Correlative validity

- Convergent validity: Does a finding correlate with another trusted variable?

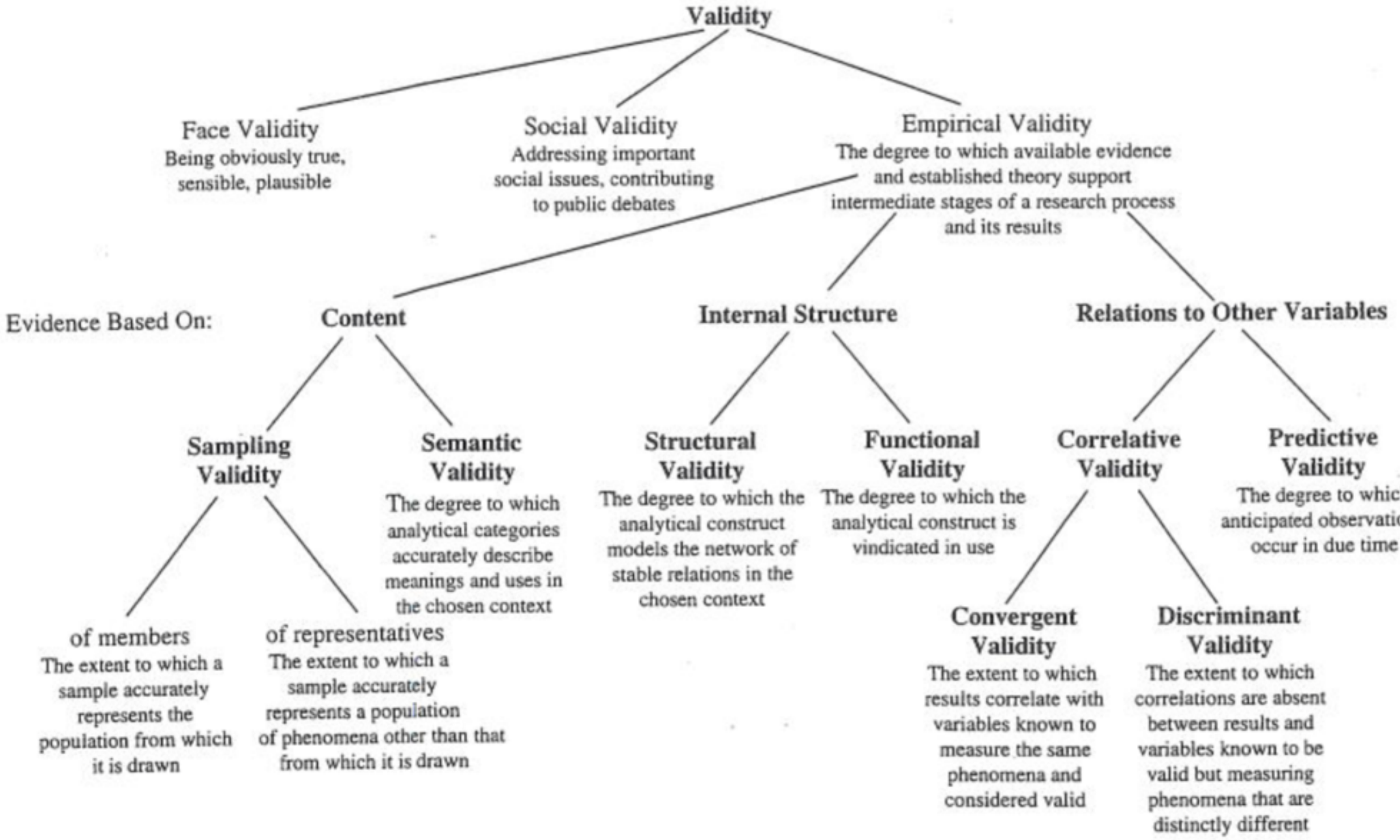
- Divergent validity: Does a finding not correlate with measures of *different* phenomena?

Predictive validity

- Does a finding make correct predictions about the future?

Validity

What other forms of validity should we add?



Krippendorff (2004)