Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 6: Validity

Feb 8, 2016

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

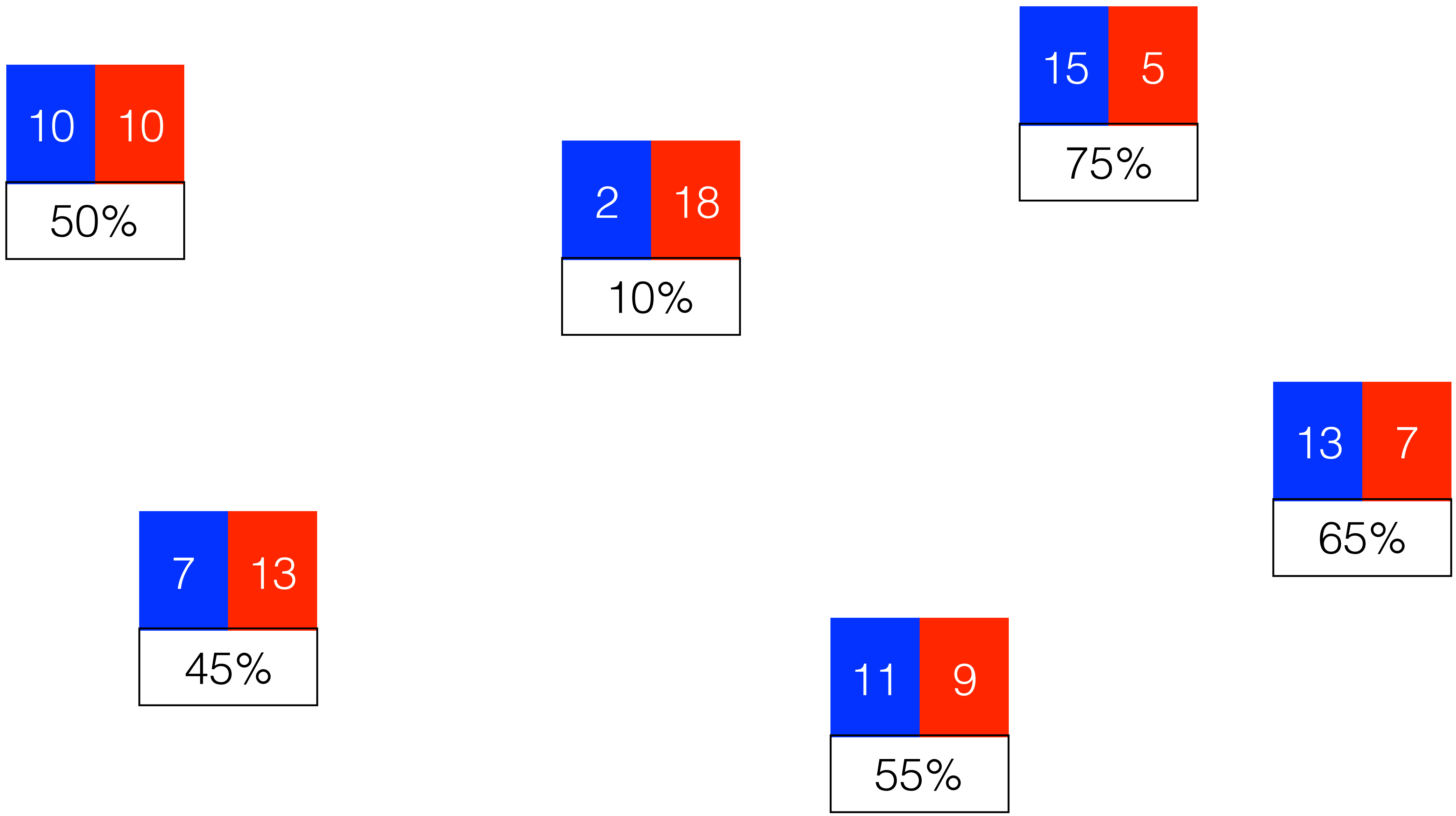
- H_0: Among all N registered {Democrat, Republican} primary voters, there are an equal number of Democrats and Republicans in Berkeley.

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

Example

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

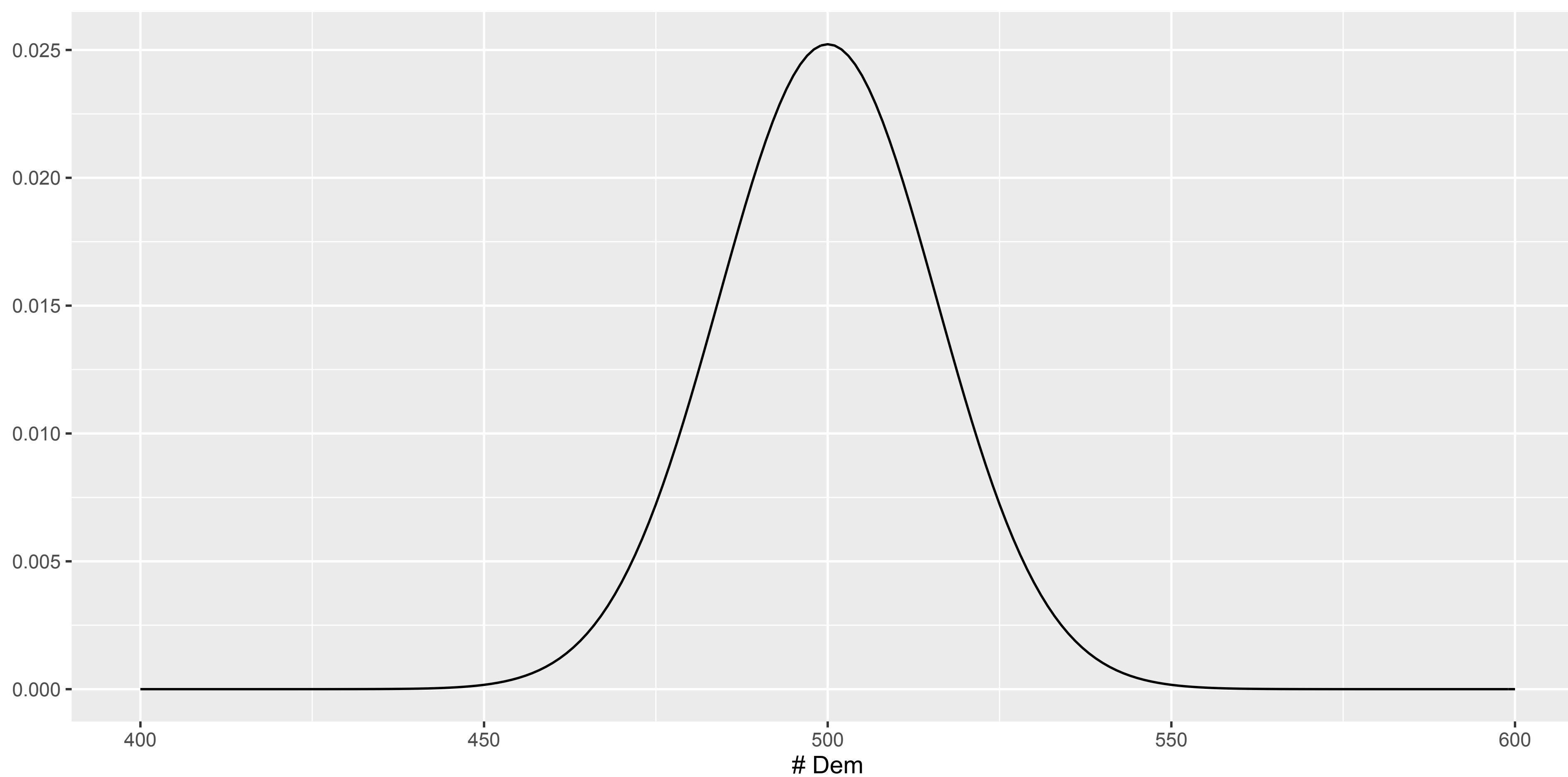
Example



Hypothesis testing

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

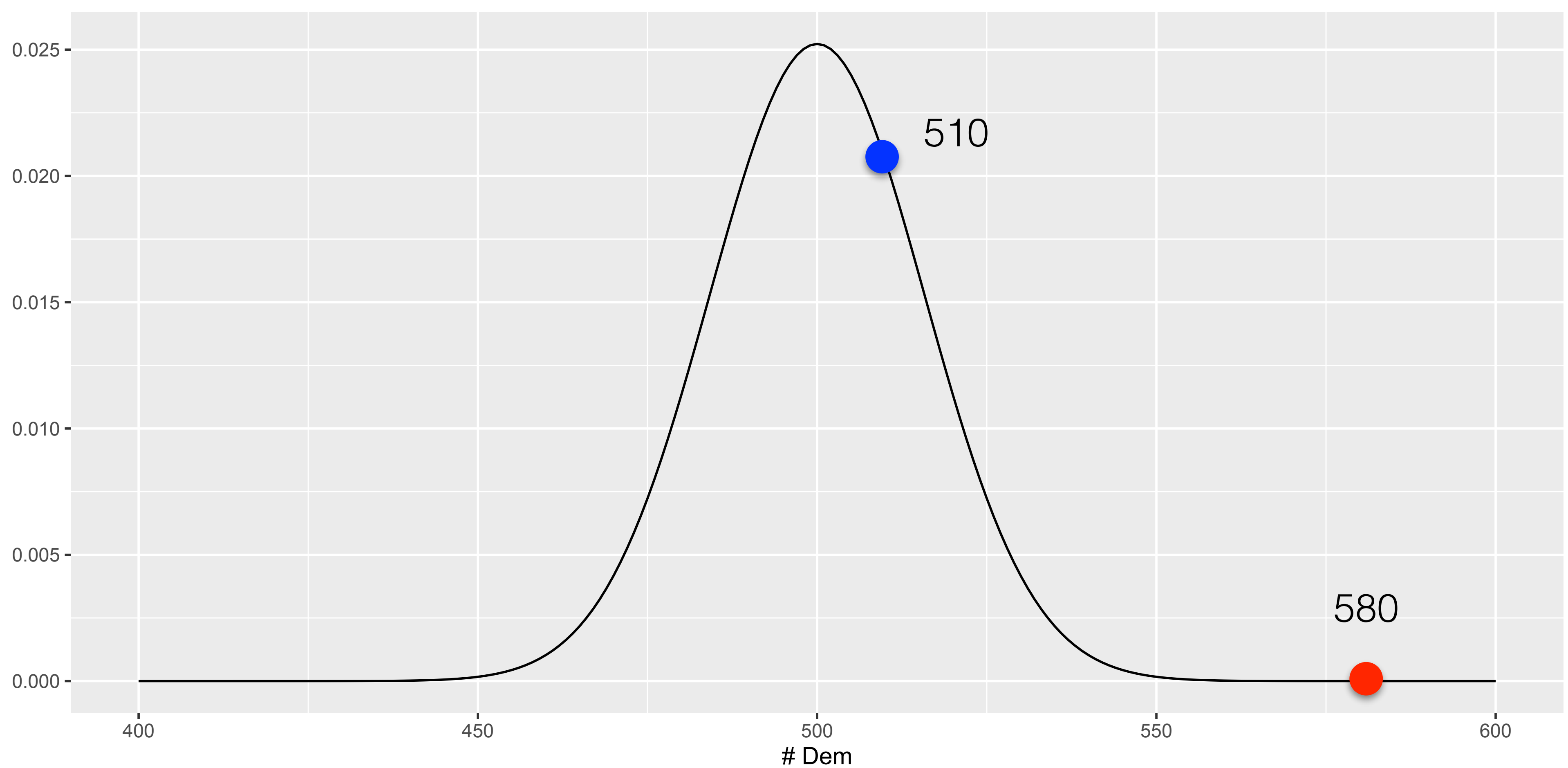
Example



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

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

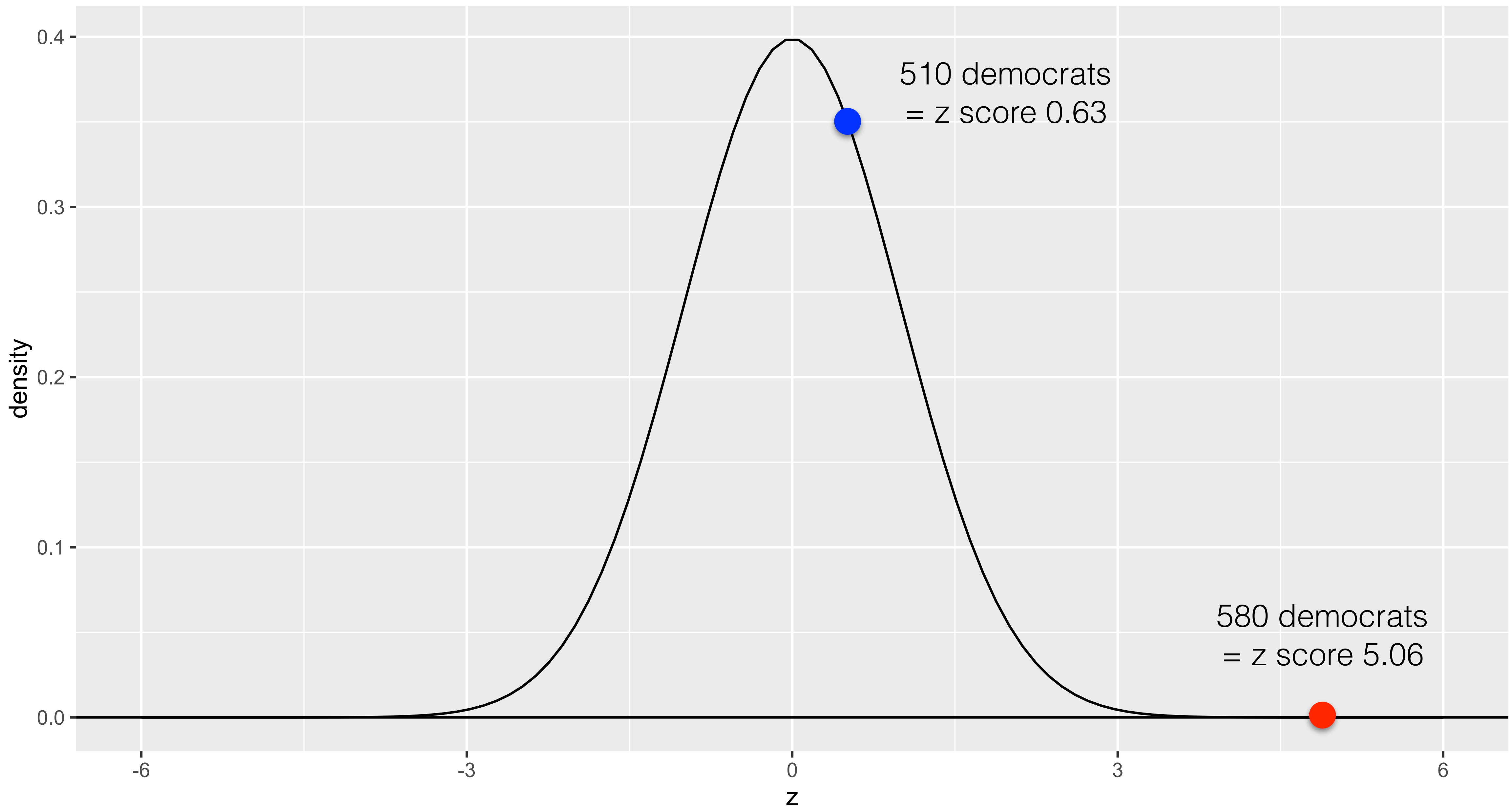
For Normal distributions, transform into standard normal (mean = 0, standard deviation =1)

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

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)

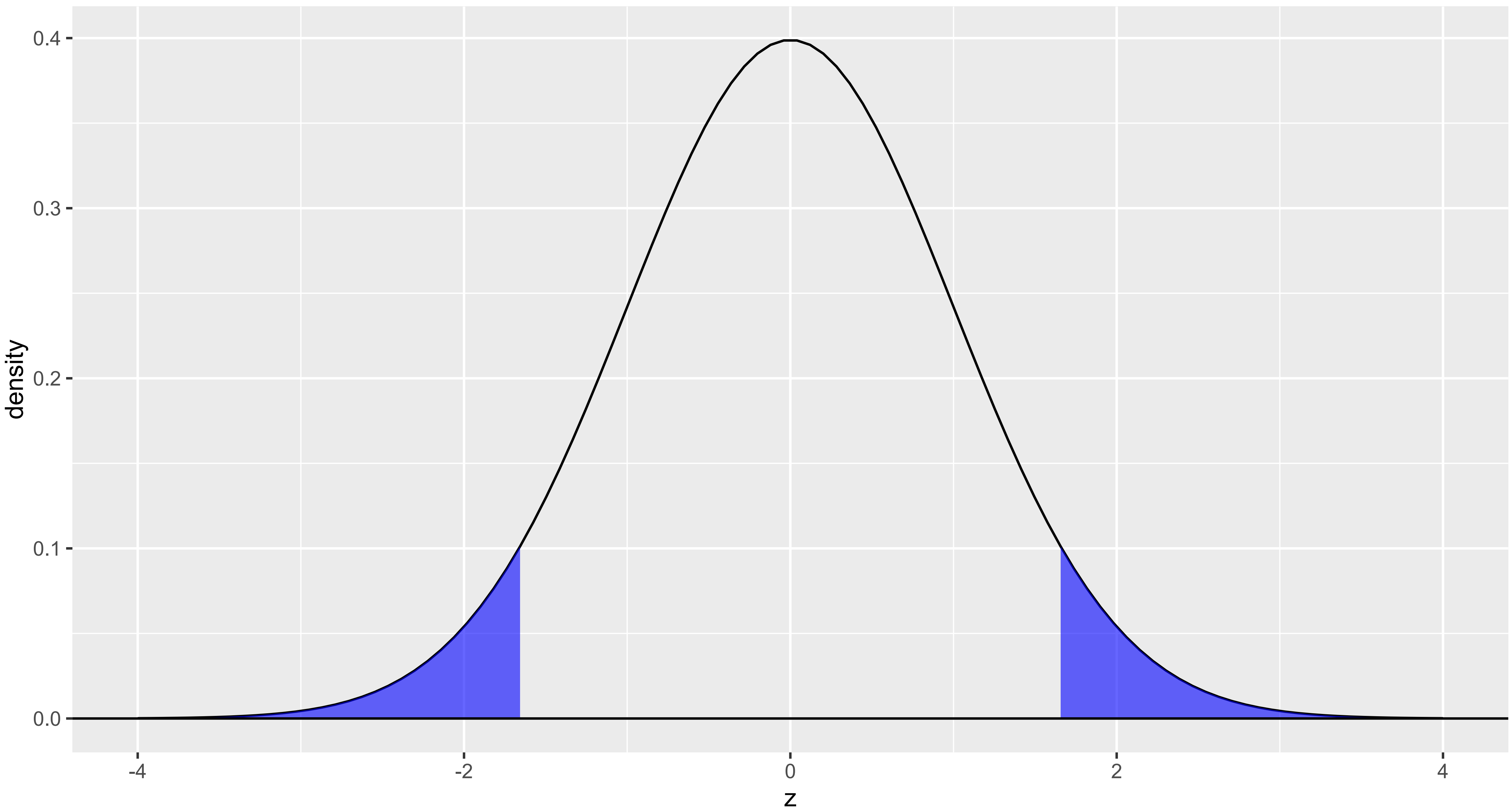
Z score


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

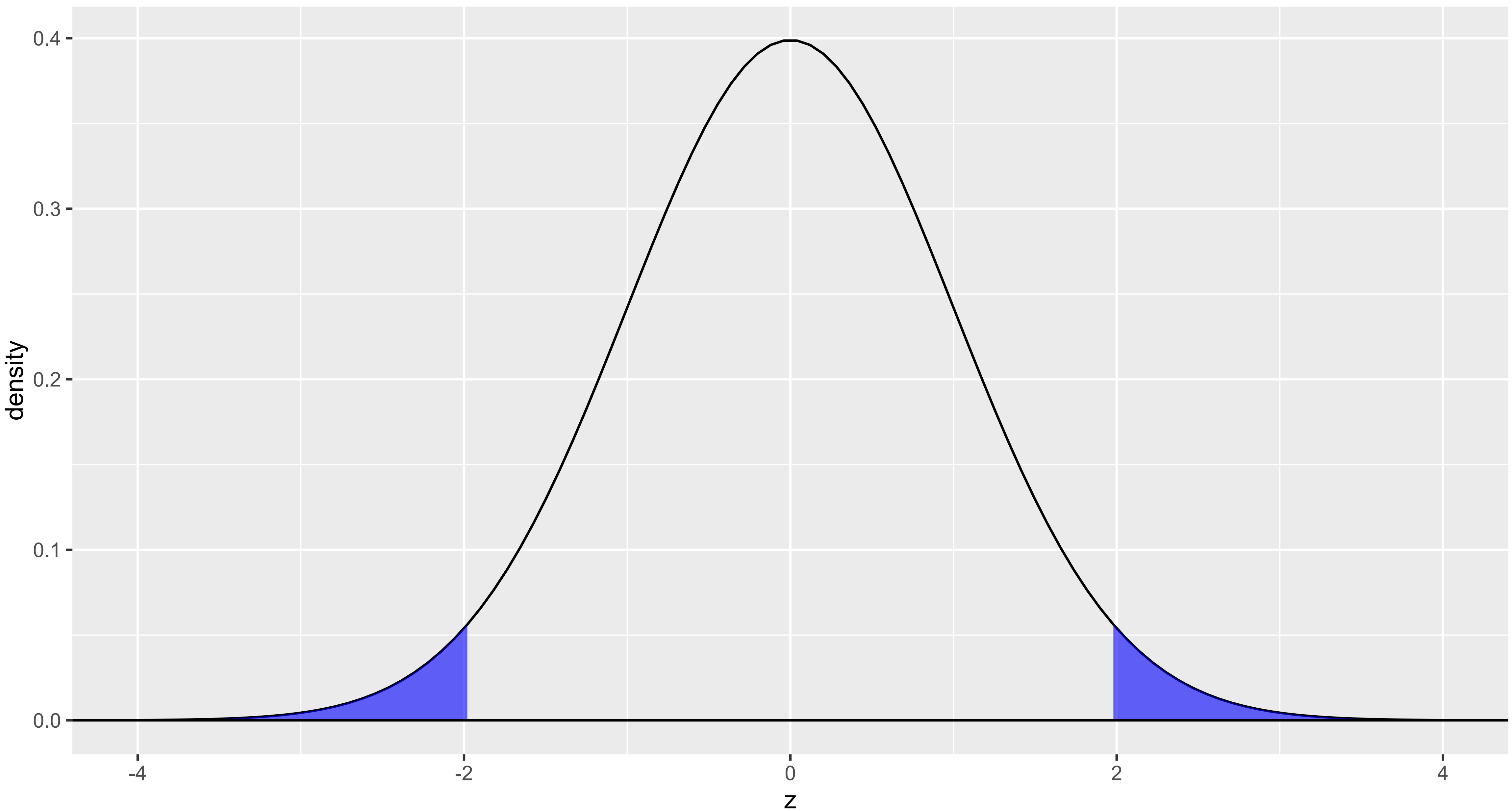
Tests

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

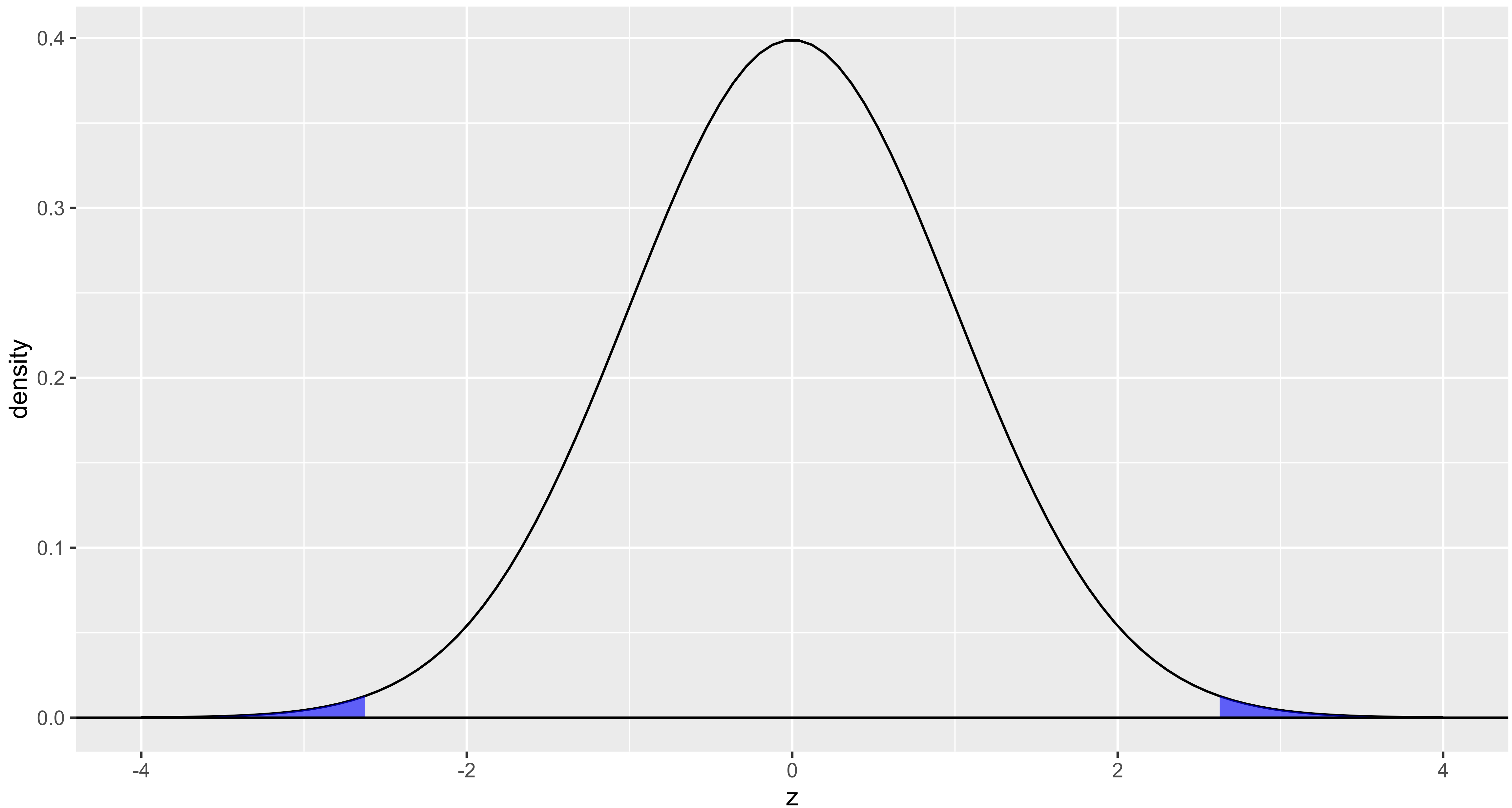
least likely 10%



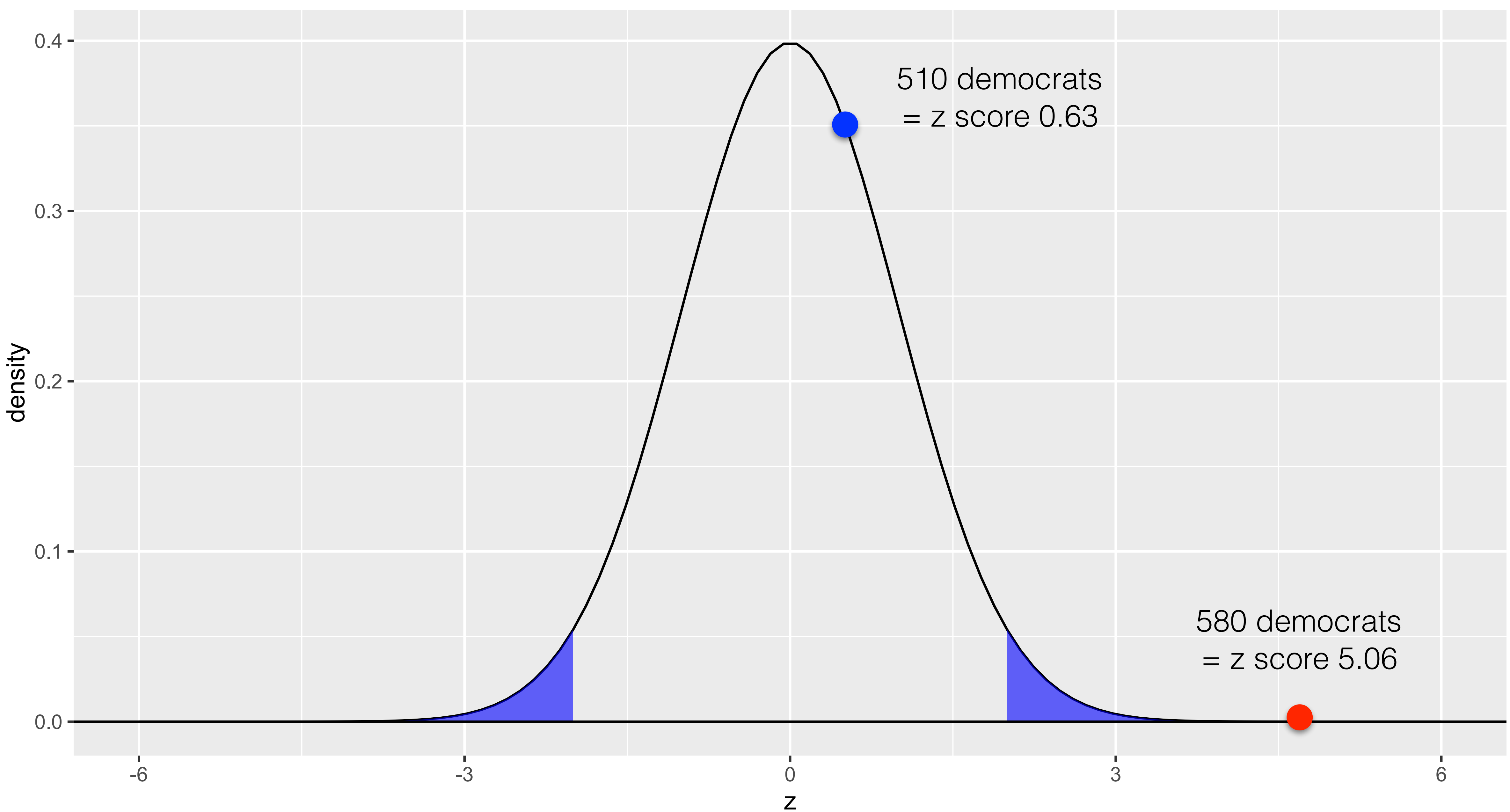
least likely 5%



least likely 1%



Tests

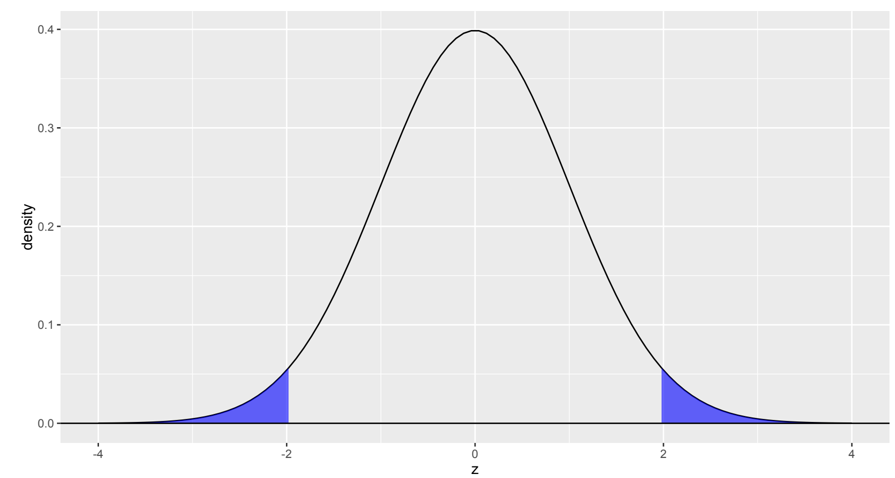

510 democrats
= z score 0.63
580 democrats
= z score 5.06
density
0.4
0.3
0.2
0.1
0.0
-6
-3
0
3
6
z

Tests

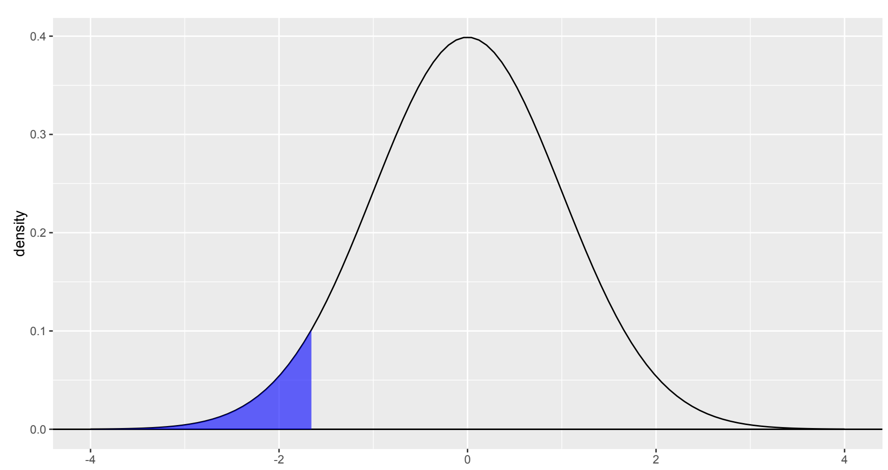
- Decide on the level of significance α. {0.05, 0.01}
- Testing is evaluating whether the sample statistic falls in the rejection region defined by α

Tails

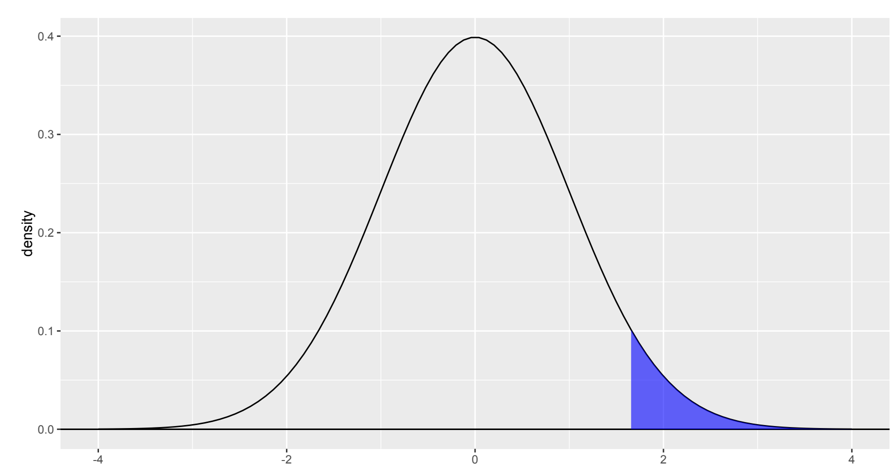
- Two-tailed tests measured whether the observed statistic is different (in either direction)
- One-tailed tests measure difference in a specific direction
- All differ in where the rejection region is located; $\alpha = 0.05$ for all.



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

Errors

		Test results	
		keep null	reject null
Truth	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 $\alpha \times n$ type I errors.
- $\alpha = 0.01$, $n = 1000$ = 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.]

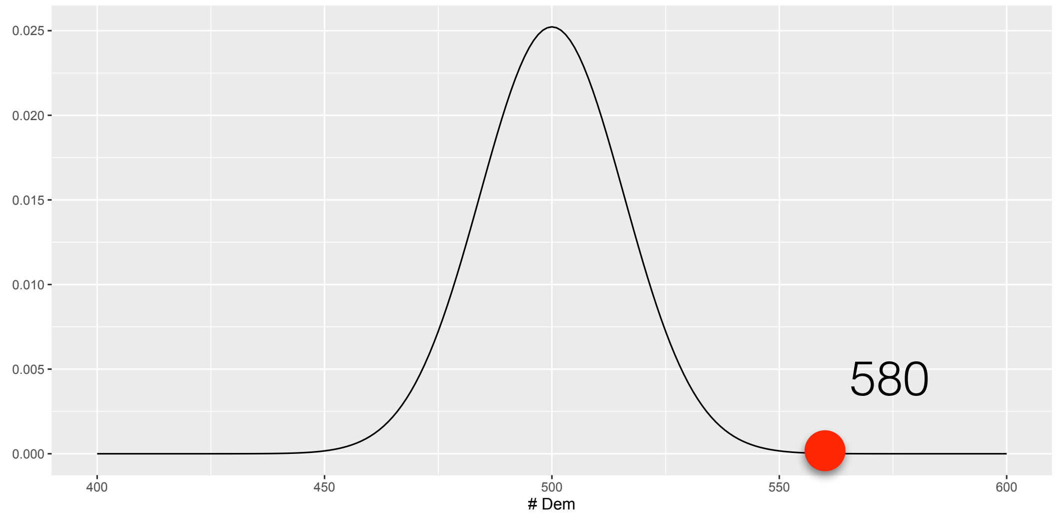
- False discovery rate

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

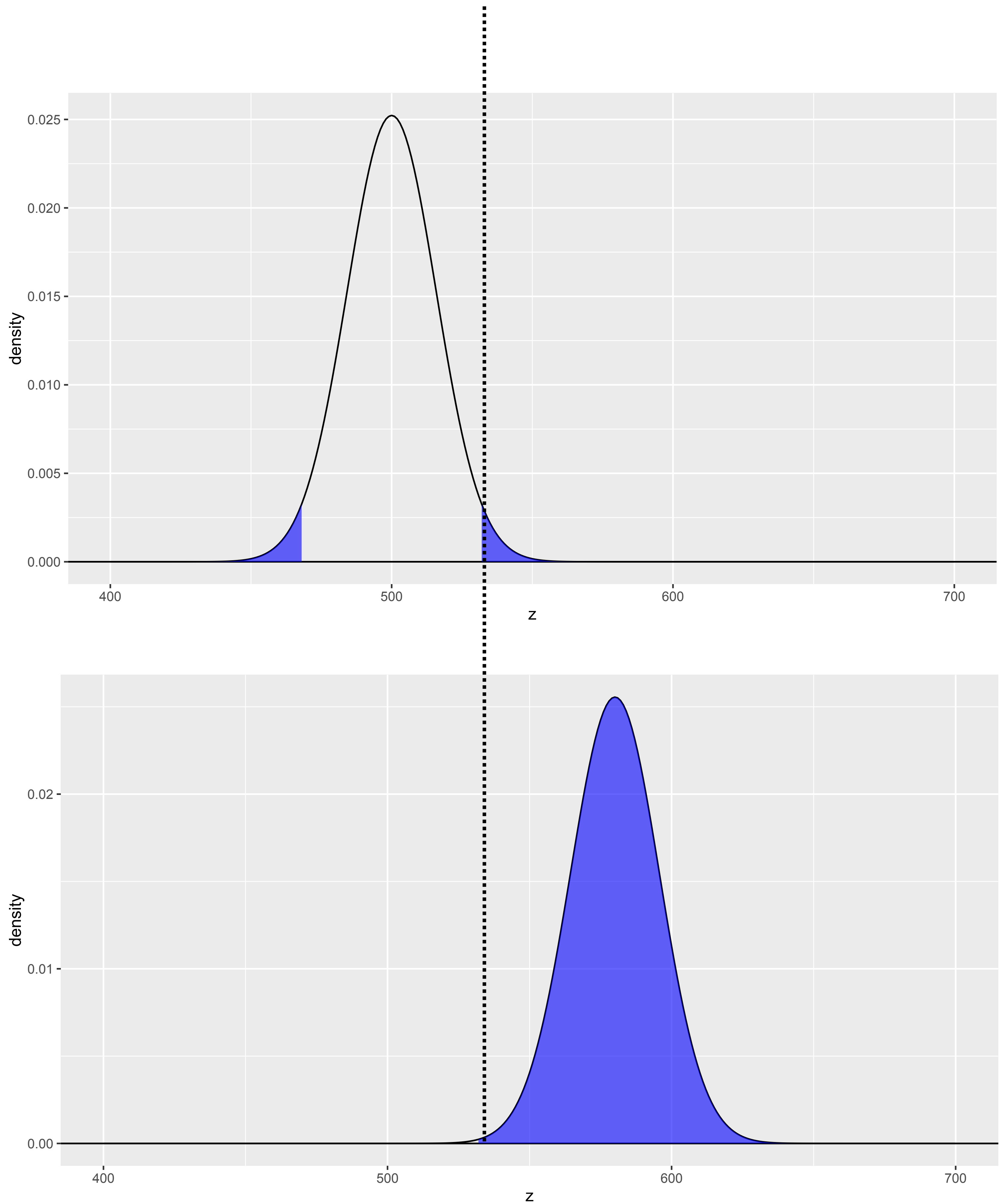
- Difference between the observed statistic and null hypothesis



null hypothesis	observed	effect size (%)	effect size (n)
0.50	0.58	0.08	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

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

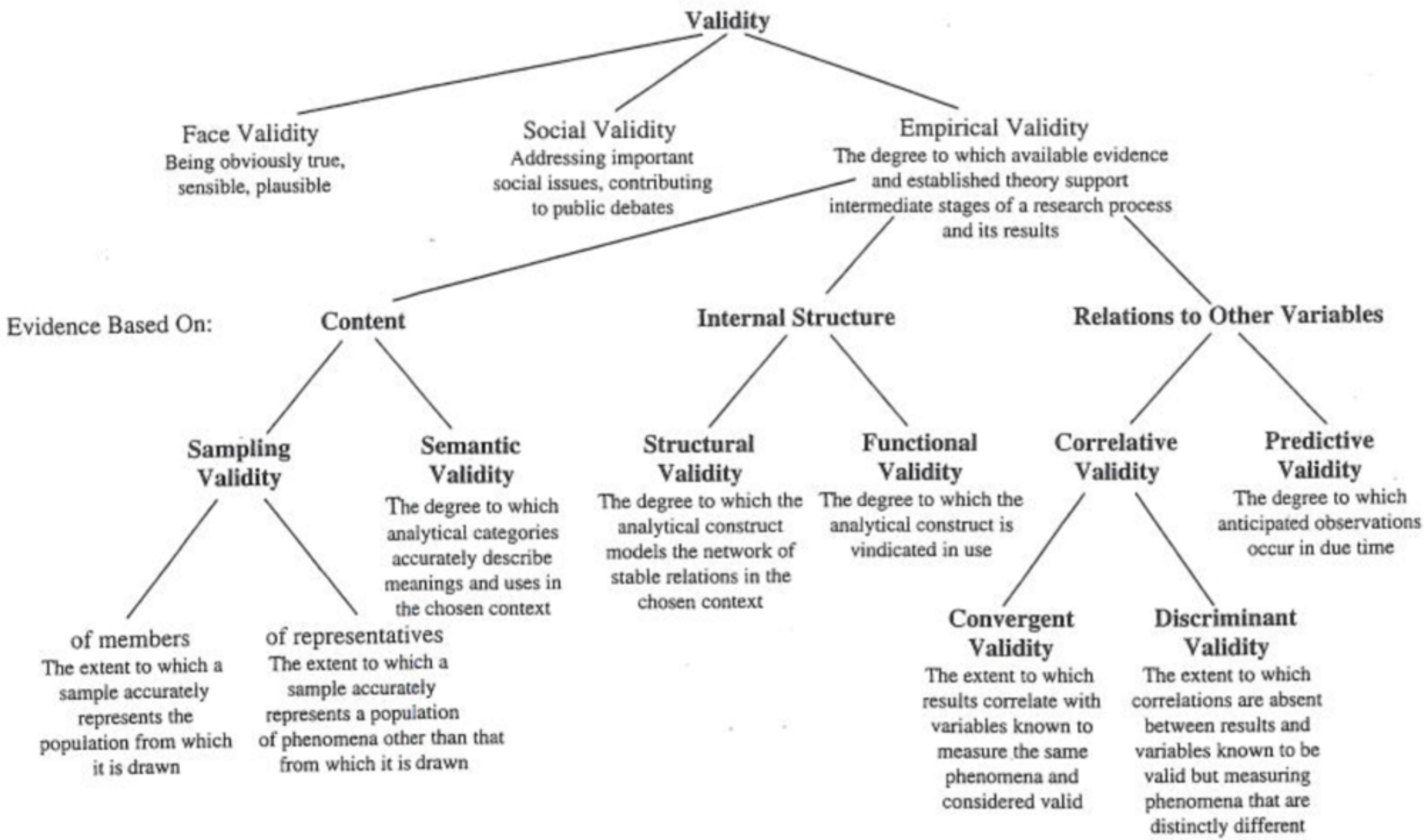
- A potential confound exists if any other variable is correlated with your intervention decision:

- 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

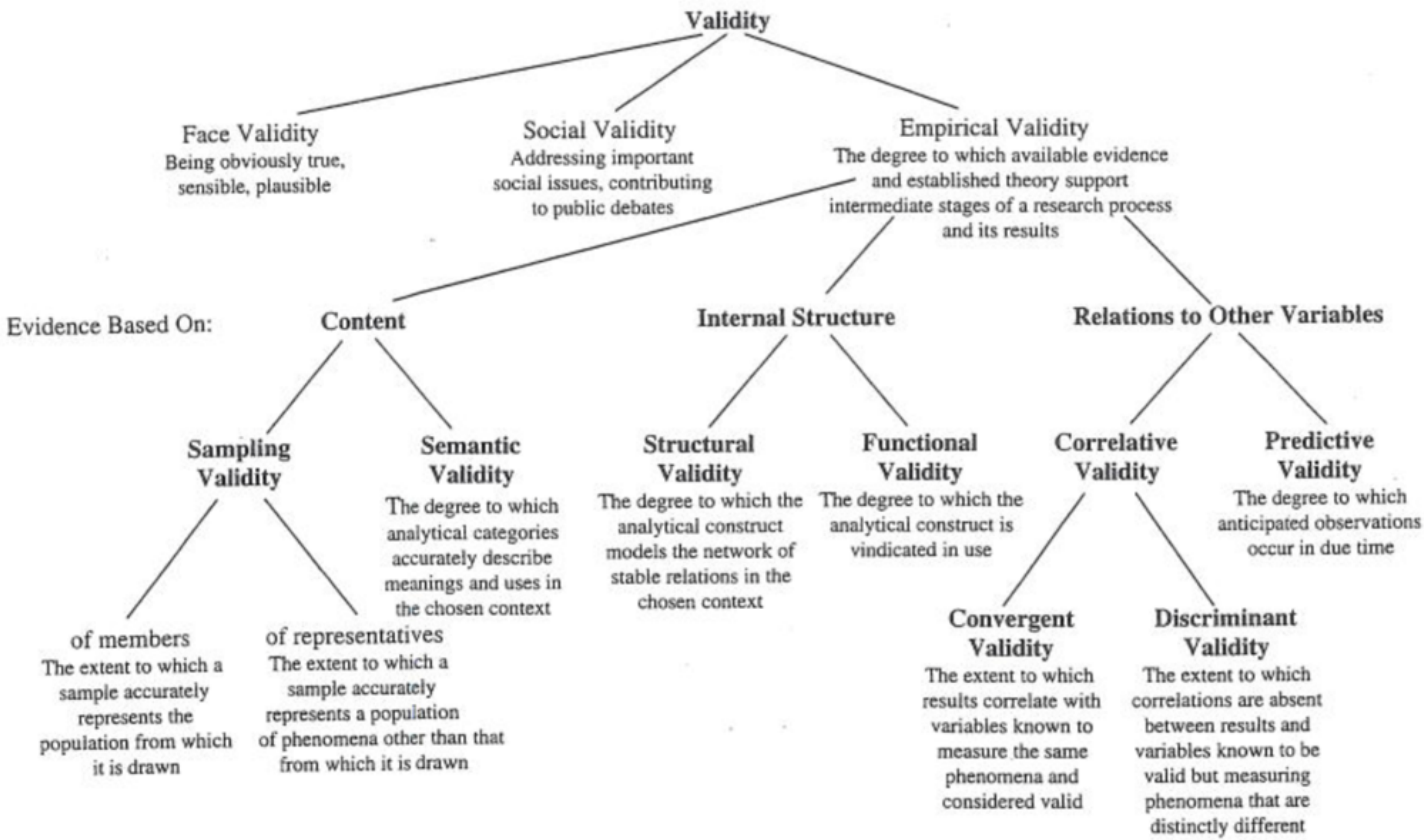
- Convergent validity: Does a finding correlate with another trusted variable?
- Divergent validity: Does a finding not correlate with measures of *different* phenomena?

Predictive validity

- Does a finding make correct predictions about the future?

Validity

What other forms of validity should we add?



Krippendorff (2004)

Homework 1, part I

- Creativity in conceptualizing what an "ideal" representation would look like, even if impractical.
- Originality in finding or imagining other types of potentially unusual data that could be included; alternatively, justification for the use of simplicity.
- Practice in the formulation of hypotheses (potential features that might be predictive) that can be justified a priori and then tested experimentally.
- Clarity in what counts as an "instance" for each of the nomination categories.
- Clarity in what counts as a "feature" that can be operationalized, and what constitutes sensible values for that feature.

Homework 1, part IIa

- Ability to operationalize the abstract features from part I into a tangible implementation.

- Ambition and creativity in the collection of data from which features can be instantiated

Homework 1, part IIb

- Understanding of the ways in which a human process can be understood as an "algorithm."

- Strong argument for the ways in which representation is consequential for learning.

- Strong argument for potential sources of bias.

- The use of specific mechanisms/techniques from data science to support your arguments