CS224C: NLP for CSS

Hypothesis Testing

Diyi Yang

Stanford CS

Lecture Overview

- Hypotheses
- Significance

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

Slides Credit to David Bamman

Hypotheses

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

This bestseller's book cover has changed a lot since 1998.

Voters in big cities prefer Clinton.

Email marketing pitch A is better than the pitch B.

Slides Credit to David Bamman

Null Hypothesis

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

Hypothesis "area	**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

Slides Credit to David Bamman

Hypothesis Testing

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

Slides Credit to David Bamman

Example

Hypothesis: Palo Alto residents tend to be politically liberal

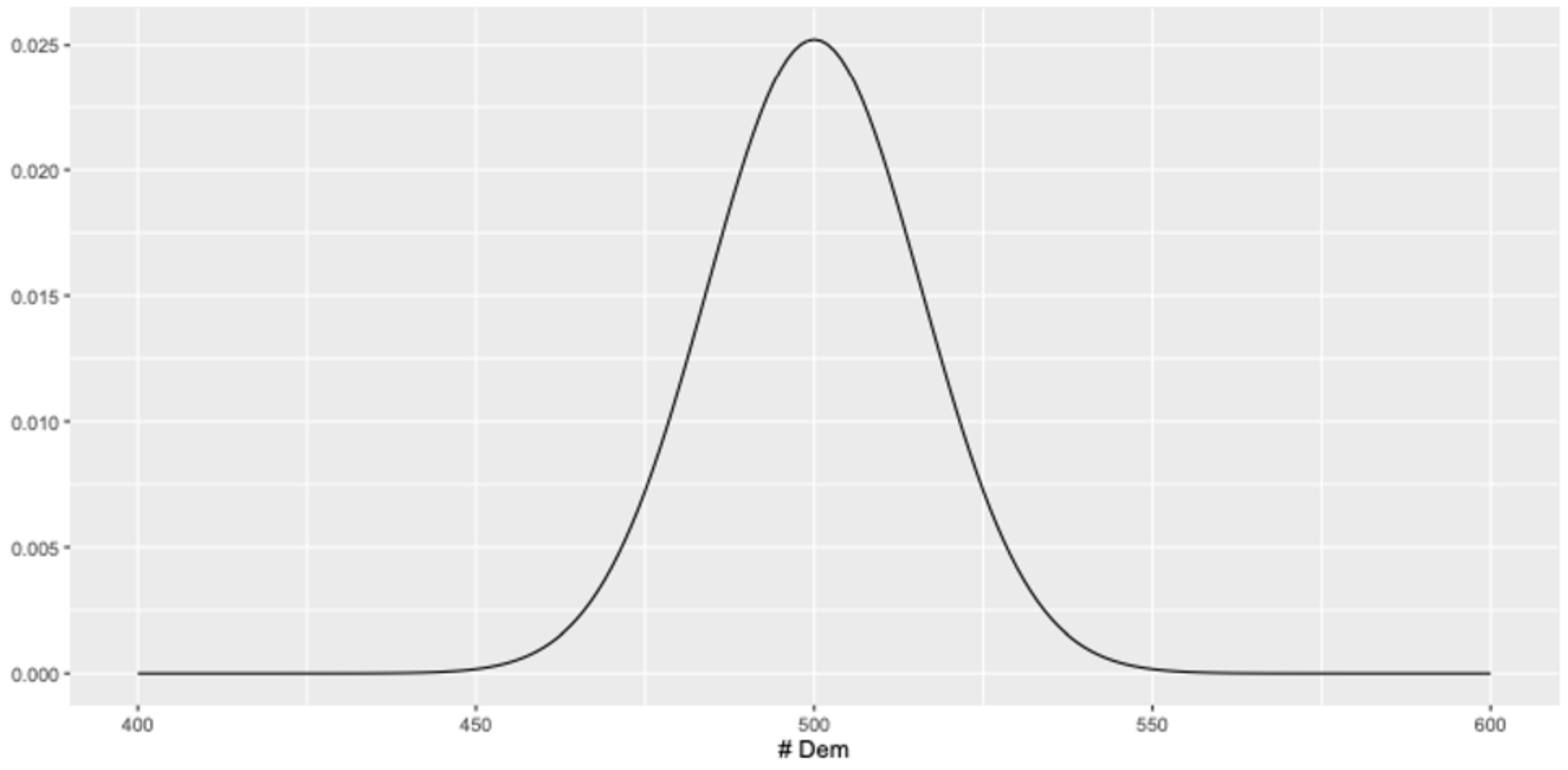
H_0**:** *Among all N registered {Democrat, Republican} primary voters, there are an equal number of Democrats and Republicans in Palo Alto*

$$\frac{N_{dem}}{N} = \frac{N_{Rep}}{N} = 0.5$$

Slides Credit to David Bamman

Hypothesis Testing

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

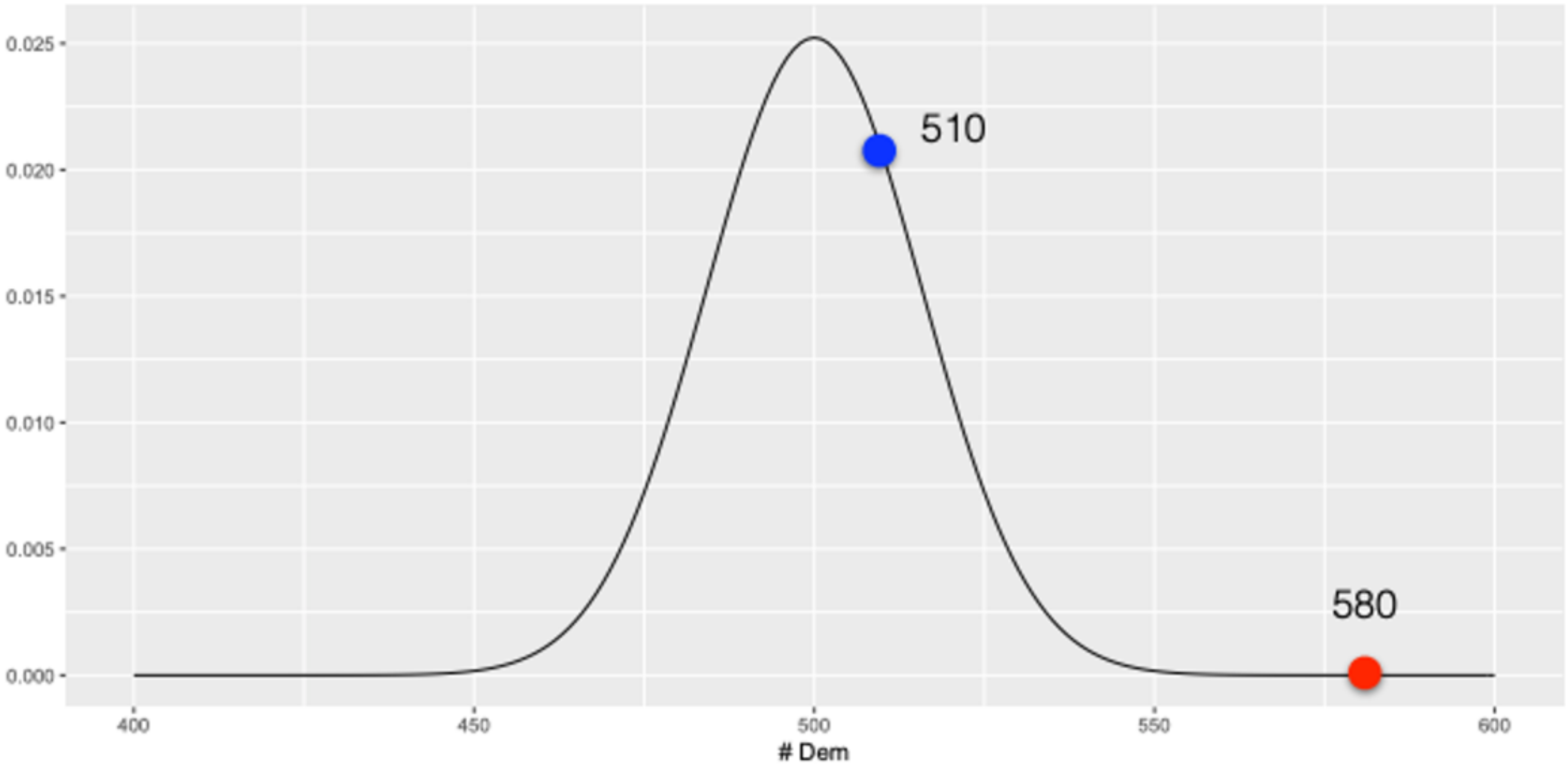


Binomial probability distribution for number of democrats in n=1000 with p = 0.5

Slides Credit to David Bamman

Example

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



Slides Credit to David Bamman

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

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

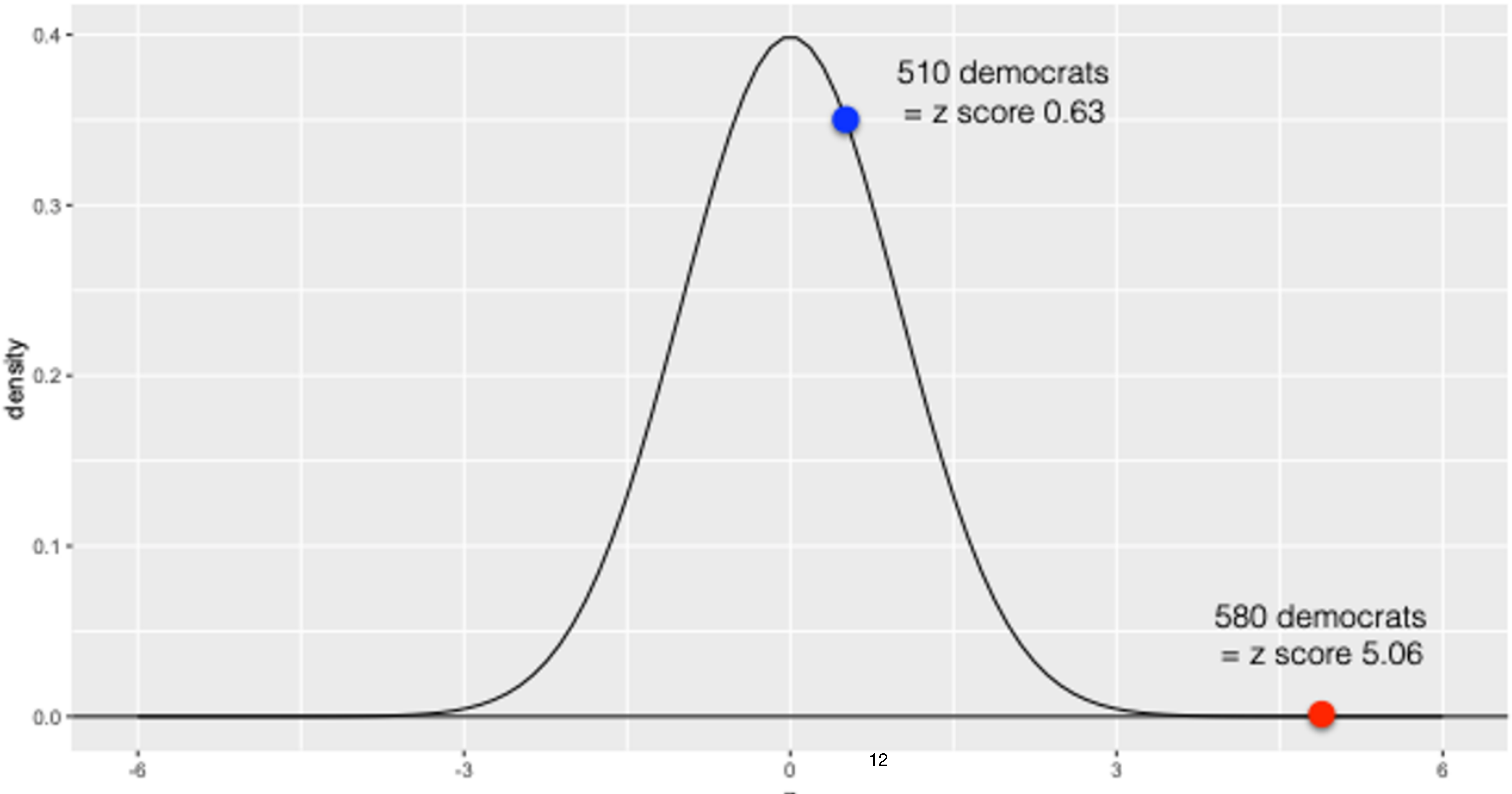
For normal distributions, transform into standard normal (mean=0, standard deviation=1)

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

For Binomial distributions, normal approximation (for large n)

Slides Credit to David Bamman

Z Score Example



Test and Significance Level

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

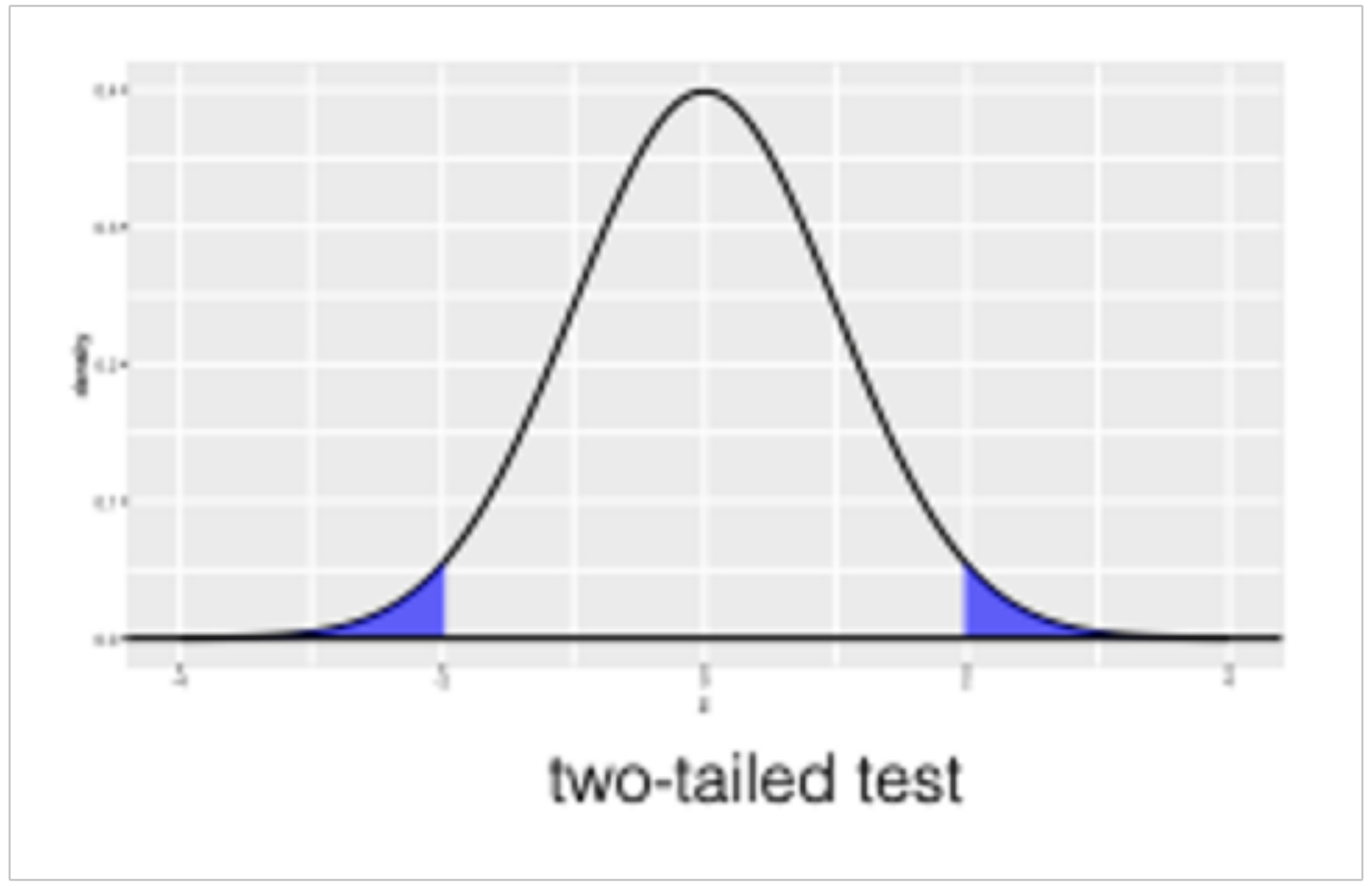
Slides Credit to David Bamman

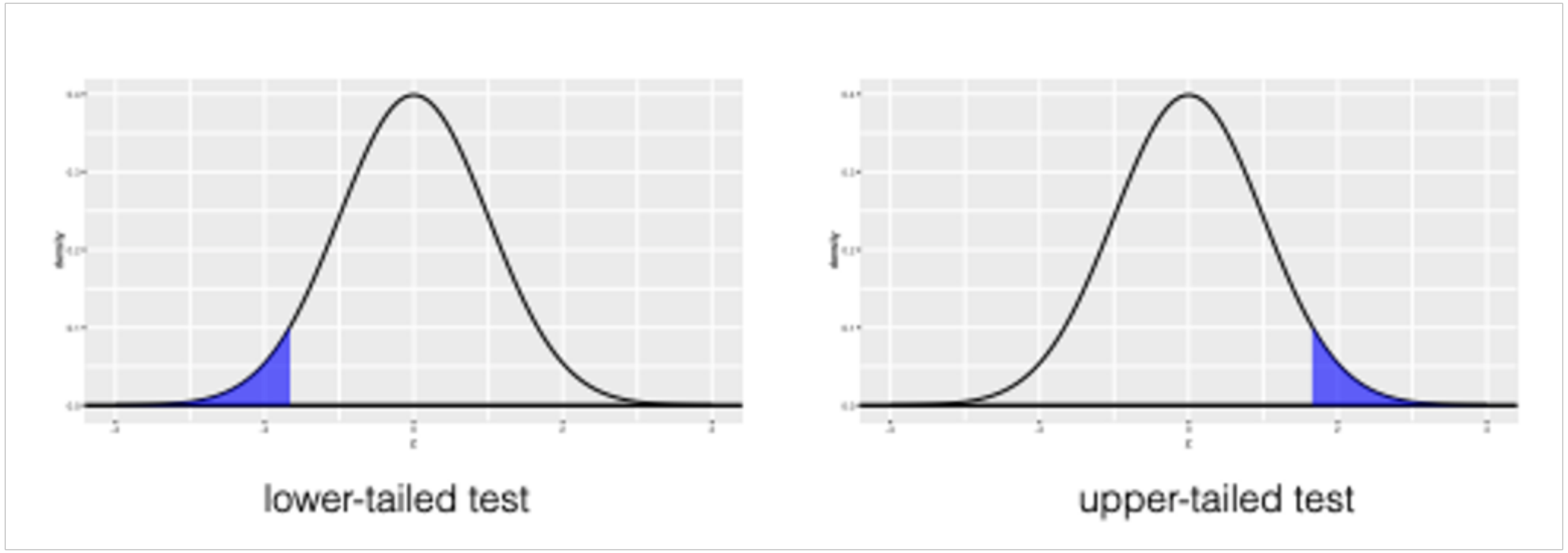
Tails

Two-tailed tests: whether the observed statistic is different (in either direction)

One-tailed tests: difference in a specific direction

All differ in where the rejection region is located: 0.05 for all





Slides Credit to David Bamman

P-Value

The p-value, or calculated probability, is the estimated probability of rejecting the null hypothesis H_0 when that hypothesis is true

Or the probability that the observed statistic occurred by chance alone

Slides Credit to David Bamman

P-Value

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

Lower-tailed test $p\text{-value}(z) = P(Z \leq z)$

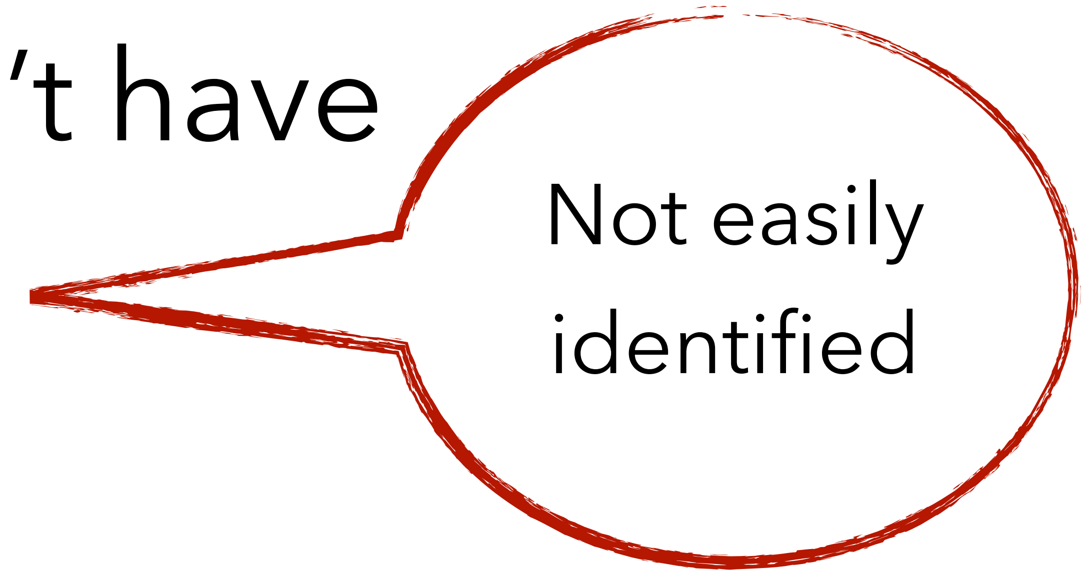
Upper-tailed test $p\text{-value}(z) = 1 - P(Z \leq z)$

Slides Credit to David Bamman

Errors are possible! **Error Types**

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



		Actual Situation	
		No Effect = H_0 True	*Effect Exists = H_0 False*
Researcher's Decision	*Reject H_0*	Type I error (α)	Decision correct
	Fail to reject H_0	Decision correct	Type II error (β)

Errors

For any significance level α and n hypothesis tests,

we can expect $\alpha \times n$ type I errors

α=0.01, n=1000 will lead to 10 “significant” results simply by chance

When would this occur in practice ?

Slides Credit to David Bamman

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

Slides Credit to David Bamman

Hypothesis Testing Summary

Step 1: State hypotheses and select alpha level

Step 2: Collect data; compute the test statistic

Step 3: Make a probability-based decision about $H_{0.}$ Reject H_0 if the test statistic is unlikely when H_0 is true ("*statistically significant*")

Reporting Significant Effect

A result is significant or statistically significant if it is very unlikely to occur when the null hypothesis is true, that is rejecting H_0

- Report that you found a significant effect
- Report value of test statistic
- Report the p-value of your test statistic

Non-parametric Tests

Many hypothesis tests rely on parametric assumptions (e.g., normality)

Alternatives that don’t rely on those assumptions

- Permutation test
- The Bootstrap

Significance of Coefficients

- A β_i value of 0 means that a 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

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

Slides Credit to David Bamman

Permutation Test

Non-parametric way of creating a null distribution 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)

Slides Credit to David Bamman

		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

Slides Credit to David Bamman

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

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

Slides Credit to David Bamman

Permutation Test

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

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

Permutation Test

To test whether the coefficients have a statistically significant effect (i.e., they are 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 test statistic β_p is more extreme than the observed test β

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

Bootstrap

Randomly sampling with replacement

1. Resample a data set x^* B times with replacement
2. Evaluate the bootstrap statistic $t(\cdot)$ each time
3. Approximate significance level via $\dfrac{t(\mathbf{x}^{*\mathbf{b}}) \geq t(\mathbf{x})}{B}$