Natural Language Processing

Info 159/259

Lecture 12: Features and hypothesis tests (Oct 3, 2017)

David Bamman, UC Berkeley

Announcements

- No office hours for DB this Friday (email if you'd like to chat)

Midterm 10/17

- In-class, 80 minutes.
- Covers all of class material and readings through 10/12; mix of multiple choice and short/long answer.
- No laptops or other devices; bring a single 8.5”x11” cheat sheet
- We’ll send out practice questions 10/12.

Midterm questions

- Email us sample questions (and answers)! We might put it in the exam. Completely voluntary/no credit, but if we pick it, you'll know the answer.

- Submit by 11:59pm next Tuesday 10/10.

MEMM Training

$$\prod_{i=1}^{n} P(y_i \mid y_{i-1}, x, \beta)$$

Locally normalized — at each time step, each conditional distribution sums to 1

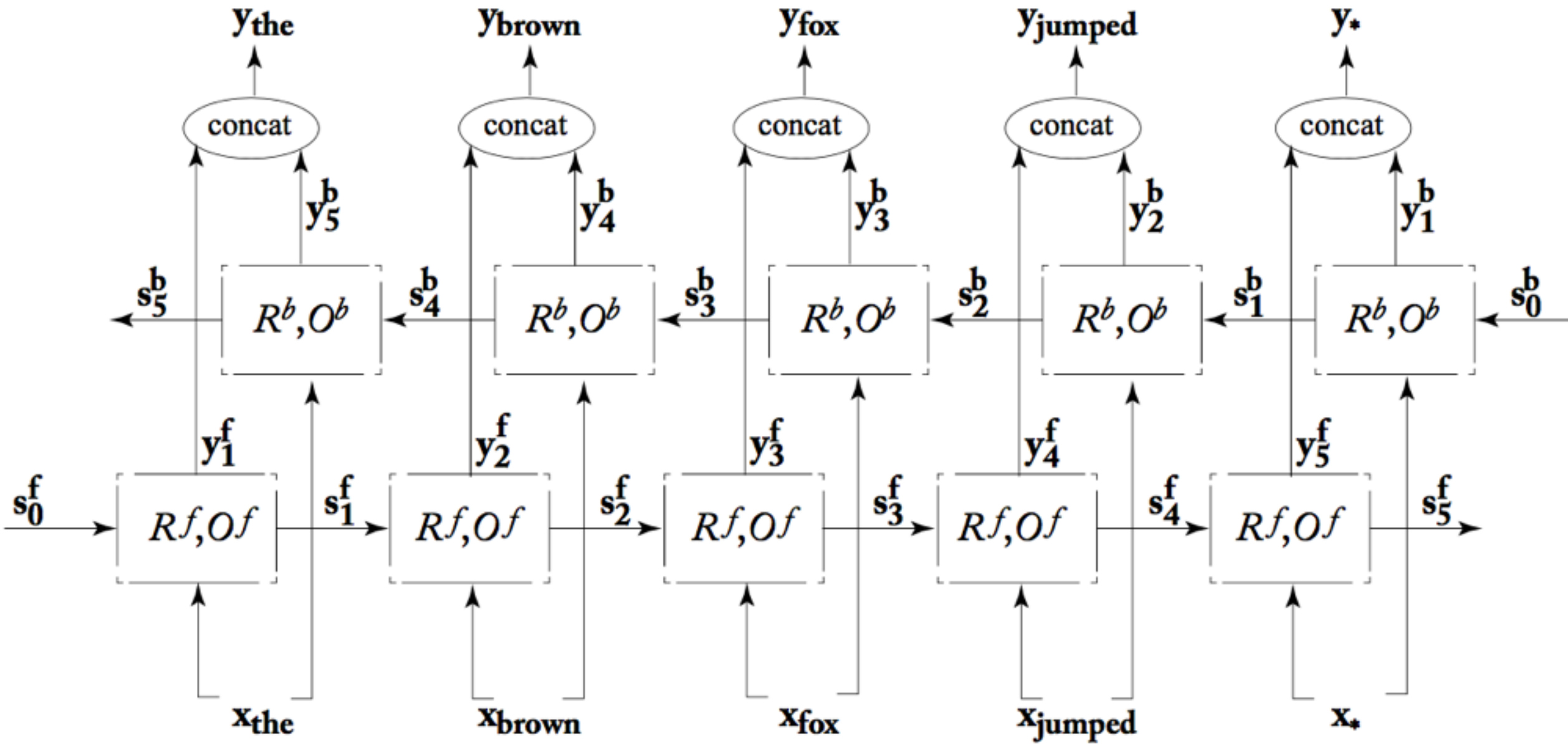
from last time

Conditional random fields

$$P(y \mid x, \beta) = \frac{\exp(\Phi(x, y)^\top \beta)}{\sum_{y' \in \mathcal{Y}} \exp(\Phi(x, y')^\top \beta)}$$

- In MEMMs, we normalize over the set of 45 POS tags
- CRFs are globally normalized, but the normalization complexity is huge — every possible sequence of labels of length n.

from last time



Goldberg 2017

from last time

Training BiRNNs

- Given this definition of an BiRNN:

$$s_b^i = R_b(x^i, s_b^{i+1}) = g(s_b^{i+1} W_b^s + x^i W_b^x + b_b)$$

$$s_f^i = R_f(x^i, s_f^{i-1}) = g(s_f^{i-1} W_f^s + x^i W_f^x + b_f)$$

$$y_i = \text{softmax}\left([s_f^i; s_b^i] W^o + b^o\right)$$

- We have 8 sets of parameters to learn (3 for each RNN + 2 for the final layer)

from last time

Significance in NLP

- You develop a new POS tagging algorithm; is it better than what comes before?

- You're developing a new model; should you include feature X? (when there is a cost to including it)

- You're developing a new model; does feature X reliably predict outcome Y?

Evaluation

- A critical part of development new algorithms and methods and demonstrating that they work

Classification

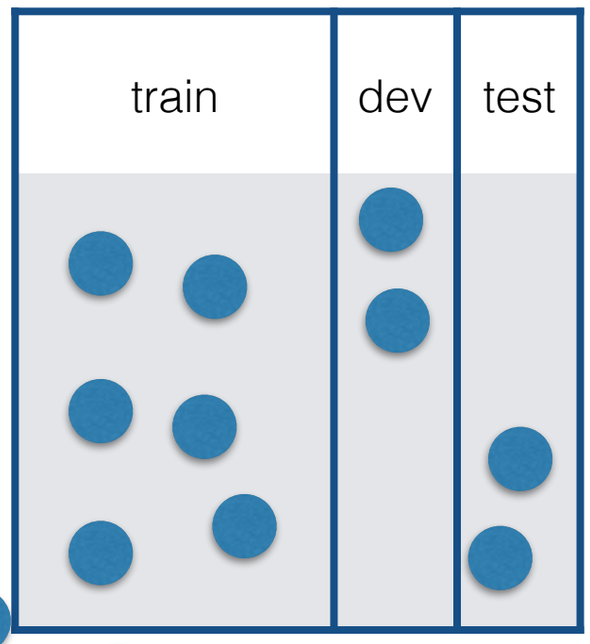
A mapping *h* from input data x (drawn from instance space $\mathcal{X}$) to a label (or labels) y from some enumerable output space $\mathcal{Y}$

$\mathcal{X}$ = set of all documents
$\mathcal{Y}$ = {english, mandarin, greek, …}

x = a single document
y = ancient greek

$\mathcal{X}$

instance space



Experiment design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Metrics

- Evaluations presuppose that you have some metric to evaluate the fitness of a model.
    - Language model: perplexity
    - POS tagging/NER: accuracy, precision, recall, F1
    - Phrase-structure parsing: PARSEVAL (bracketing overlap)
    - Dependency parsing: Labeled/unlabeled attachment score
    - Machine translation: BLEU, METEOR
    - Summarization: ROUGE

Metrics

- Downstream tasks that use NLP to predict the natural world also have metrics:
    - Predicting presidential approval rates from tweets.
    - Predicting the type of job applicants from a job description.
    - Conversational agent

Multiclass confusion matrix

True (y) \ Predicted (ŷ)	NN	VBZ	JJ
NN	100	2	15
VBZ	0	104	30
JJ	30	40	70

Accuracy

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

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

Predicted (ŷ)

True (y)	NN	VBZ	JJ
NN	100	2	15
VBZ	0	104	30
JJ	30	40	70

Precision

Precision(NN) =

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

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

True (y) \ Predicted (ŷ)	NN	VBZ	JJ
NN	100	2	15
VBZ	0	104	30
JJ	30	40	70

Recall

Recall(NN) =

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

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

Predicted (ŷ)

True (y)

	NN	VBZ	JJ
NN	100	2	15
VBZ	0	104	30
JJ	30	40	70

F score

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

Ablation test

- To test how important individual features are (or components of a model), conduct an ablation test
  - Train the full model with all features included, conduct evaluation.
  - Remove feature, train reduced model, conduct evaluation.

Ablation test

	Dev.	Test	
Our tagger, all features	88.67	89.37	
independent ablations:			
−DISTSIM	87.88	88.31	(−1.06)
−TAGDICT	88.28	88.31	(−1.06)
−TWORTH	87.51	88.37	(−1.00)
−METAPH	88.18	88.95	(−0.42)
−NAMES	88.66	89.39	(+0.02)
Our tagger, base features	82.72	83.38	
Stanford tagger	85.56	85.85	
Annotator agreement	92.2		

Table 2: Tagging accuracies on development and test data, including ablation experiments. Features are ordered by importance: test accuracy decrease due to ablation (final column).

Gimpel et al. 2011, “Part-of-Speech Tagging for Twitter”

Significance

Your work	58%
Current state of the art	50%

- If we observe difference in performance, what's the cause? Is it because one system is better than another, or is it a function of randomness in the data? If we had tested it on other data, would we get the same result?

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
Your system X is better than state-of-the-art system Y

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
Your system X is better than state-of-the-art system Y	There is no difference in the two systems.

Hypothesis testing

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

Hypothesis testing

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

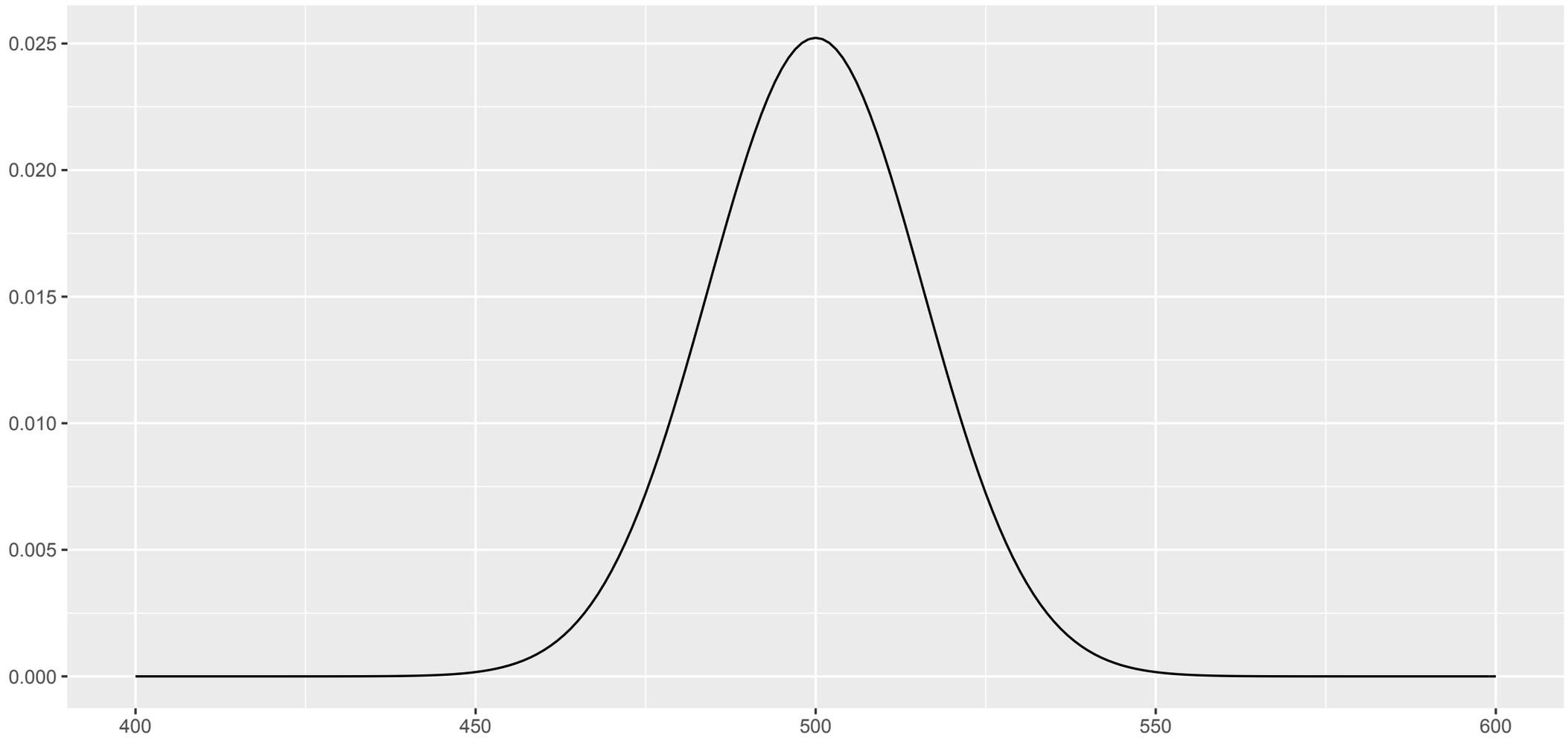
Hypothesis testing

- Current state of the art = 50%; your model = 58%. Both evaluated on the same test set of 1000 data points.

- Null hypothesis = there is no difference, so we would expect your model to get 500 of the 1000 data points right.

- If we make parametric assumptions, we can model this with a Binomial distribution (number of successes in n trials)

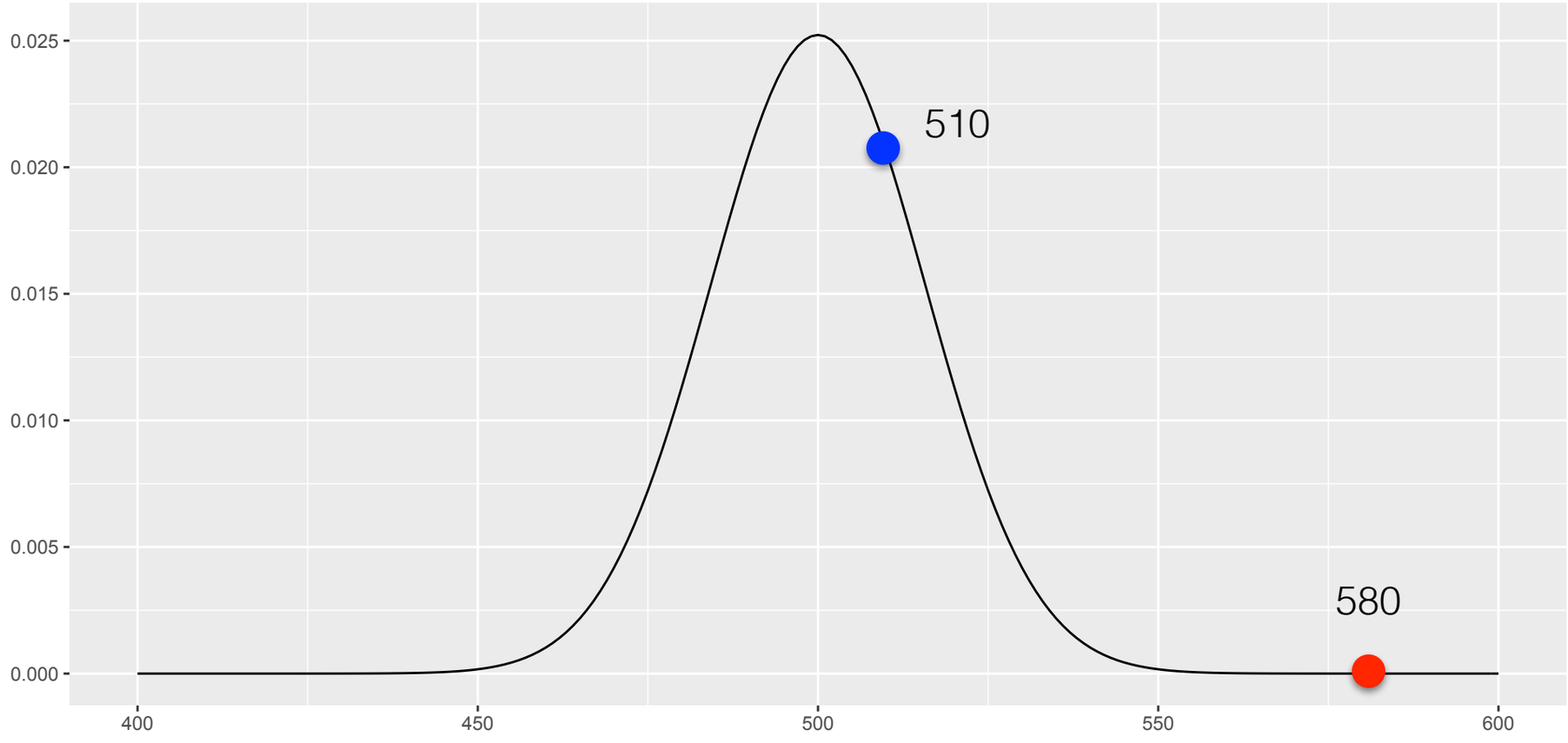
Example



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?



Example

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

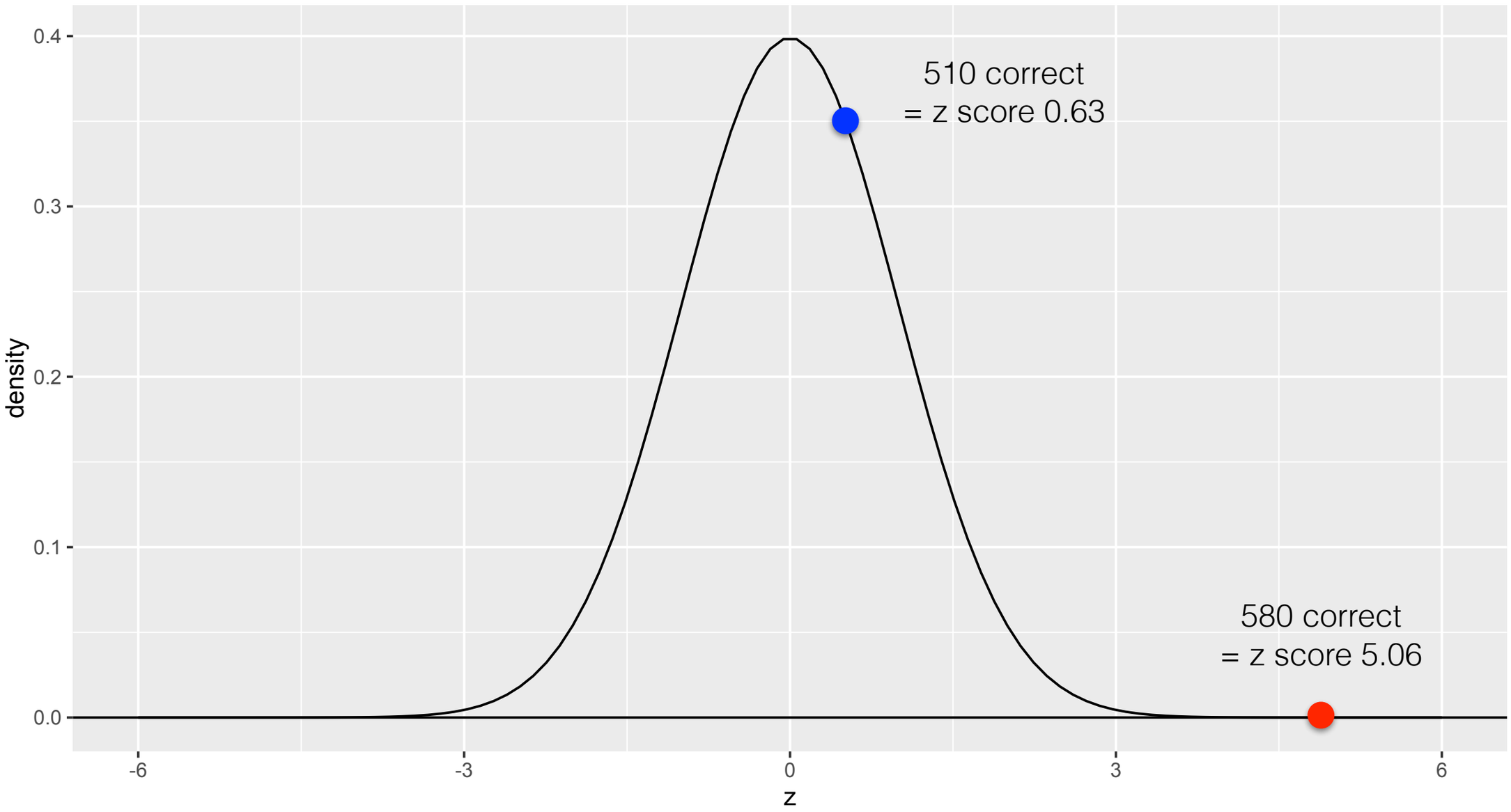
- We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size n; for large n, we can often make a normal approximation.

Z score

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

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

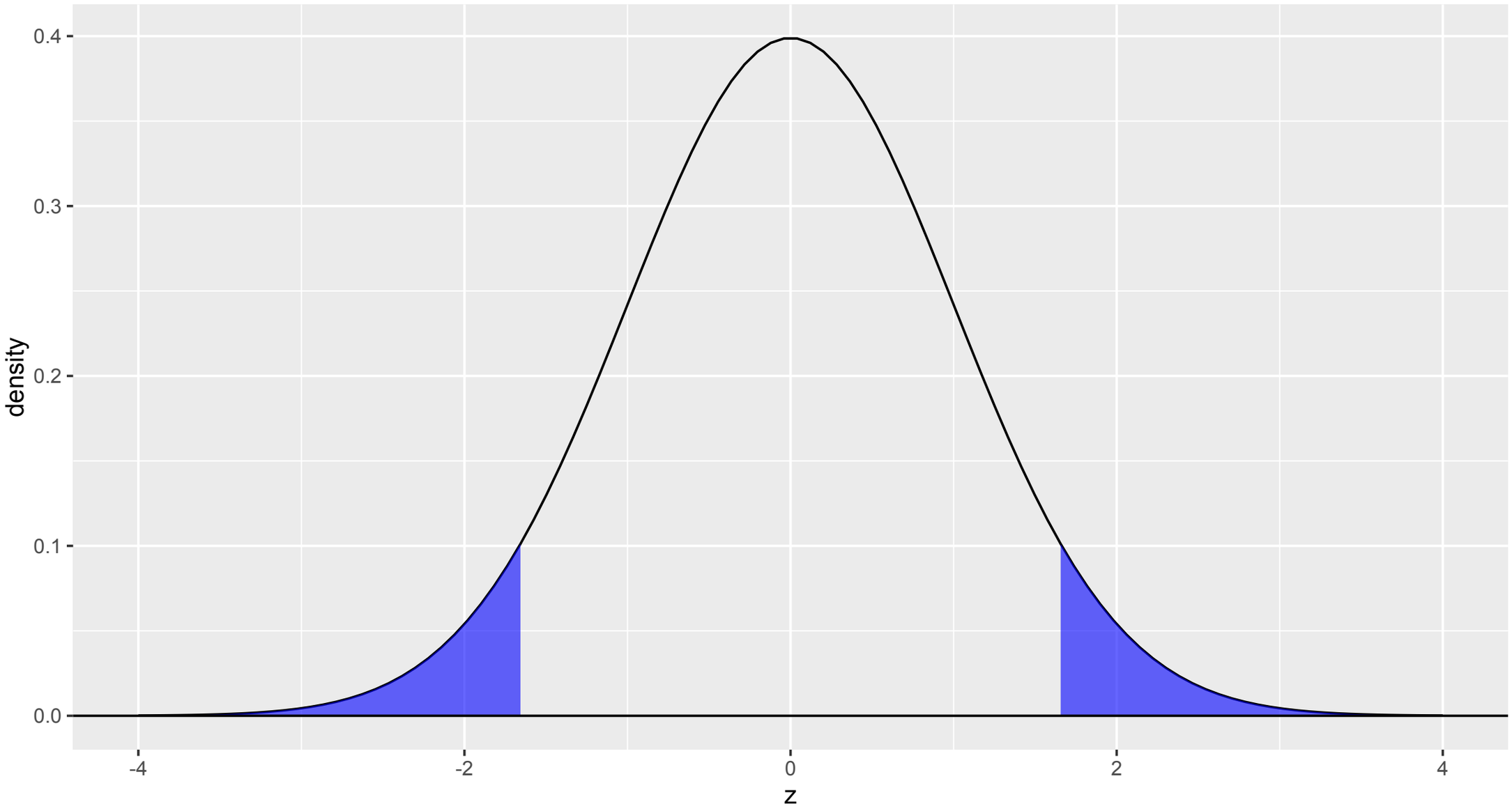
Z score


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

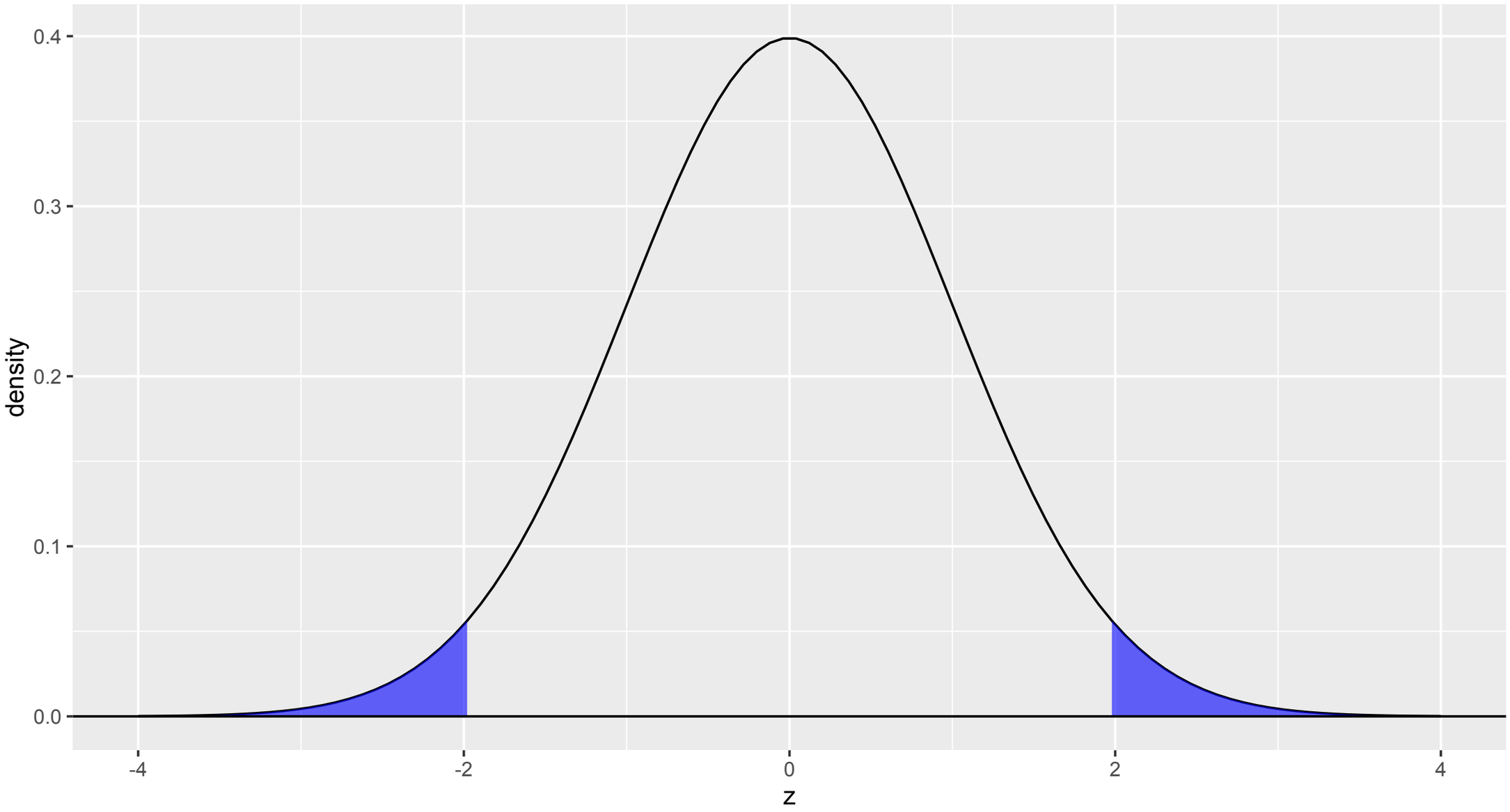
Tests

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

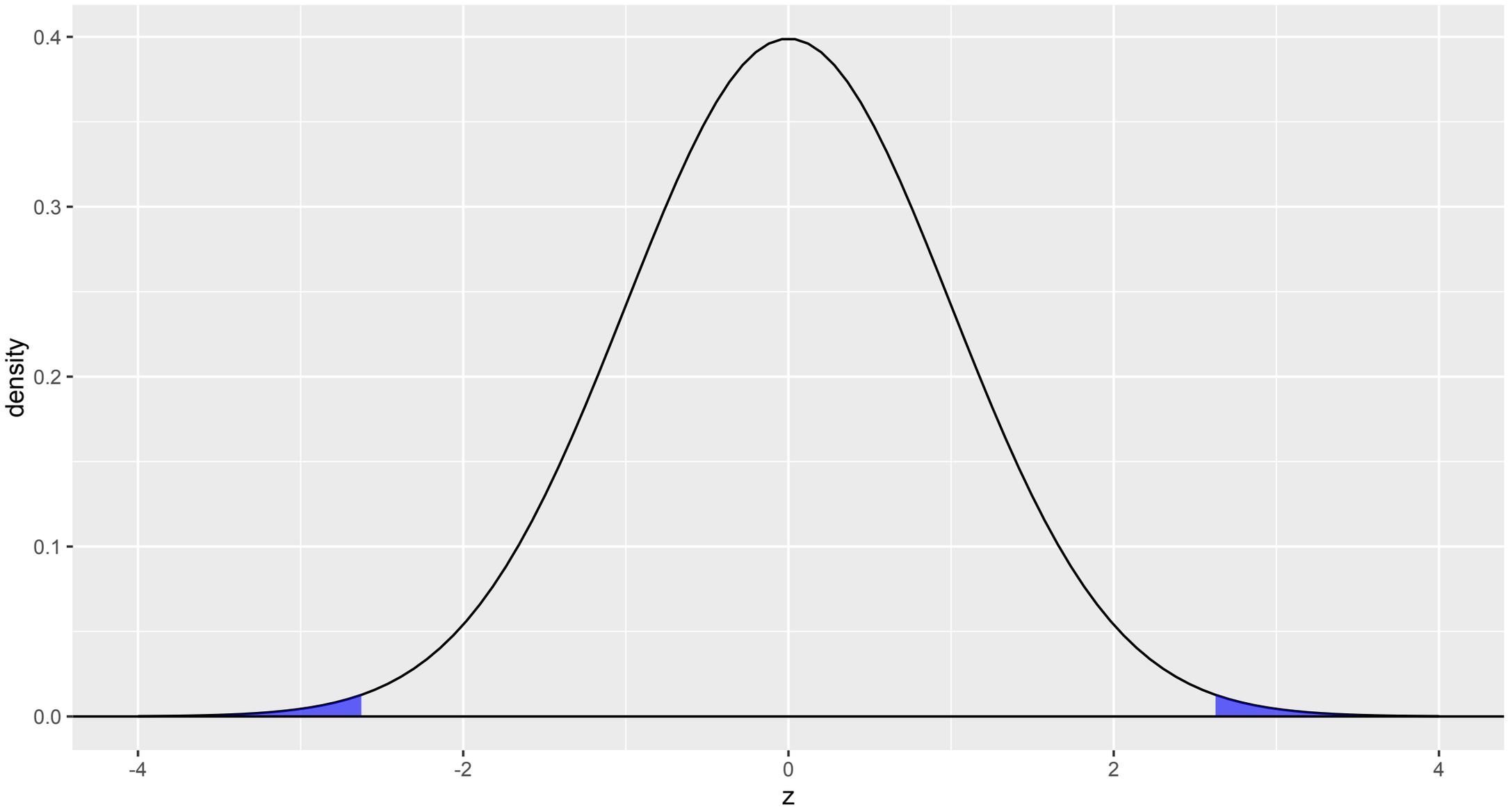
least likely 10%



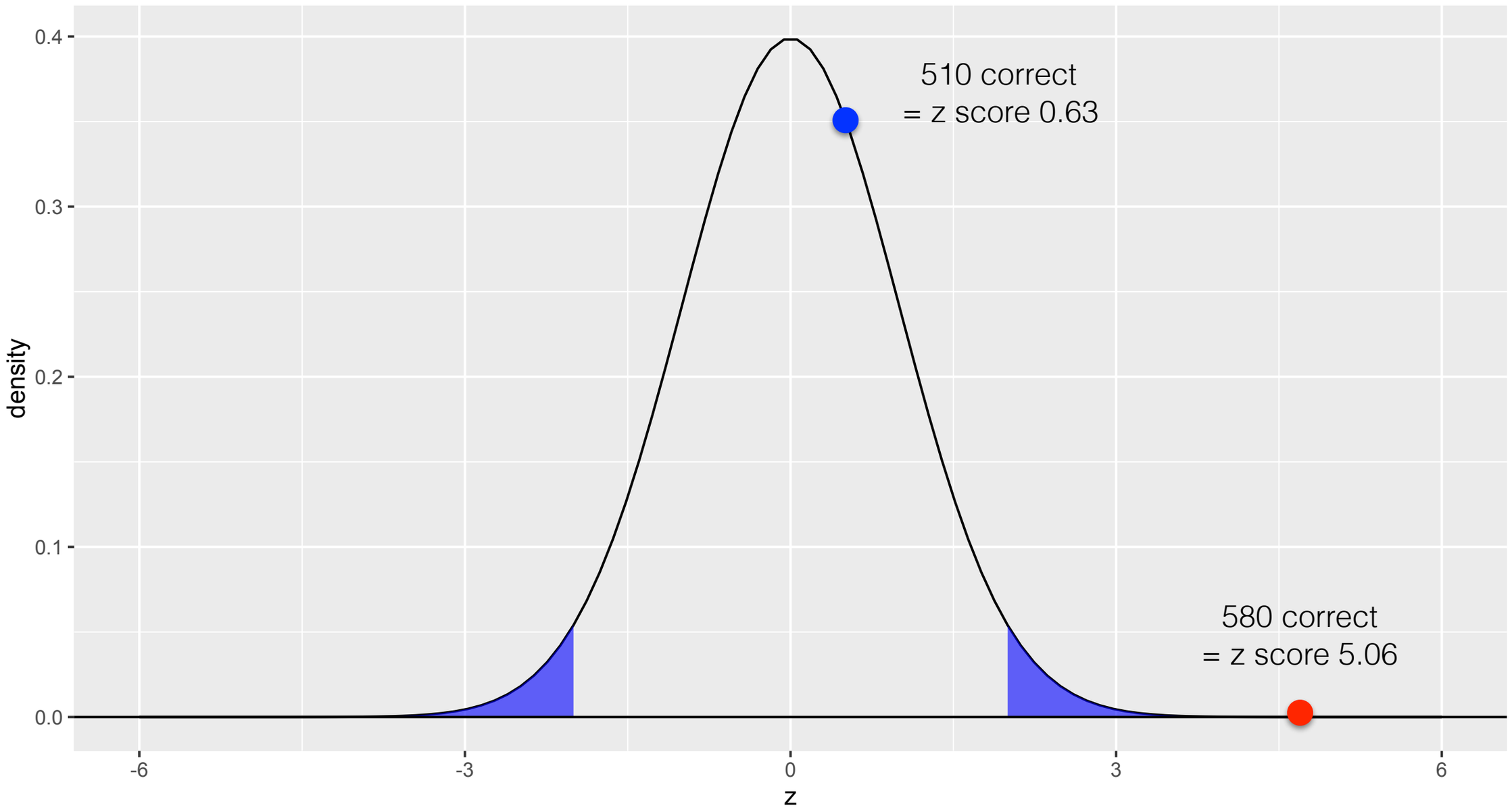
least likely 5%



least likely 1%



Tests


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

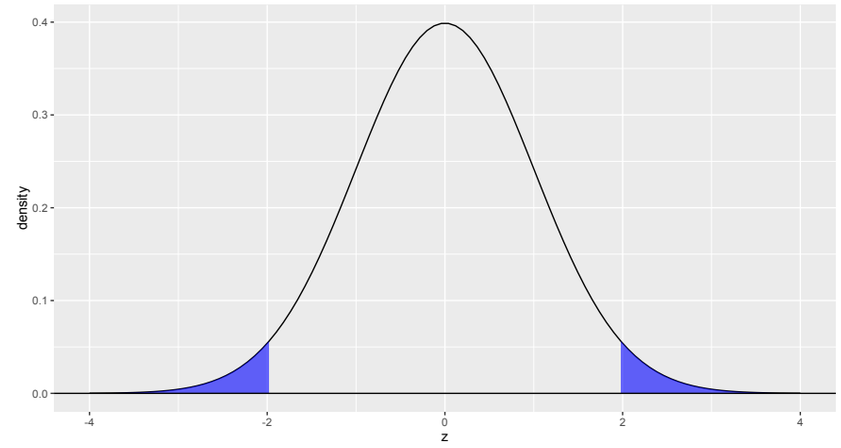
Tests

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

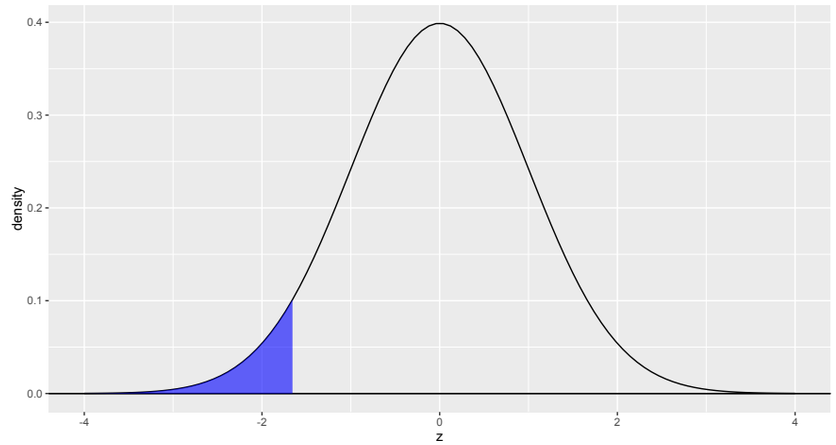
- Testing is evaluating whether the sample statistic falls in the rejection region defined by α

Tails

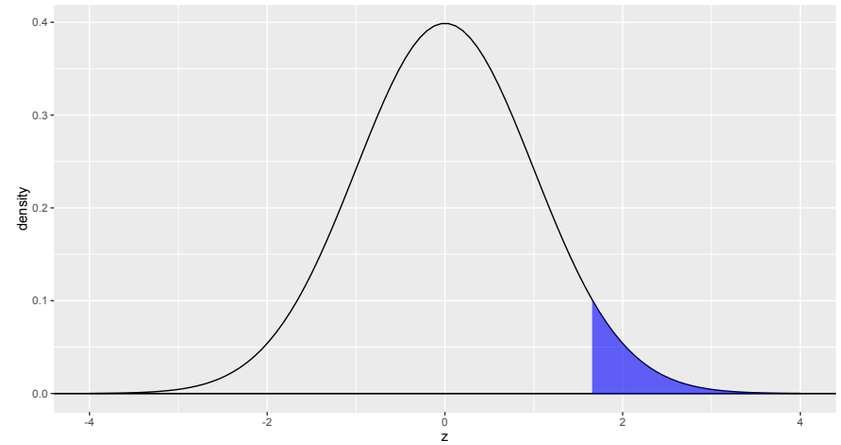
- Two-tailed tests measured whether the observed statistic is different (in either direction)
- One-tailed tests measure difference in a specific direction
- All differ in where the rejection region is located; $\alpha = 0.05$ for all.



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	“jobs” is predictive of presidential approval rating
2	“job” is predictive of presidential approval rating
3	“war” is predictive of presidential approval rating
4	“car” is predictive of presidential approval rating
5	“the” is predictive of presidential approval rating
6	“star” is predictive of presidential approval rating
7	“book” is predictive of presidential approval rating
8	“still” is predictive of presidential approval rating
9	“glass” is predictive of presidential approval rating
…	…
1,000	“bottle” is predictive of presidential approval rating

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

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

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

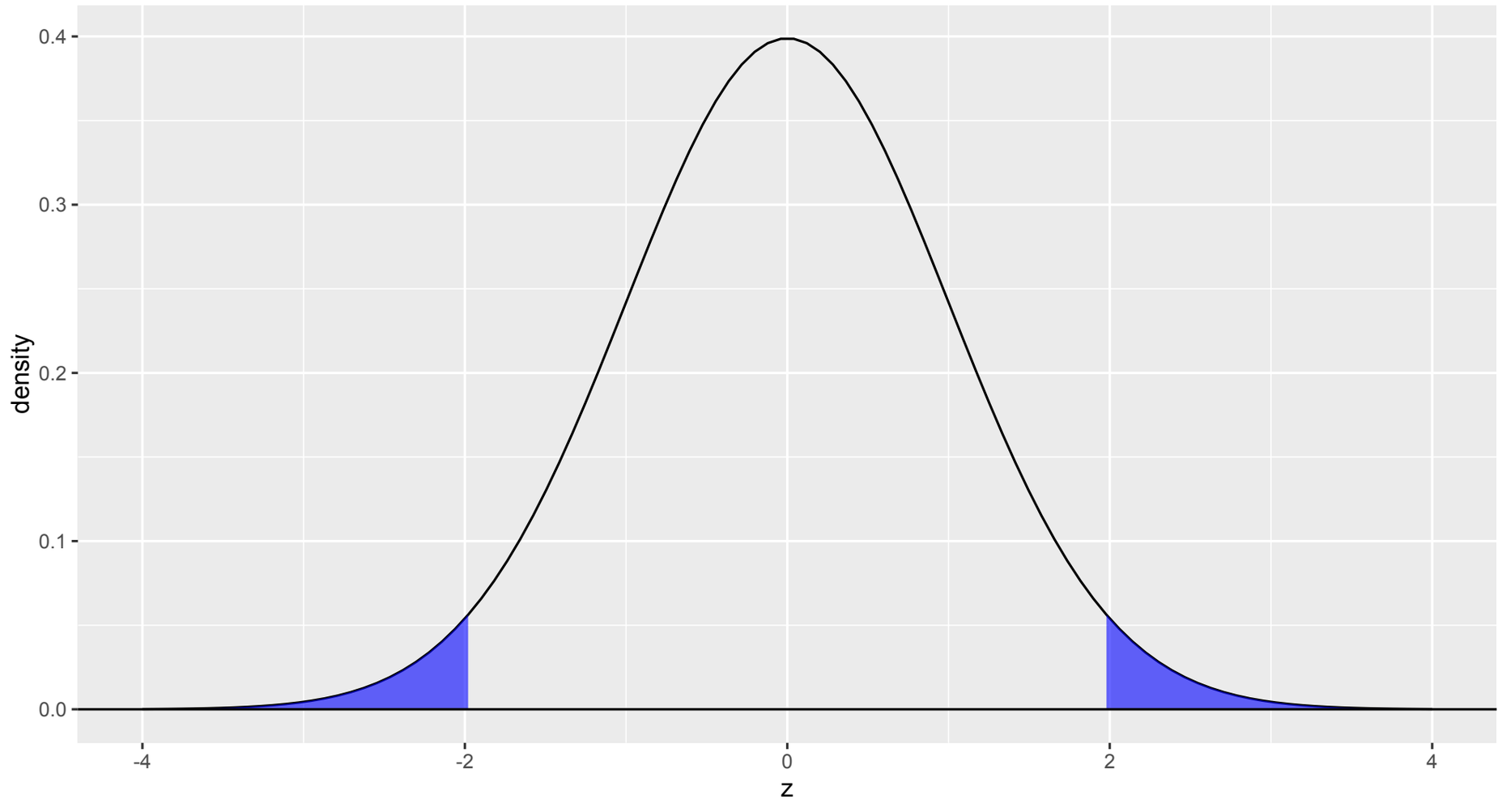
Significance of coefficients

- 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?

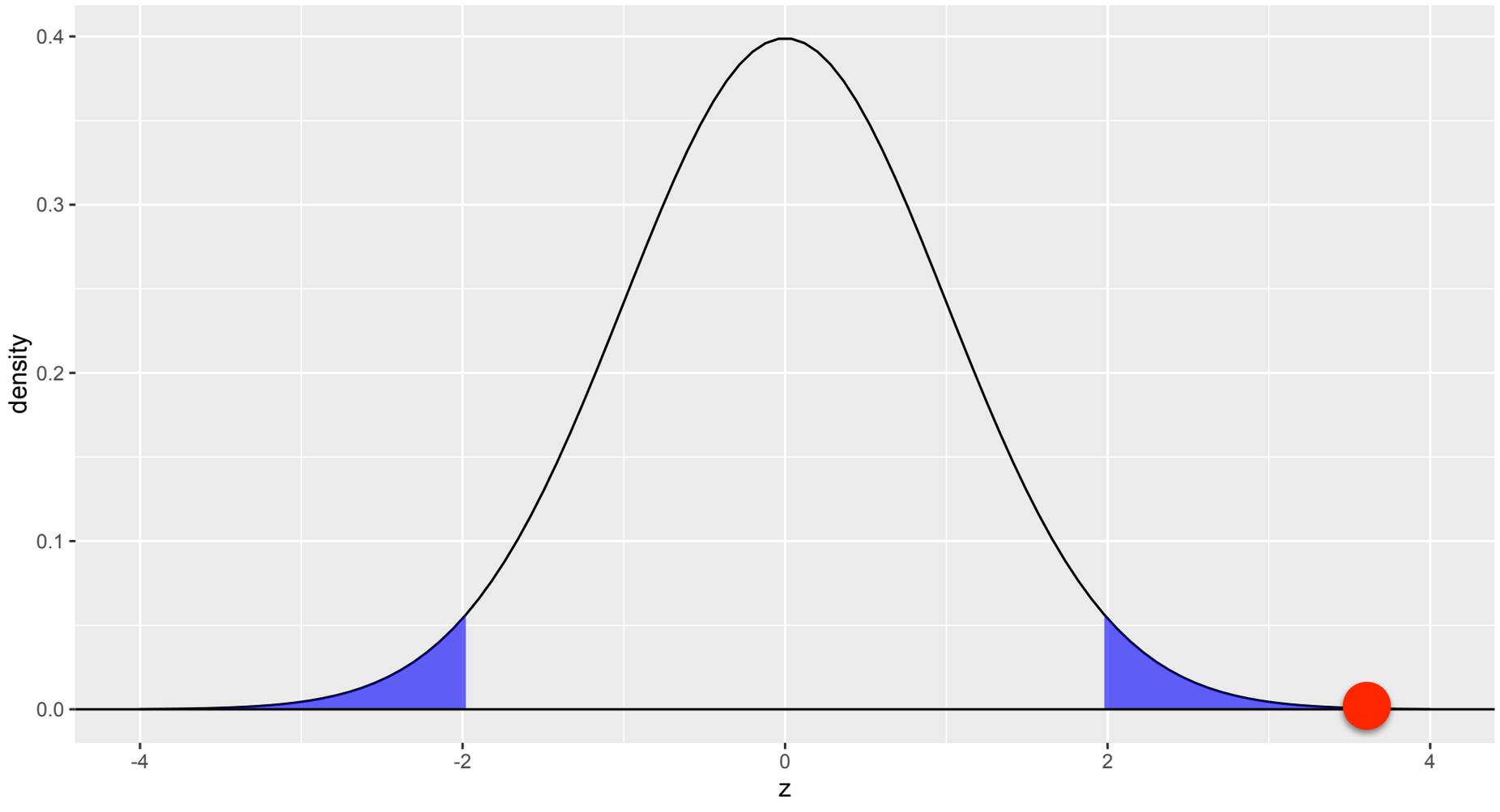
- People often use parametric tests (coefficients are drawn from a normal distribution) to assess this for logistic regression, but we can use it to illustrate another more robust test.

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)

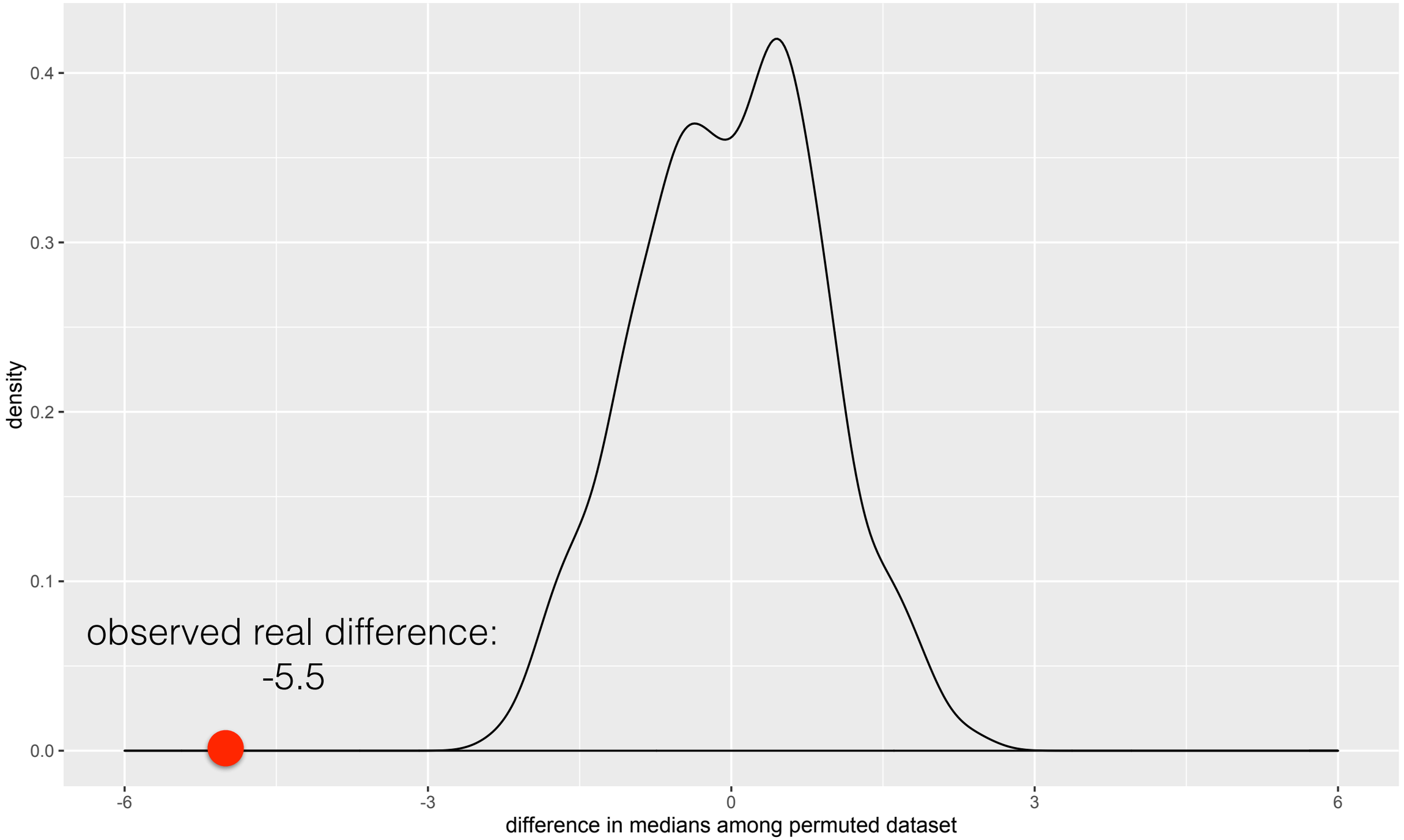
- We shuffle the labels of the data under the null assumption that the labels don't matter (the null is that A = B)

		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:

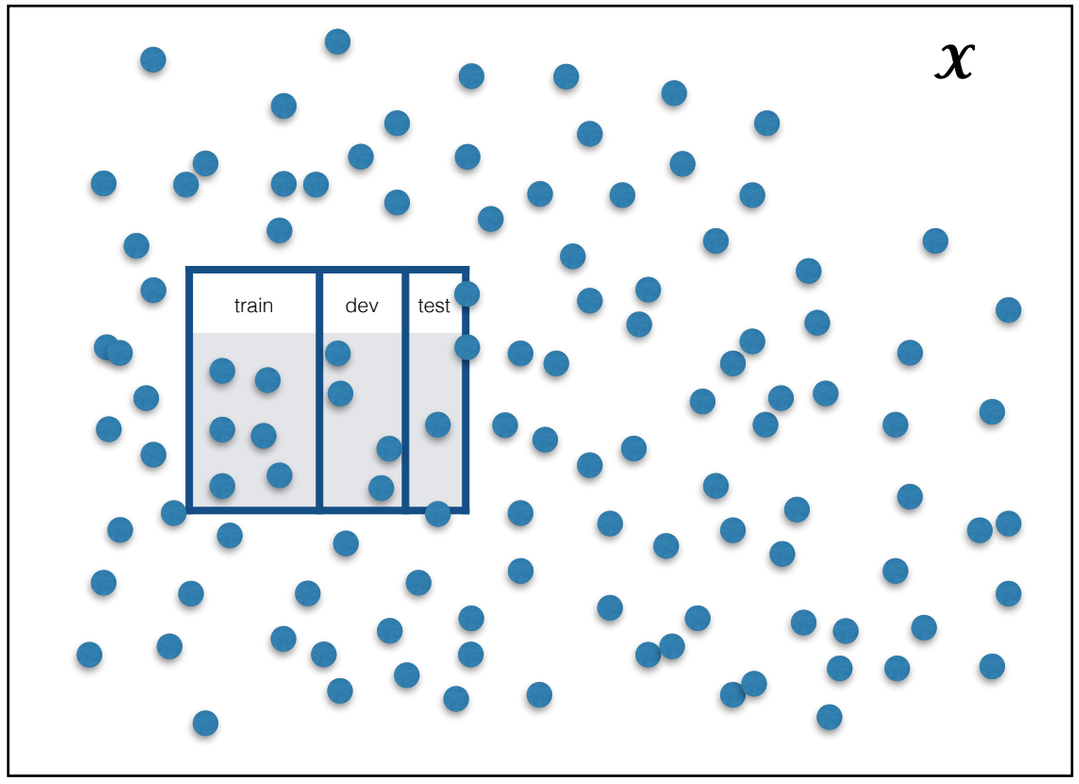
Bootstrap

- 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 from all data that could exist; let's sample from our sample to estimate the variability.



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)

Bootstrap

- At the end of the process, you end up with a vector of values m = [m(1), …, m(b)] (for b bootstrap samples)

m	Utility
POS accuracy for System A	Estimate confidence intervals
I[System A > System B]	Calculate significance of difference

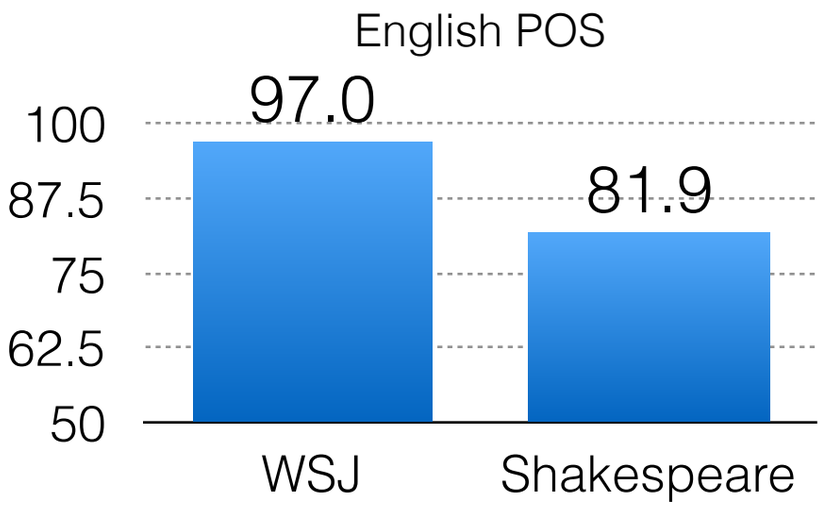
Issues

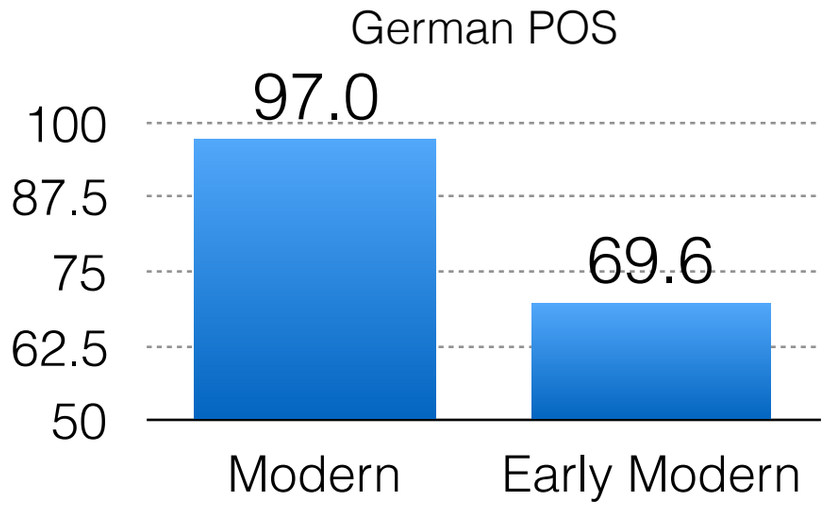
- Evaluation performance may not hold across domains (e.g., WSJ →literary texts)

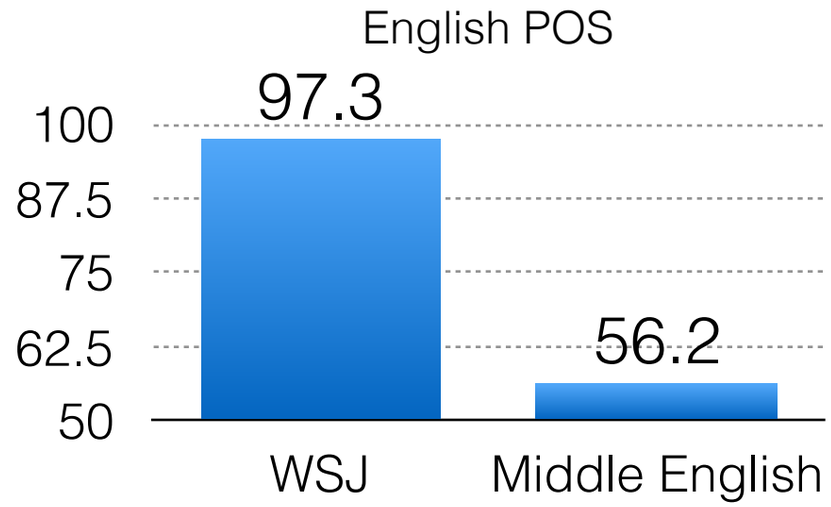
- Covariates may explain performance (MT/parsing, sentences up to length n)

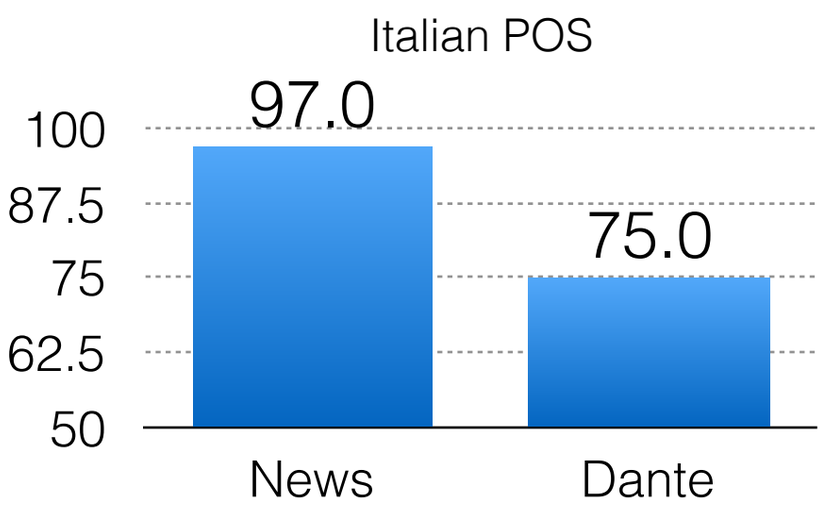
- Multiple metrics may offer competing results

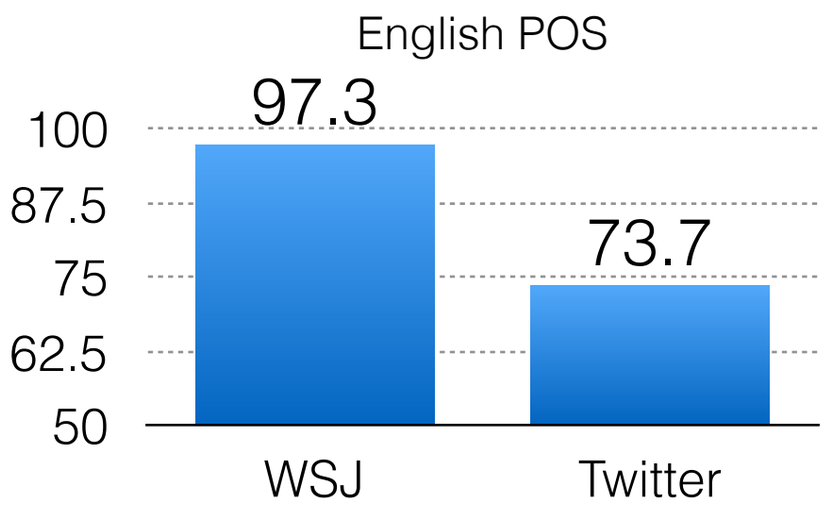
Søgaard et al. 2014

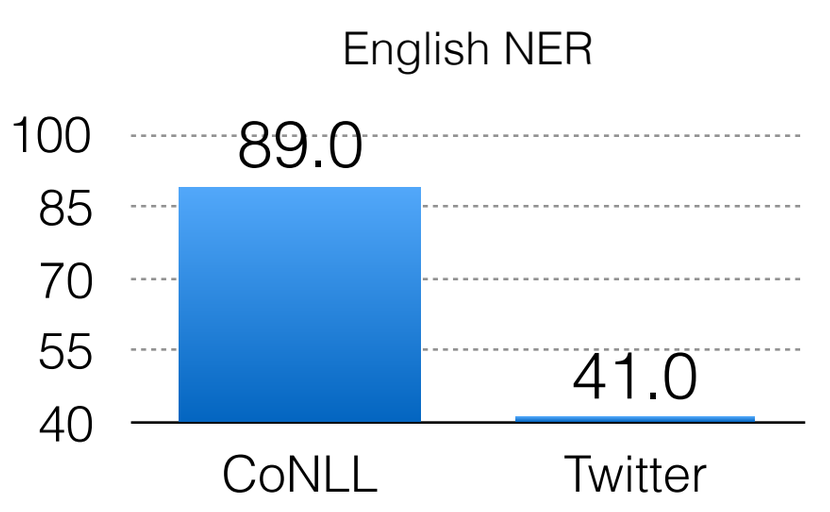

English POS
97.0
81.9
100
87.5
75
62.5
50
WSJ
Shakespeare

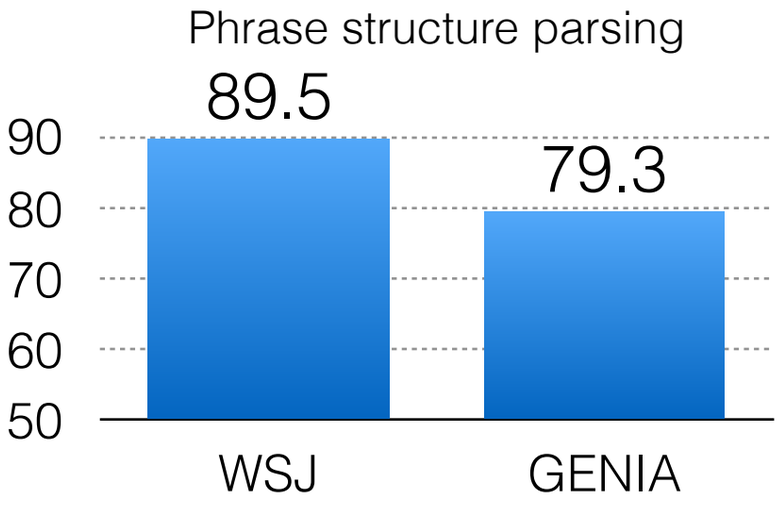

German POS
97.0
69.6
100
87.5
75
62.5
50
Modern
Early Modern

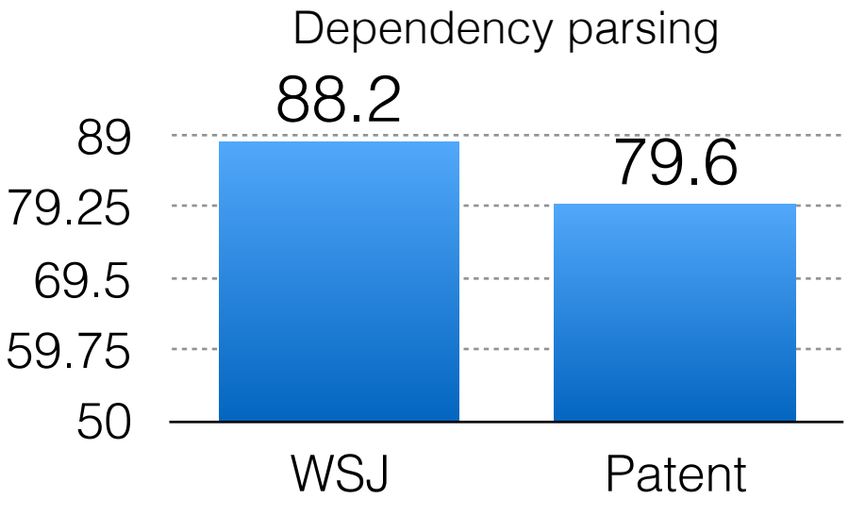

English POS
97.3
56.2
100
87.5
75
62.5
50
WSJ
Middle English

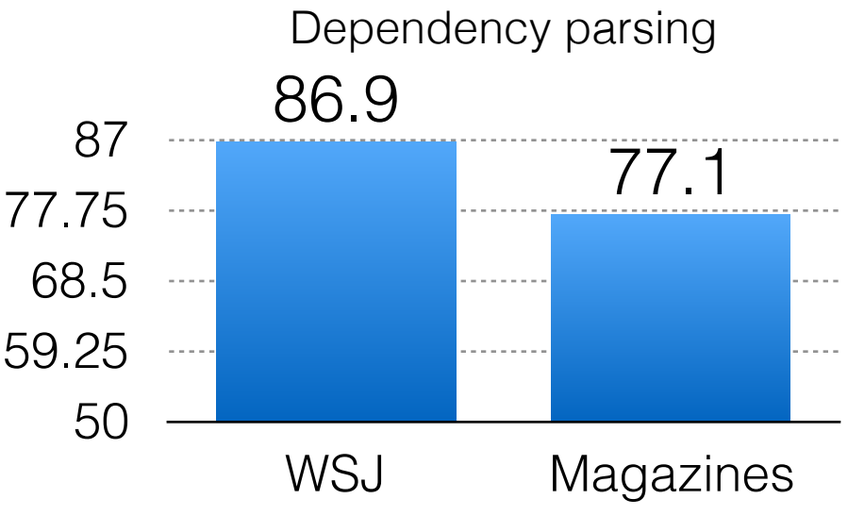

Italian POS
97.0
75.0
100
87.5
75
62.5
50
News
Dante


English POS
97.3
73.7
100
87.5
75
62.5
50
WSJ
Twitter


English NER
89.0
41.0
100
85
70
55
40
CoNLL
Twitter


Phrase structure parsing
89.5
79.3
90
80
70
60
50
WSJ
GENIA


Dependency parsing
88.2
79.6
89
79.25
69.5
59.75
50
WSJ
Patent


Dependency parsing
86.9
77.1
87
77.75
68.5
59.25
50
WSJ
Magazines

Takeaways

- At a minimum, always evaluate a method on the domain you're using it on

- When comparing the performance of models, quantify your uncertainty with significant tests/ confidence bounds

- Use ablation tests to identify the impact that a feature class has on performance.