Applied Natural Language Processing

Info 256

Lecture 9: Testing (Sept. 25, 2023)

David Bamman, UC Berkeley

Significance in NLP

- You develop a new method for text classification; 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

Metrics

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

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

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

Multiclass confusion matrix

Predicted (ŷ)

True (y)

	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	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)

	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	30	40	70

Precision

Precision(Dem) =

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

Predicted (ŷ)

True (y)

	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	30	40	70

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

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

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?

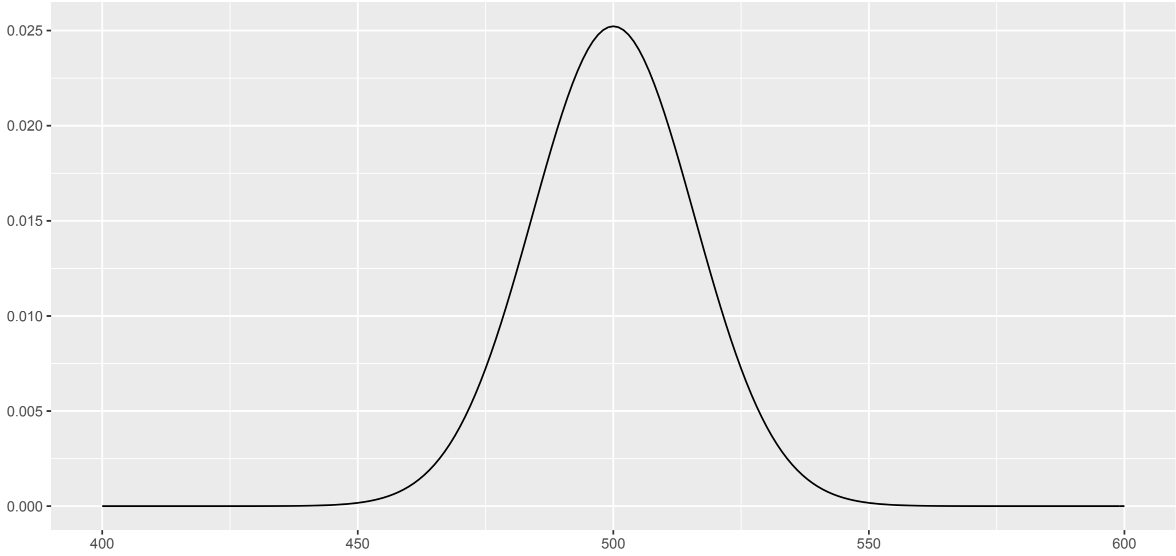
Hypothesis testing

- 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)

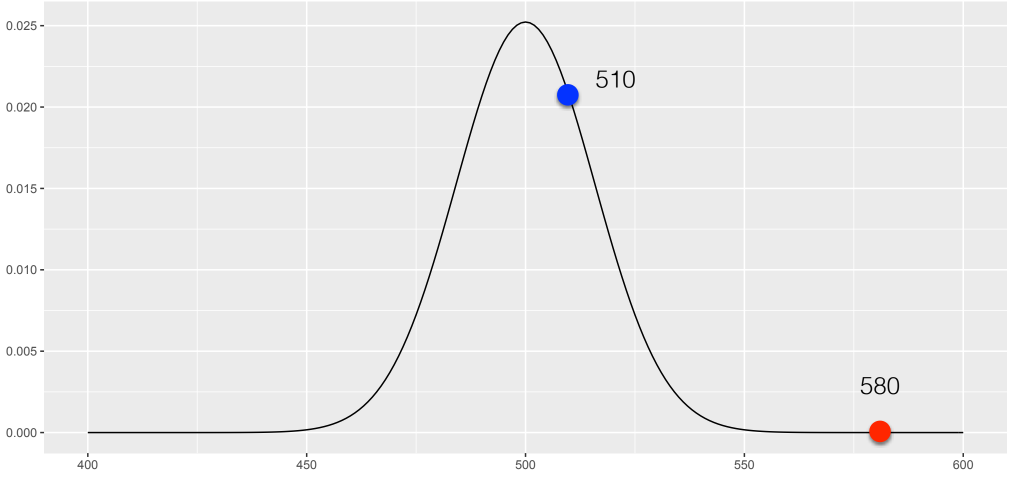
Example



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.

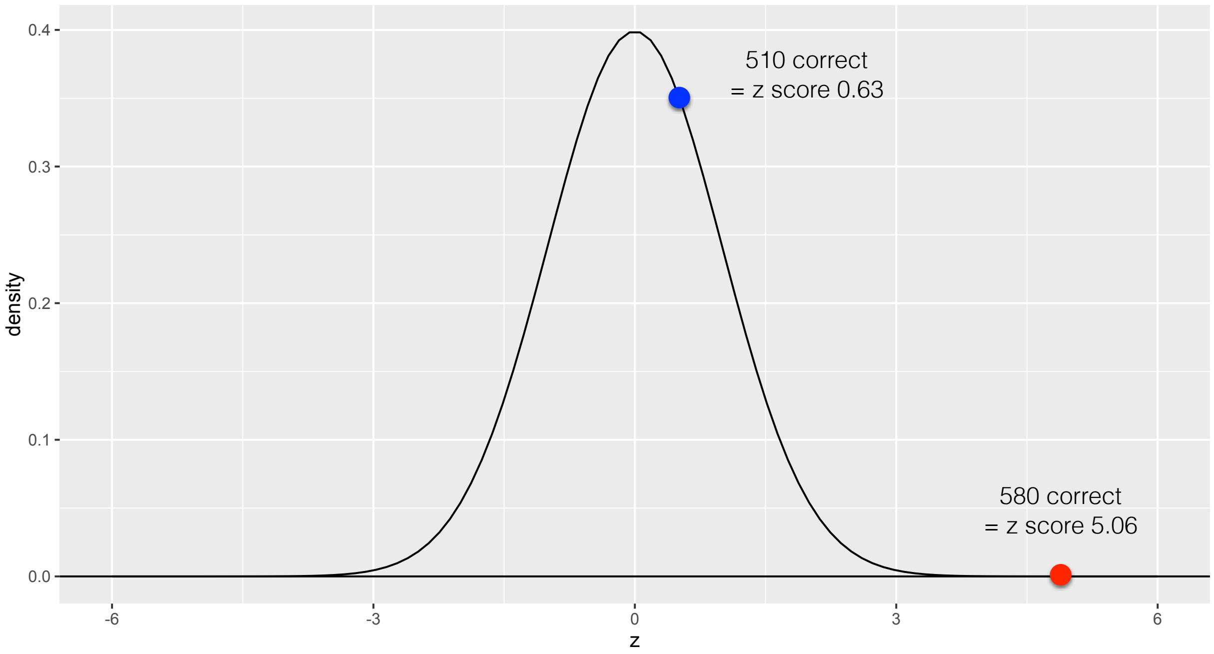
- We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size n; for large n, we can often make a normal approximation.

Z score

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

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

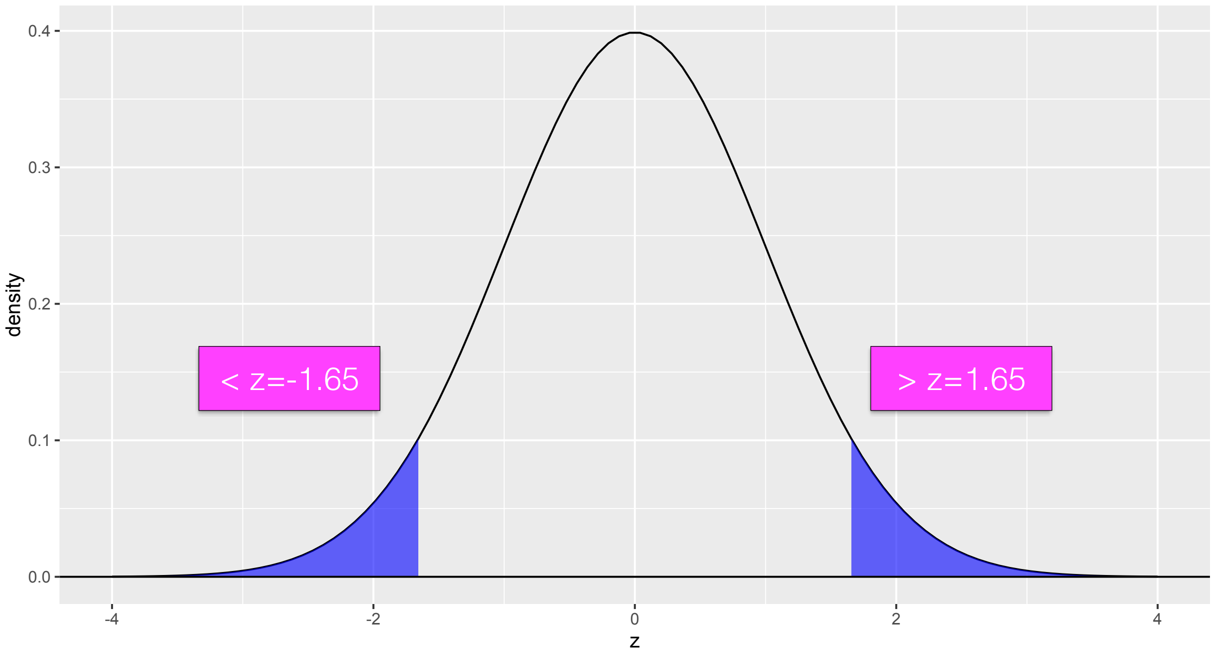
Z score


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

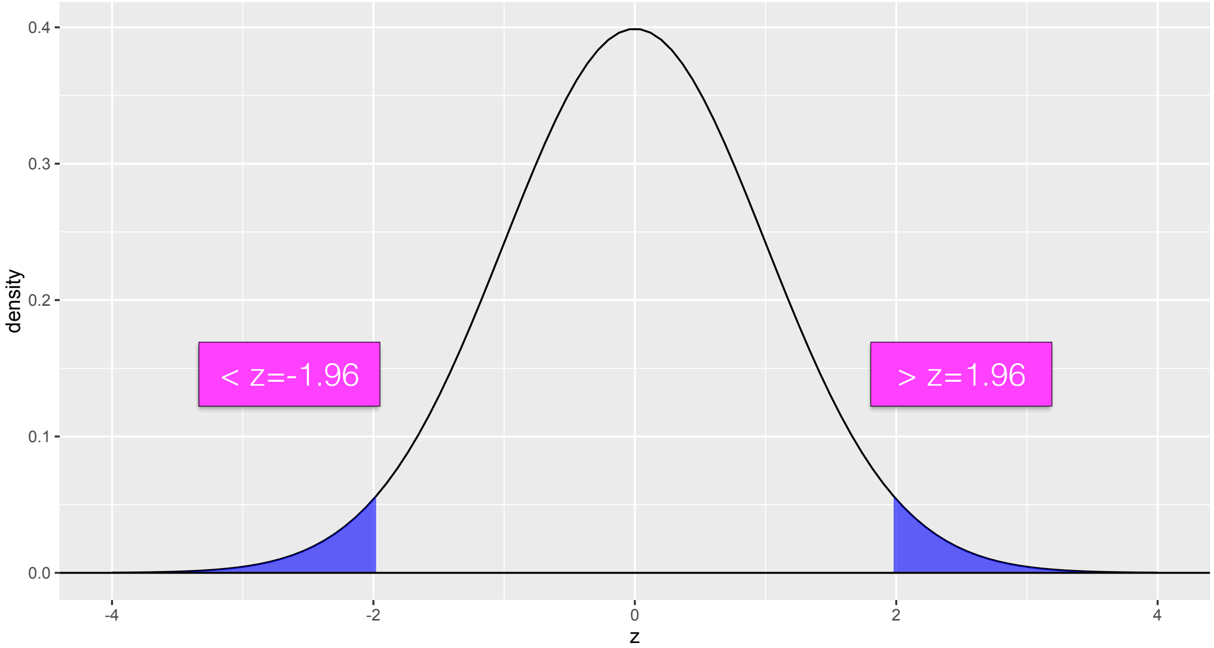
Tests

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

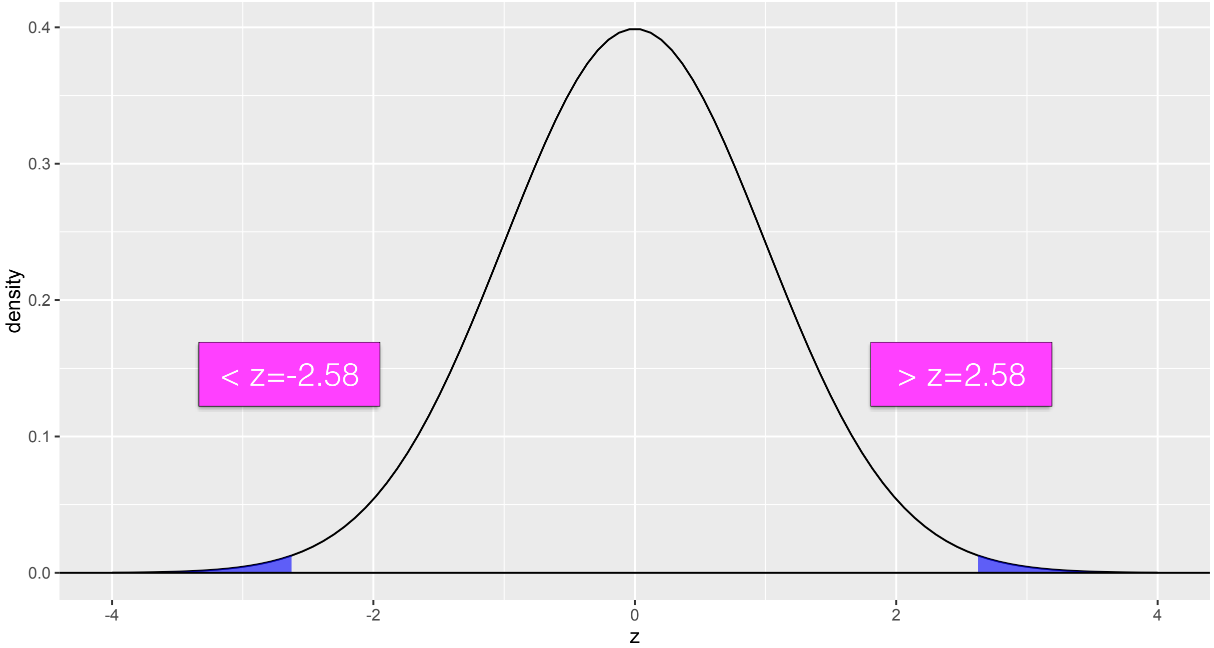
least likely 10%



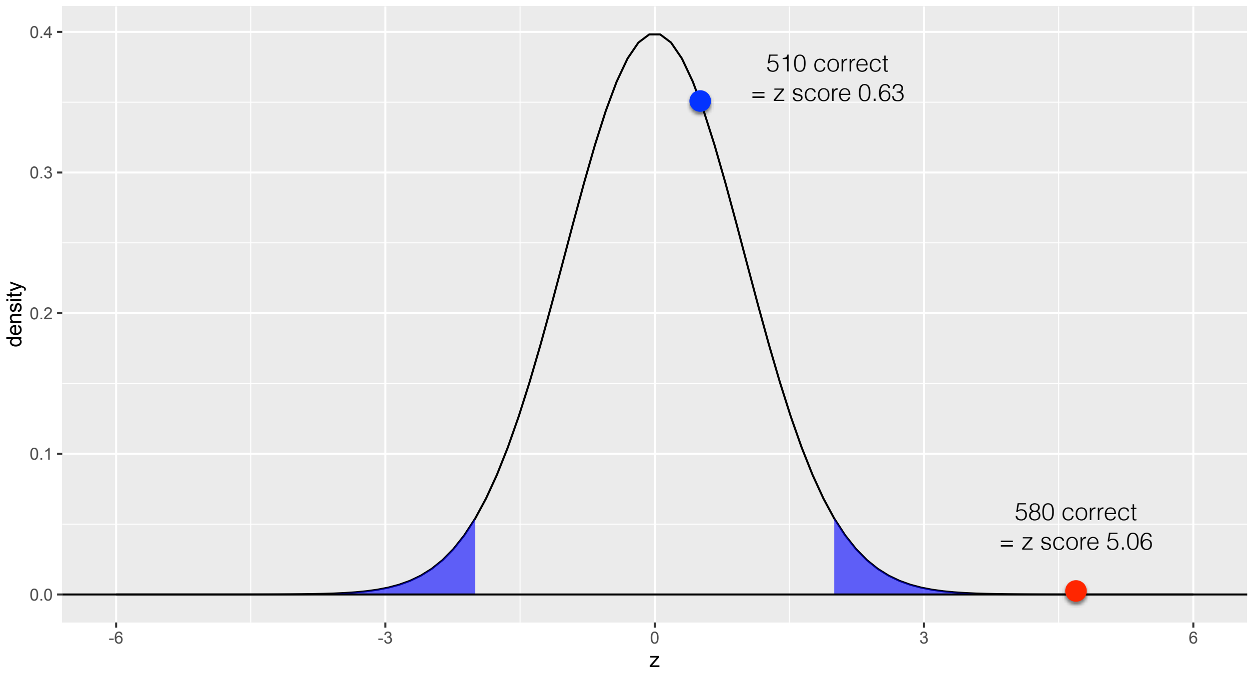
least likely 5%



least likely 1%



Tests



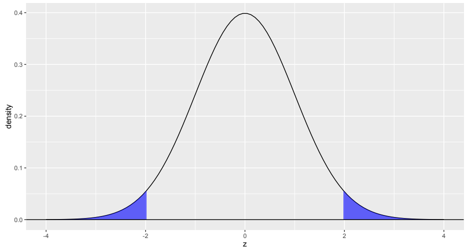
510 correct
= z score 0.63
580 correct
= z score 5.06
0.4
0.3
0.2
0.1
0.0
density
-6
-3
0
3
6
z

Tests

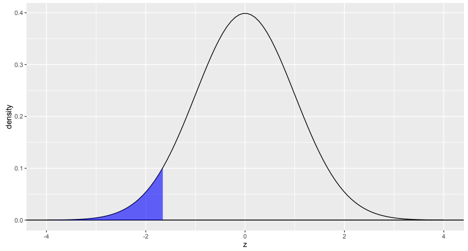
- Decide on the level of significance α. {0.05, 0.01}
- Testing is evaluating whether the sample statistic falls in the rejection region defined by α

Tails

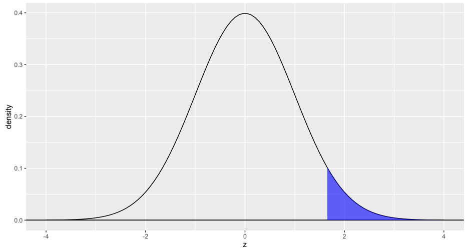
- Two-tailed tests measured whether the observed statistic is different (in either direction)
- One-tailed tests measure difference in a specific direction
- All differ in where the rejection region is located; α = 0.05 for all.



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

- 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.

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

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 $\alpha \times n$ type I errors.
- $\alpha=0.01$, $n=1000$ = 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

Confidence intervals

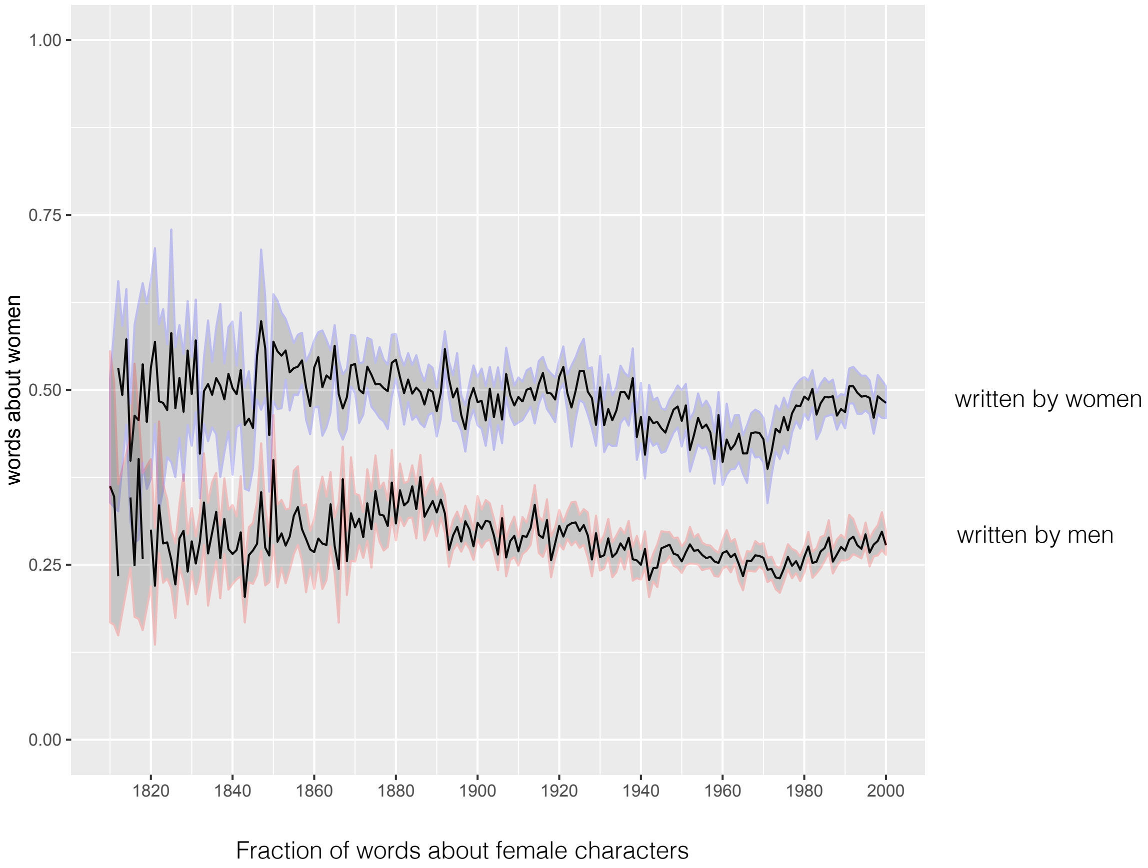
- Even in the absence of specific test, we want to quantify our uncertainty about any metric.
- Confidence intervals specify a range that is likely to contain the (unobserved) population value from a measurement in a sample.

Confidence intervals

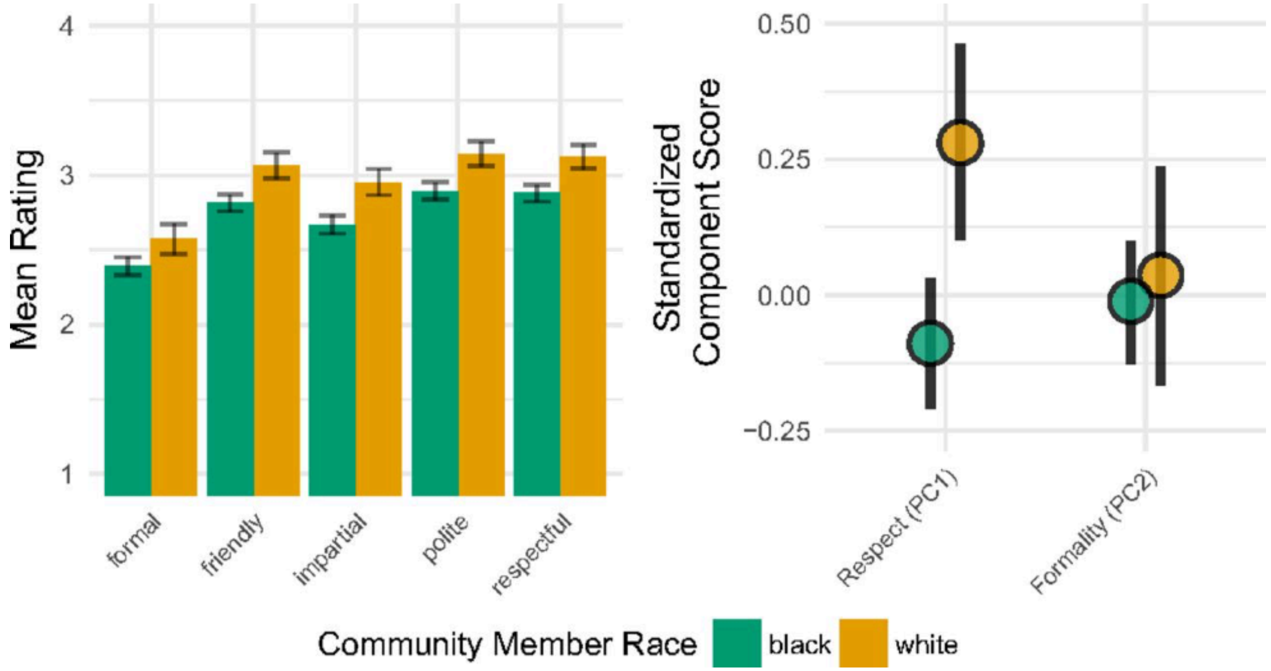
Binomial confidence intervals (again using Normal approximation):

- p = rate of success (e.g., for binary classification, the accuracy).
- n = the sample size (e.g., number of data points in test set).
- z_α = the critical value at significance level α.
    - 95% confidence interval: $\alpha = 0.05$; $z_\alpha = 1.96$
    - 99% confidence interval: $\alpha = 0.01$; $z_\alpha = 2.58$

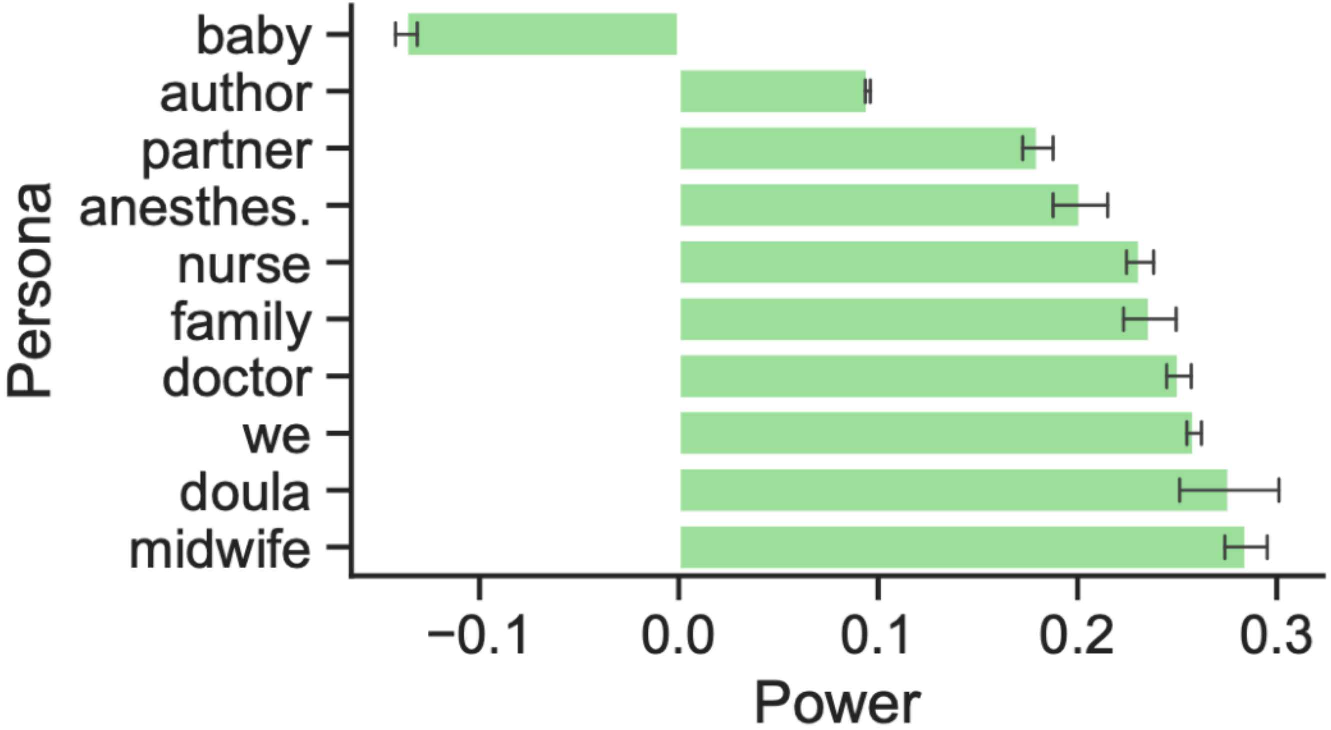
$$p \pm z_\alpha \sqrt{\frac{p(1-p)}{n}}$$



Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," (*Cultural Analytics*)



Voigt et al. 2017, “Language from police body camera footage shows racial disparities in officer respect”

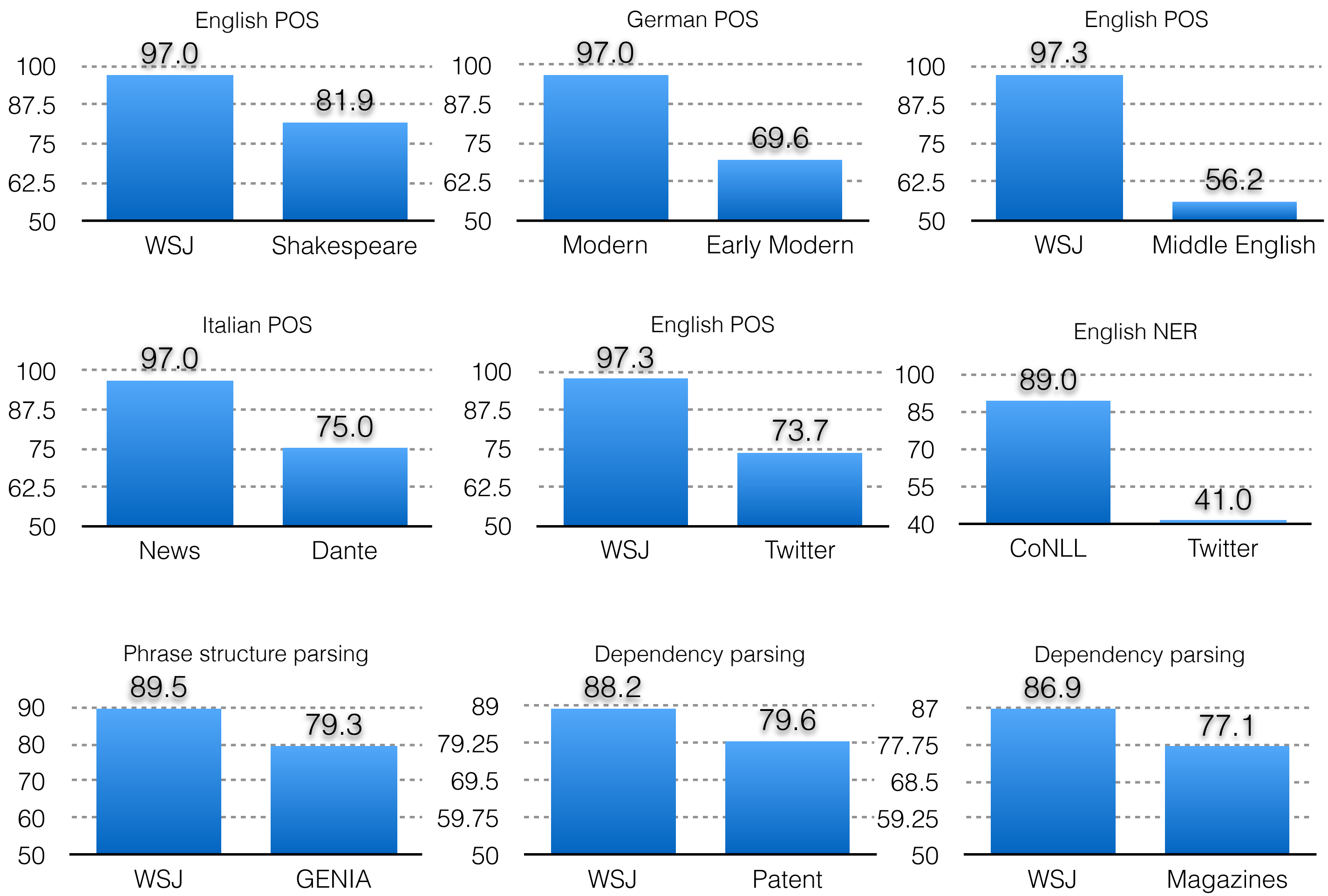


Antoniak et al. 2019, "Narrative Paths and Negotiation of Power in Birth Stories"

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.

Takeaways

- Whenever you calculate a metric from some data, report confidence intervals around it!
- → Accuracy and other measures of validity; any inferred statistics from a sample you're using to make arguments about — anything where there may be variability in that measures as a result of the sample of data you have.

Activity

`7.tests/ParametricTest`

- Explore a simple hypothesis test checking whether the accuracy of a trained model for classification in your last homework is meaningfully different from a majority class baseline