How to test a hypothesis?

Sandeep Soni

09/17/2024

Question for the day

"How to statistically test claims using text?"

Hypotheses

Hypotheses

- Many claims can be framed as hypotheses

Hypotheses

- Many claims can be framed as hypotheses

Examples
Gender bias is decreasing in books
Reframing a tweet can increase its retweet rate
Institutional mistrust is predictive of hate speech
Classifier A is better than classifier B

Null hypothesis

Null hypothesis

- Null hypothesis is a claim that is assumed to be true

Null hypothesis

- Null hypothesis is a claim that is assumed to be true

Hypothesis (H)	H_0
Gender bias is decreasing over time in books	Gender bias remains the same over time in books
Reframing a tweet can increase its retweet rate	Reframing a tweet has no effect on retweet rate
Institutional mistrust is predictive of hate speech	Institutional mistrust is not predictive of hate speech
Classifier A is better than classifier B	Classifier A is the same as classifier B

Hypothesis testing

Hypothesis testing

- H_0 helps us get an expected result

Hypothesis testing

- H_0 helps us get an expected result
- Testing if hypothesis H holds ==> if H_0 is assumed to be true, how likely does the observed data match our expectation?

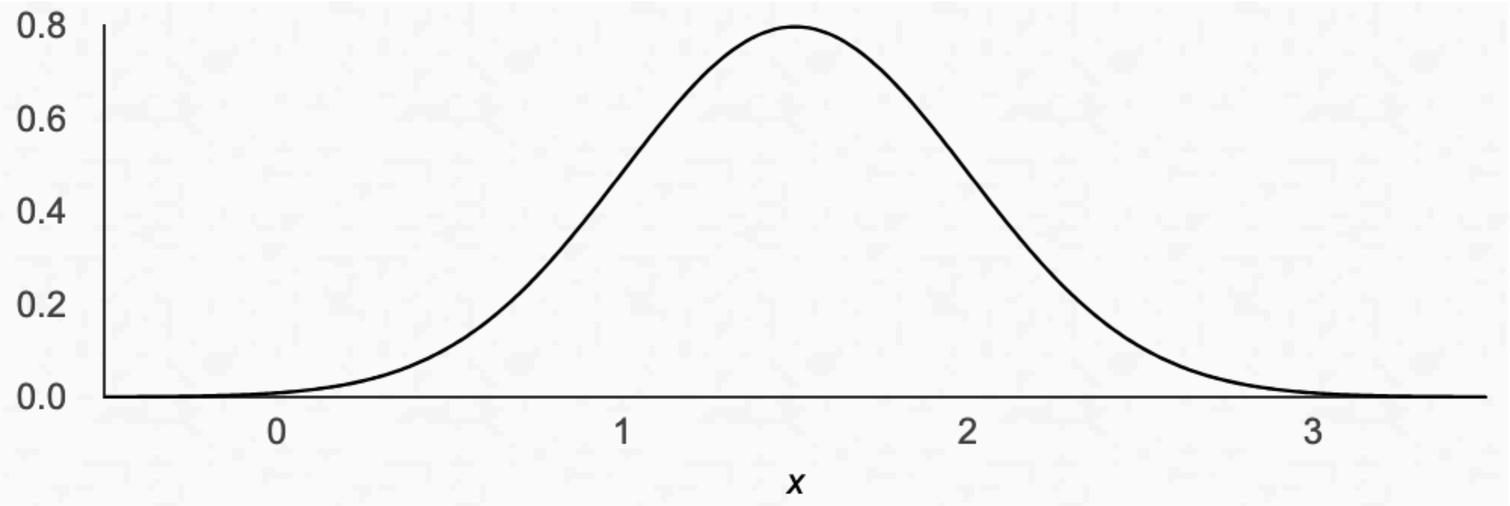
Hypothesis testing

Hypothesis testing

According to the null hypothesis, I expect the statistic to be normally distributed with some mean and variance

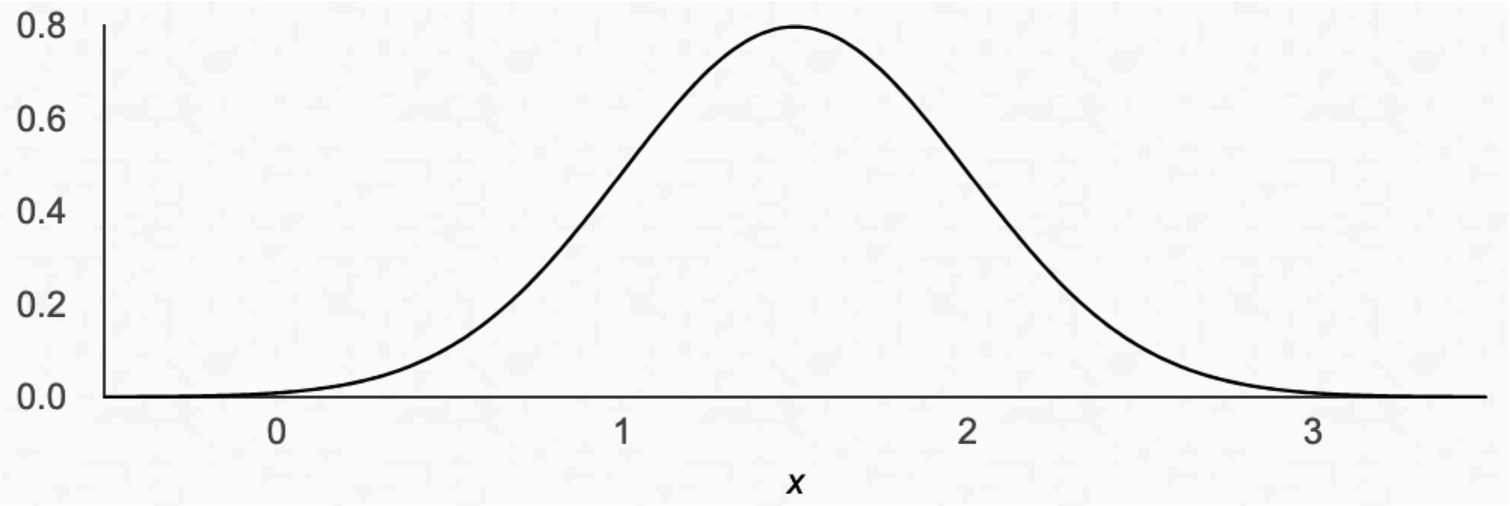
Hypothesis testing

According to the null hypothesis, I expect the statistic to be normally distributed with some mean and variance



Hypothesis testing

According to the null hypothesis, I expect the statistic to be normally distributed with some mean and variance



This sets up expectation about the shape and position of the distribution

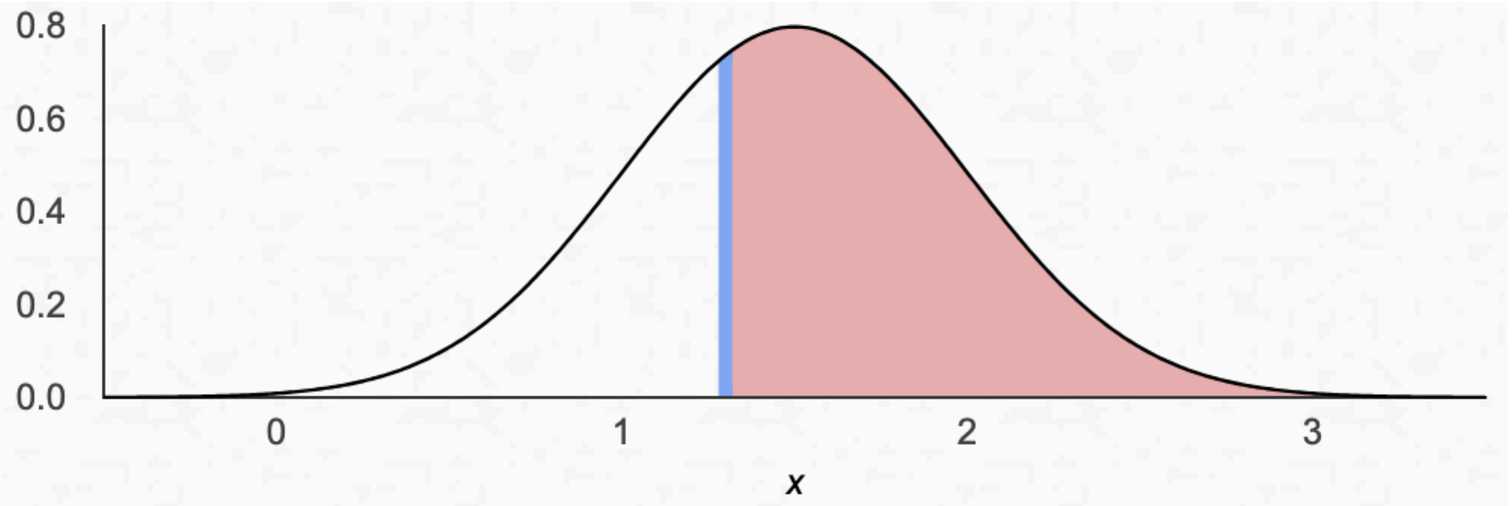
Hypothesis testing

Hypothesis testing

If we observe x to be 1.3, how likely are we to see that if the null hypothesis holds?

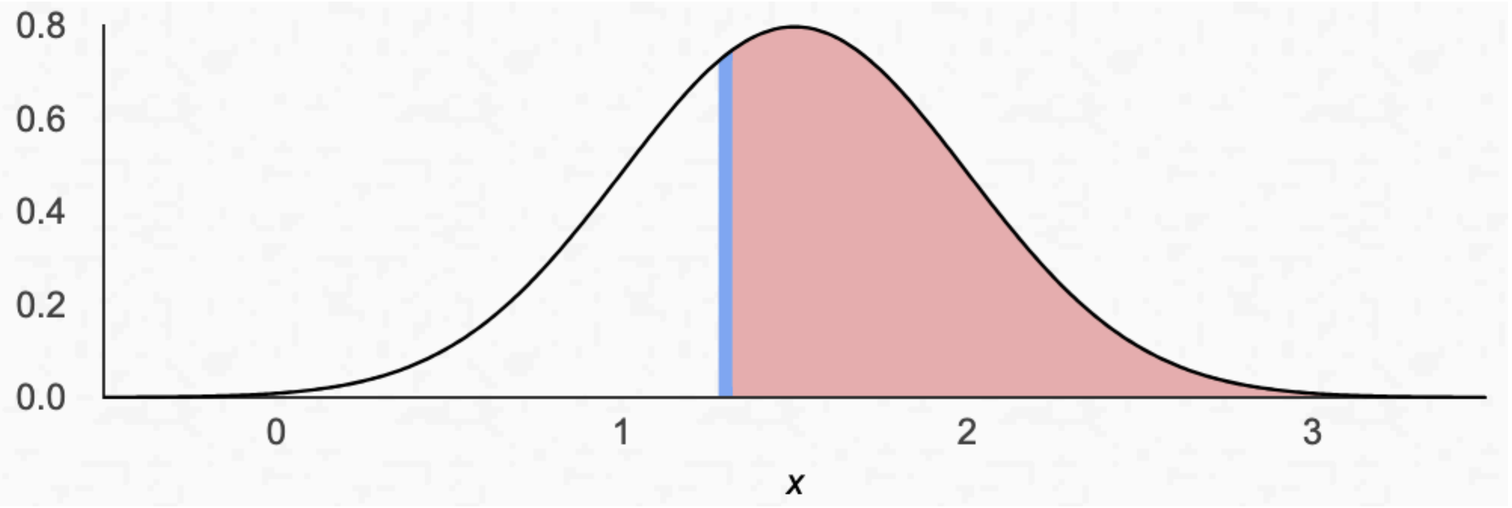
Hypothesis testing

If we observe x to be 1.3, how likely are we to see that if the null hypothesis holds?



Hypothesis testing

If we observe x to be 1.3, how likely are we to see that if the null hypothesis holds?



We can quantify this by calculating the probability of the shaded area

Hypothesis testing

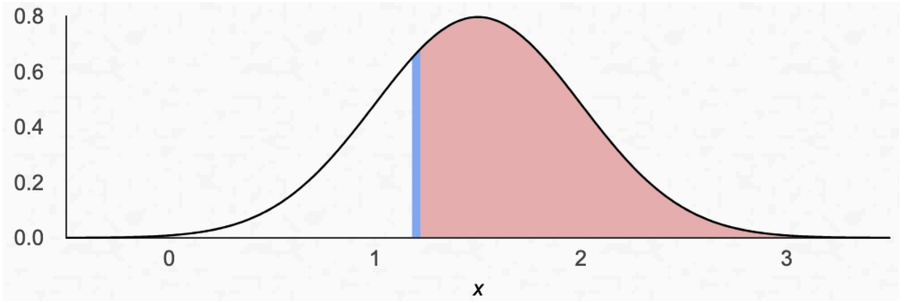
Hypothesis testing

Which observation would you say is surprising if the null distribution holds?

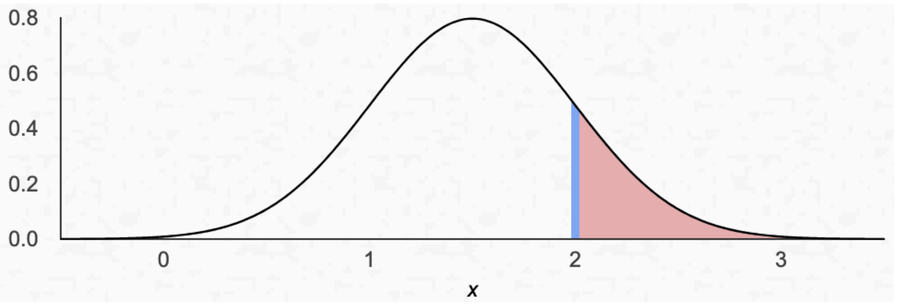
Hypothesis testing

Which observation would you say is surprising if the null distribution holds?

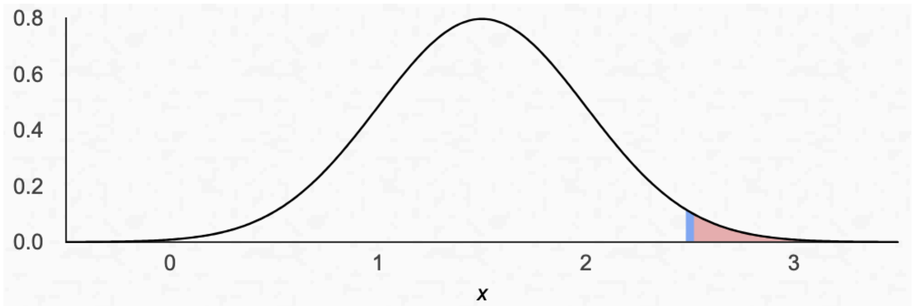
x=1.25



x=2



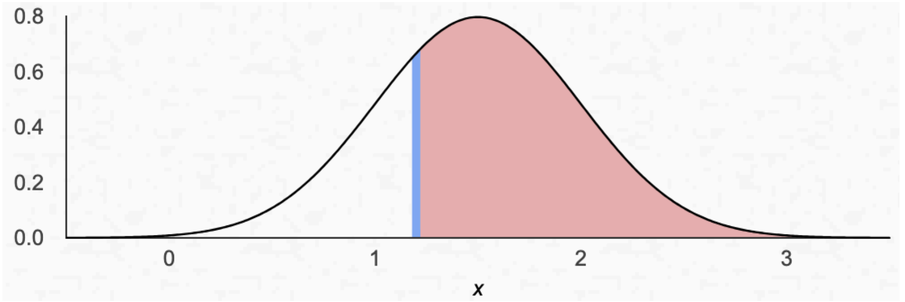
x=2.5



Hypothesis testing

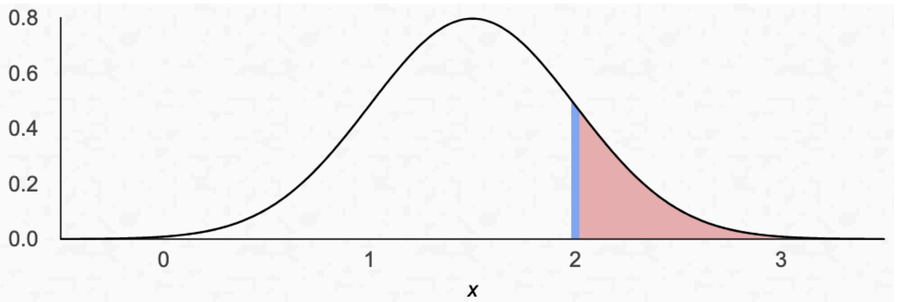
Which observation would you say is surprising if the null distribution holds?

x=1.25



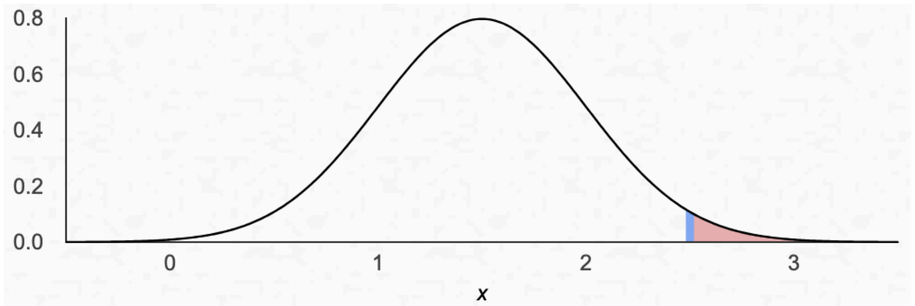
p=0.69

x=2



p=0.16

x=2.5



p=0.02

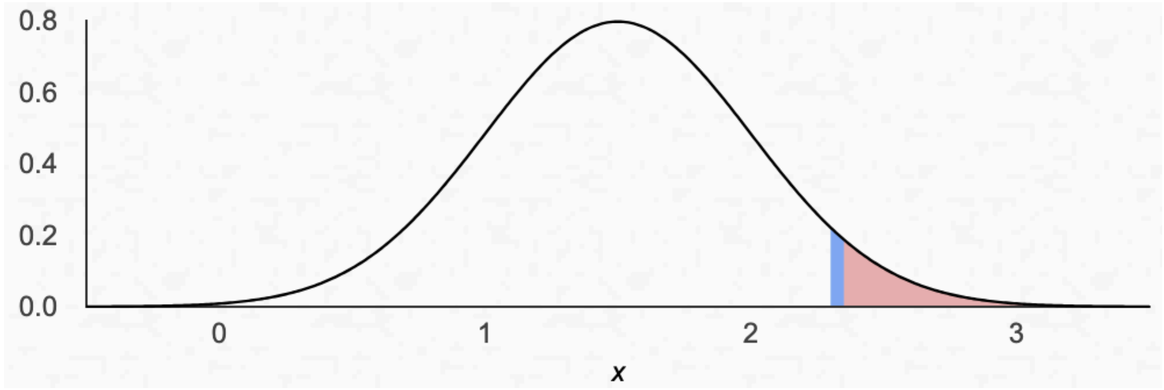
Significance level

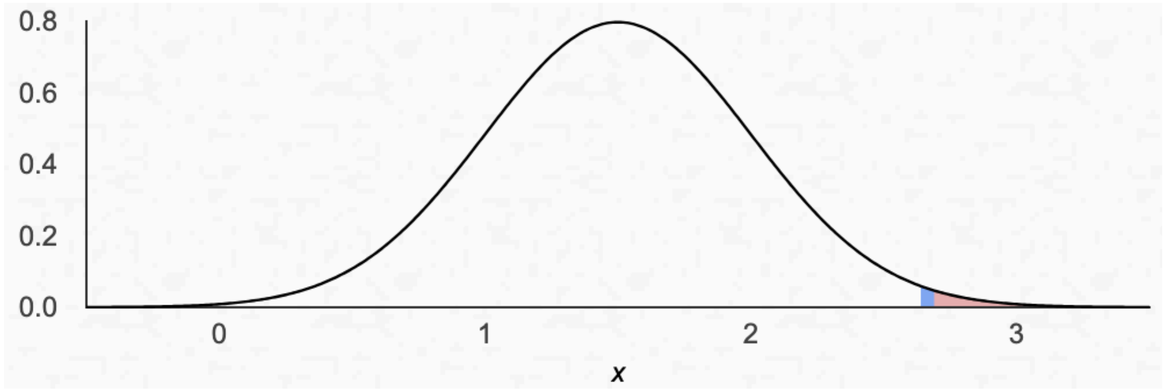
Significance level

To make a decision, set up a rejection region by choosing the value of α

Significance level

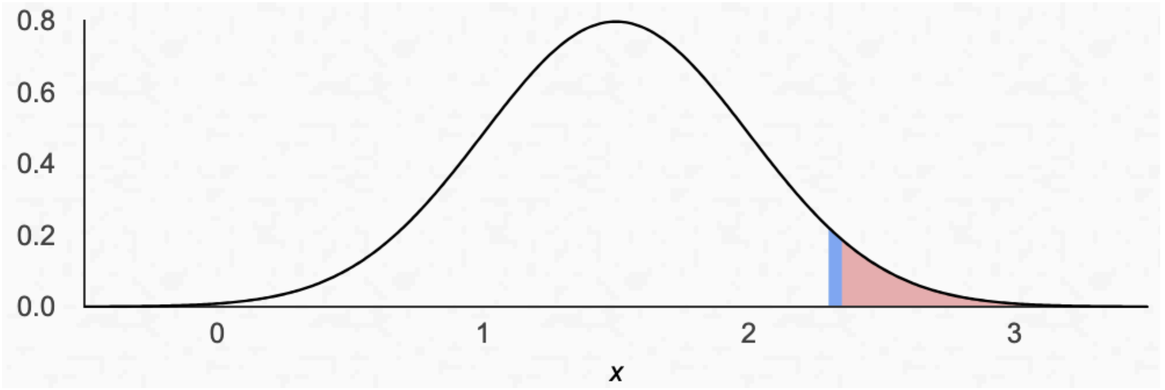
To make a decision, set up a rejection region by choosing the value of α



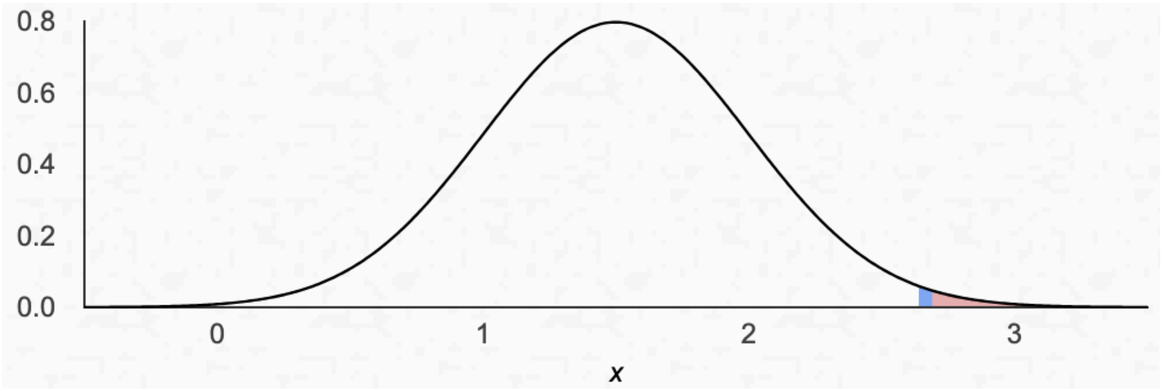


Significance level

To make a decision, set up a rejection region by choosing the value of α

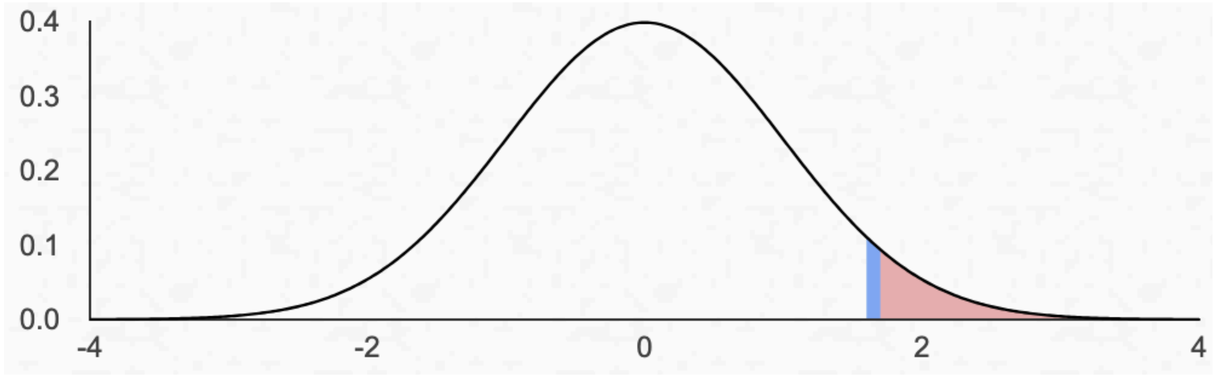


α=0.05



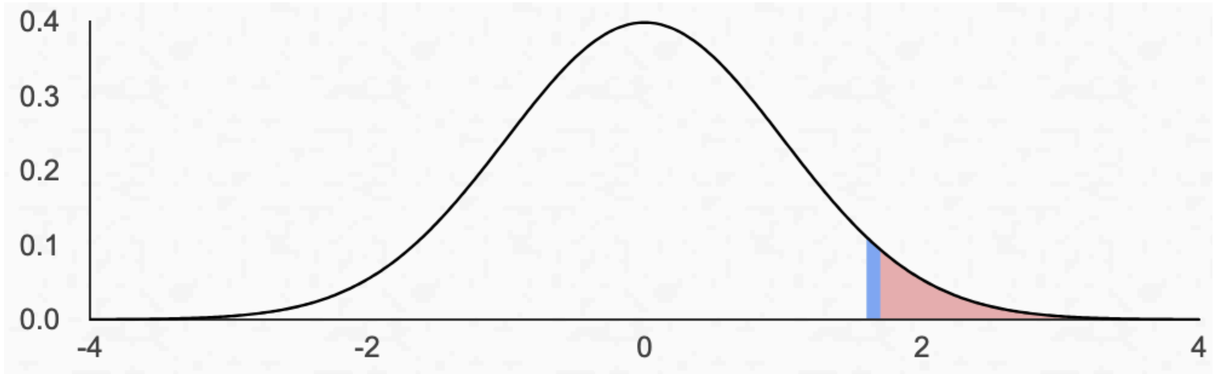
α=0.01

Z-score



Z-score

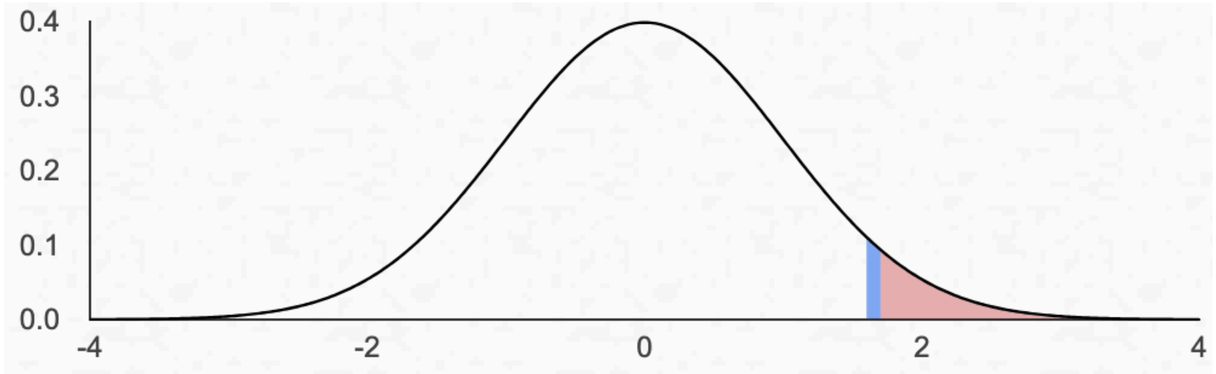
Calculating probabilities and comparing with the significance level can be tedious



Z-score

Calculating probabilities and comparing with the significance level can be tedious

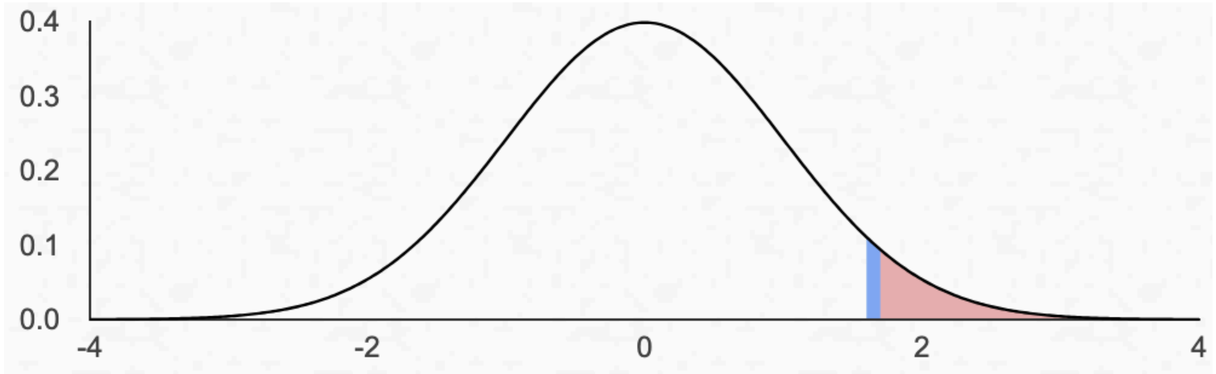
$$z = \frac{x - \mu}{\sigma / \sqrt{n}}$$



Z-score

Calculating probabilities and comparing with the significance level can be tedious

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

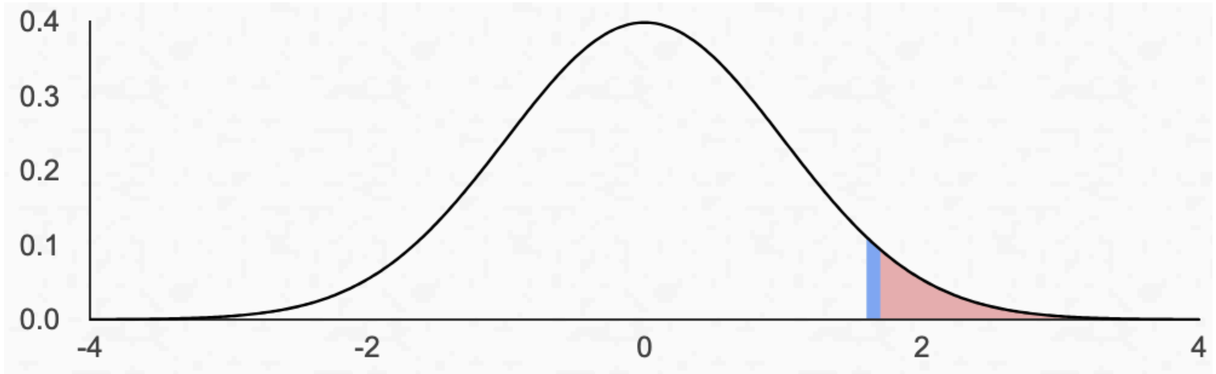


$$z = 1.65$$

Z-score

Calculating probabilities and comparing with the significance level can be tedious

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$



$z = 1.65$

We can simply calculate the z-score of the statistic and compare it to the z-score for the significance levels

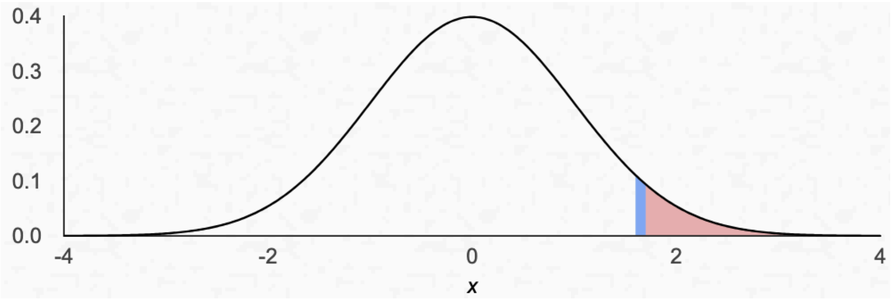
Tails

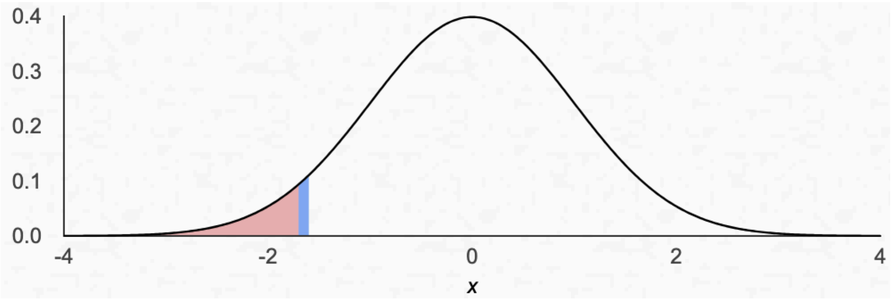
Tails

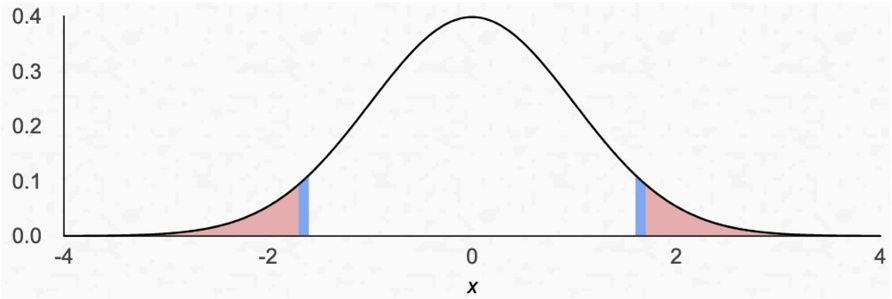
What is considered surprising is dependent on the problem at hand

Tails

What is considered surprising is dependent on the problem at hand

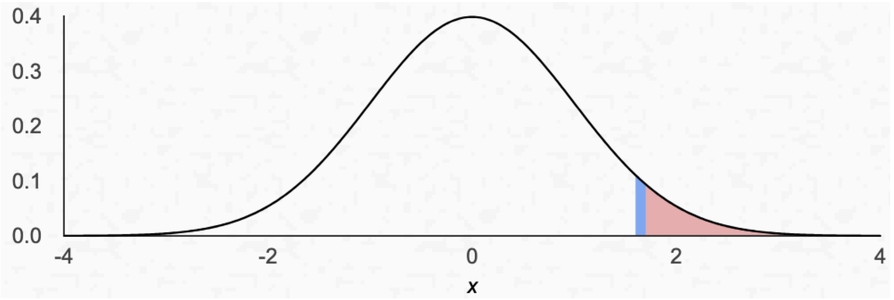




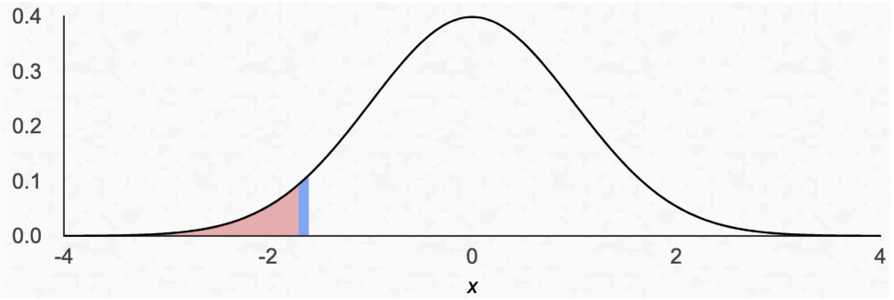


Tails

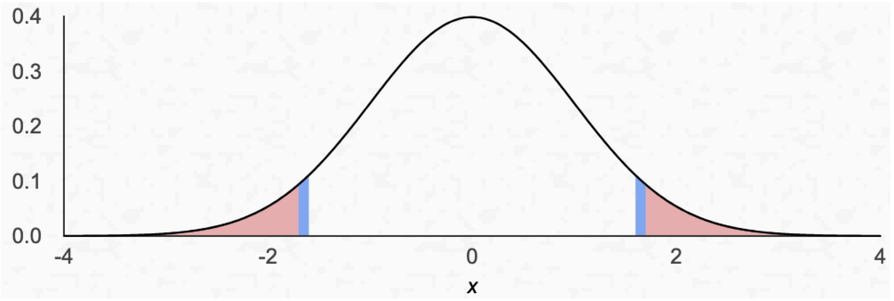
What is considered surprising is dependent on the problem at hand



upper-tail: is x significantly greater?



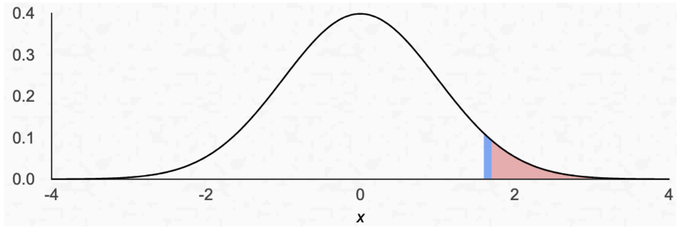
lower-tail: is x significantly smaller?



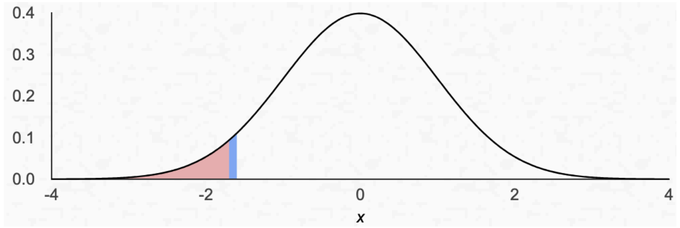
two-tail: is x significantly different?

P VALUES

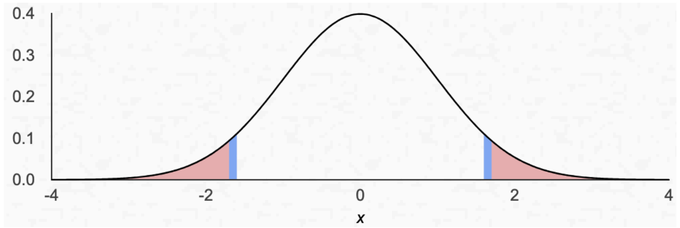
A p value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.



p-value (x) = $P(X \geq x \mid H_0) = 1 - P(X \leq x \mid H_0)$



p-value (x) = $1 - P(X \leq x \mid H_0)$



p-value (x) = $2 \times P(X \leq -\,|x|\,|H_0)$

Recipe for hypothesis testing

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$
- Calculate some statistic x whose conditional distribution is given by $P(X|H_0)$

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$
- Calculate some statistic x whose conditional distribution is given by $P(X|H_0)$
- Calculate p-value

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$
- Calculate some statistic x whose conditional distribution is given by $P(X|H_0)$
- Calculate p-value
- If p-value falls in rejection region then null hypothesis can be rejected; else null hypothesis cannot be rejected

Errors

Errors

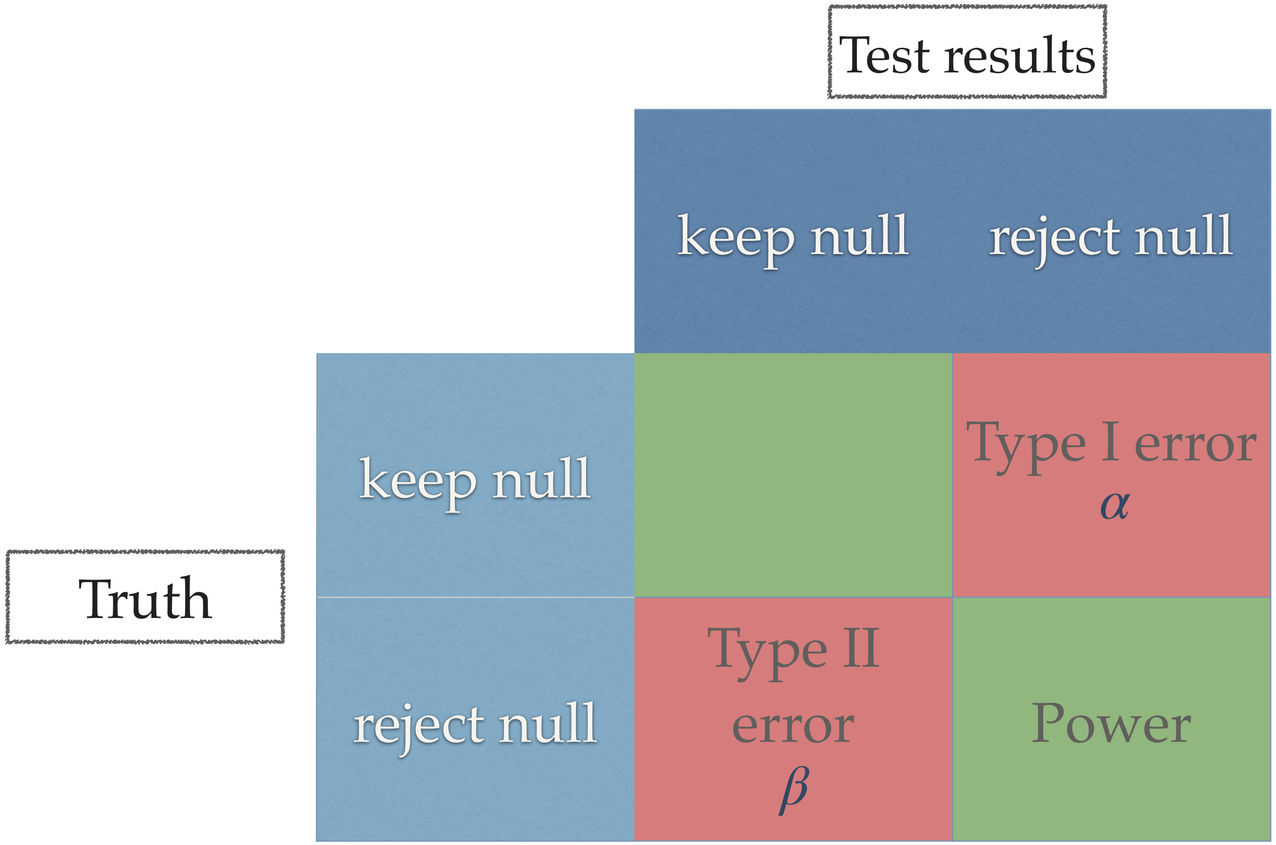
- Type I error: We incorrectly rejected the null hypothesis

Errors

- Type I error: We incorrectly rejected the null hypothesis
- Type II error: We incorrectly failed to reject the null hypothesis

Errors

- Type I error: We incorrectly rejected the null hypothesis
- Type II error: We incorrectly failed to reject the null hypothesis



The Boy who Cried Wolf

- In the Aesop’s fable, assuming that there typically is no wolf, the villagers make two types of errors
  - They believe there was a wolf when there was none
  - They believe there was no wolf when there was one

The Boy who Cried Wolf

The Boy who Cried Wolf

- In the Aesop's fable – The Boy who Cried Wolf – the villagers make two types of errors
  - They believe there was a wolf when there was none [TYPE I]
  - They believe there was no wolf when there was one [TYPE II]

The Boy who Cried Wolf

Errors

Errors

- A well-calibrated statistical test should have acceptable Type I error rate and high statistical power

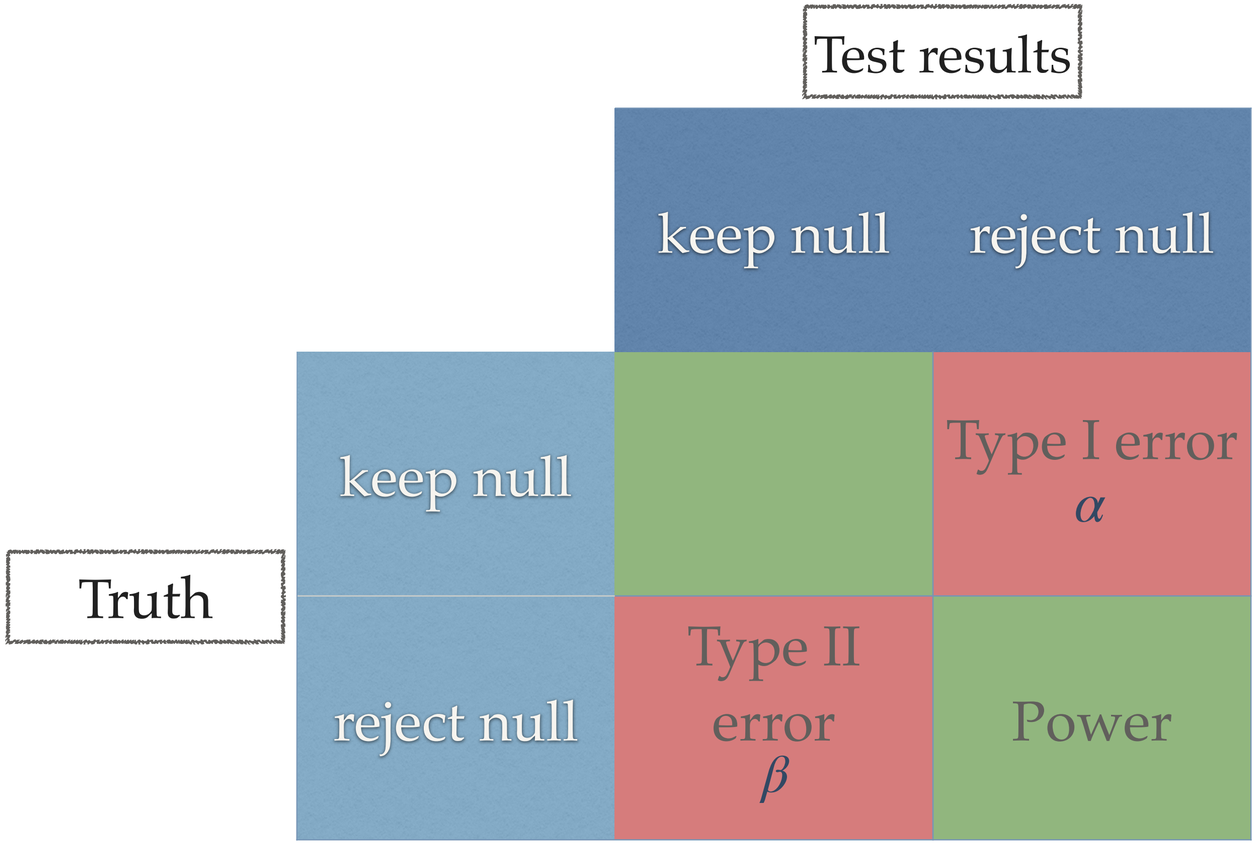
Errors

- A well-calibrated statistical test should have acceptable Type I error rate and high statistical power

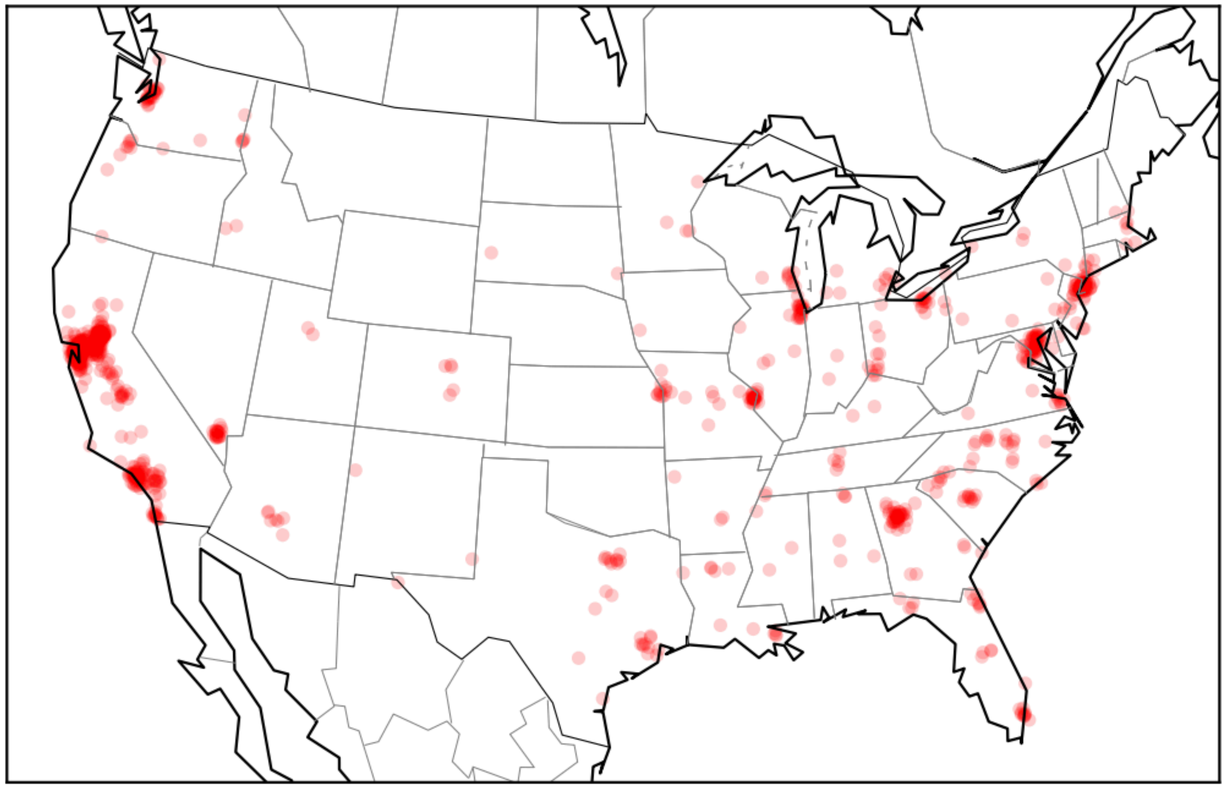
- If $\alpha = 0.05$ and we do 100 tests, we expect to make 5 mistakes

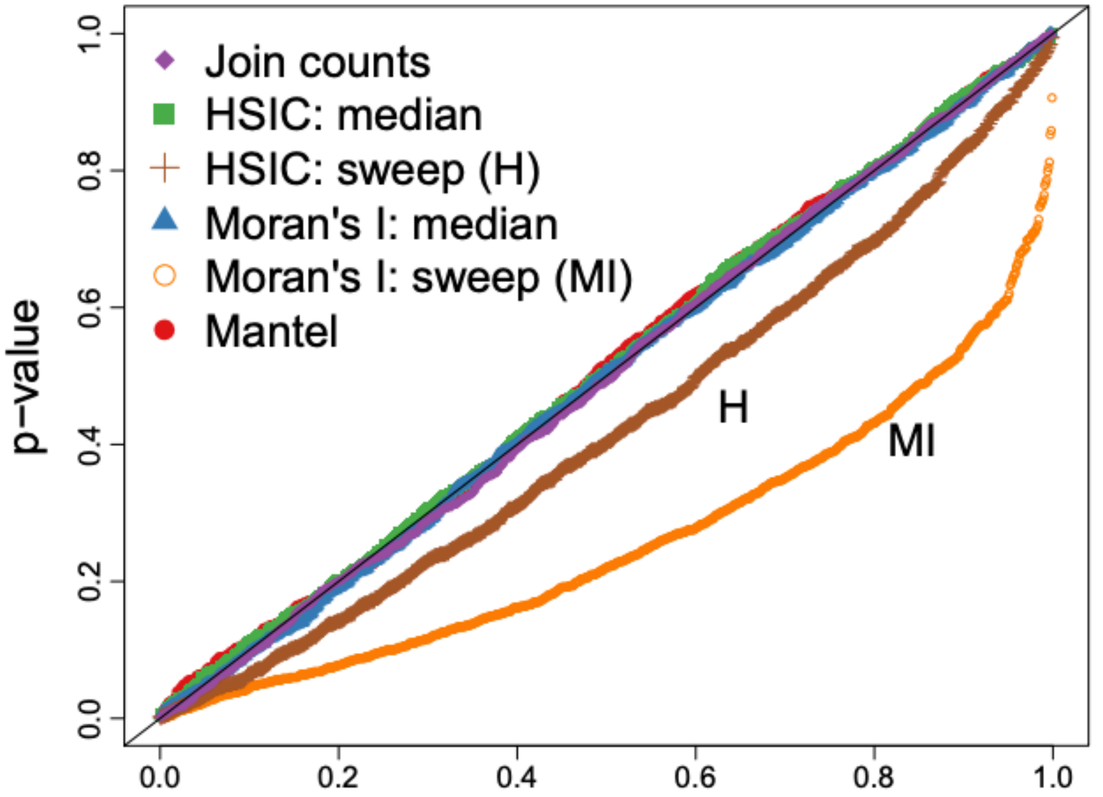
Errors

- A well-calibrated statistical test should have acceptable Type I error rate and high statistical power
- If $\alpha = 0.05$ and we do 100 tests, we expect to make 5 mistakes



Test if lexical variation is independent of geography





Dong Nguyen and Jacob Eisenstein. 2017. A Kernel Independence Test for Geographical Language Variation. Computational Linguistics, 43(3):567–592.

Multiple hypothesis corrections

Multiple hypothesis corrections

- When we do multiple tests, we want to correct for the likelihood of getting statistical significance by chance

Multiple hypothesis corrections

- When we do multiple tests, we want to correct for the likelihood of getting statistical significance by chance
- Apply bonferroni correction which conservatively sets a lower significance threshold based on number of tests

Multiple hypothesis corrections

- When we do multiple tests, we want to correct for the likelihood of getting statistical significance by chance
- Apply bonferroni correction which conservatively sets a lower significance threshold based on number of tests

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

Here the significance level α_0 is adjusted

Confidence intervals

Confidence intervals

- In many instances, instead of doing a statistical test, we want to bound the error of the metric

Confidence intervals

- In many instances, instead of doing a statistical test, we want to bound the error of the metric
- Confidence intervals helps us quantify the range in which the observed metric will lie for the unobserved population

Confidence intervals

Confidence intervals

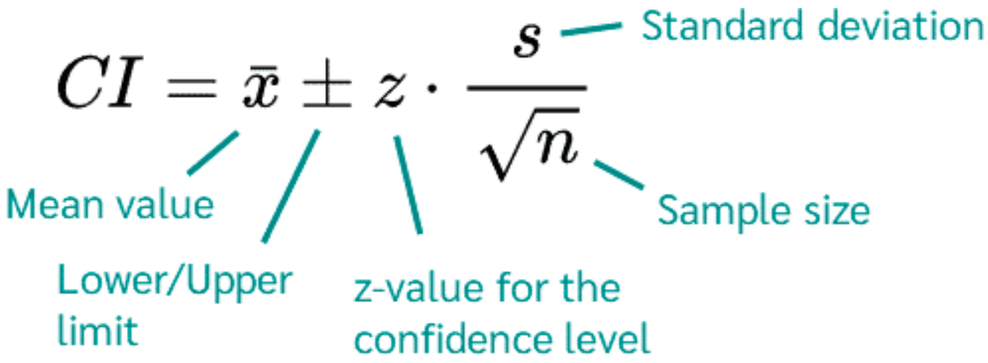
- The CI is statistically determined with some parametric assumptions

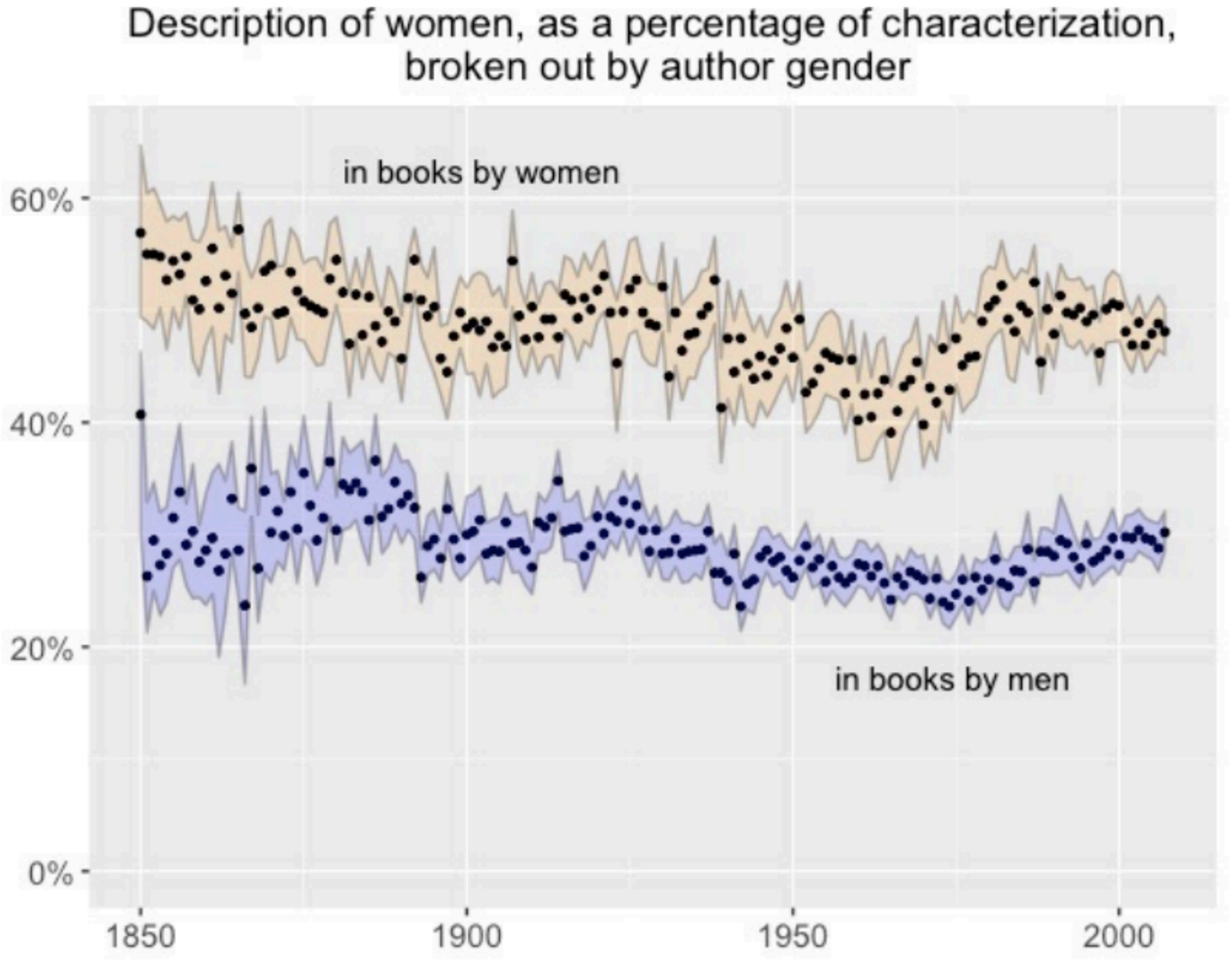
Confidence intervals

- The CI is statistically determined with some parametric assumptions
- The observed value is assumed to be a point estimate of the mean

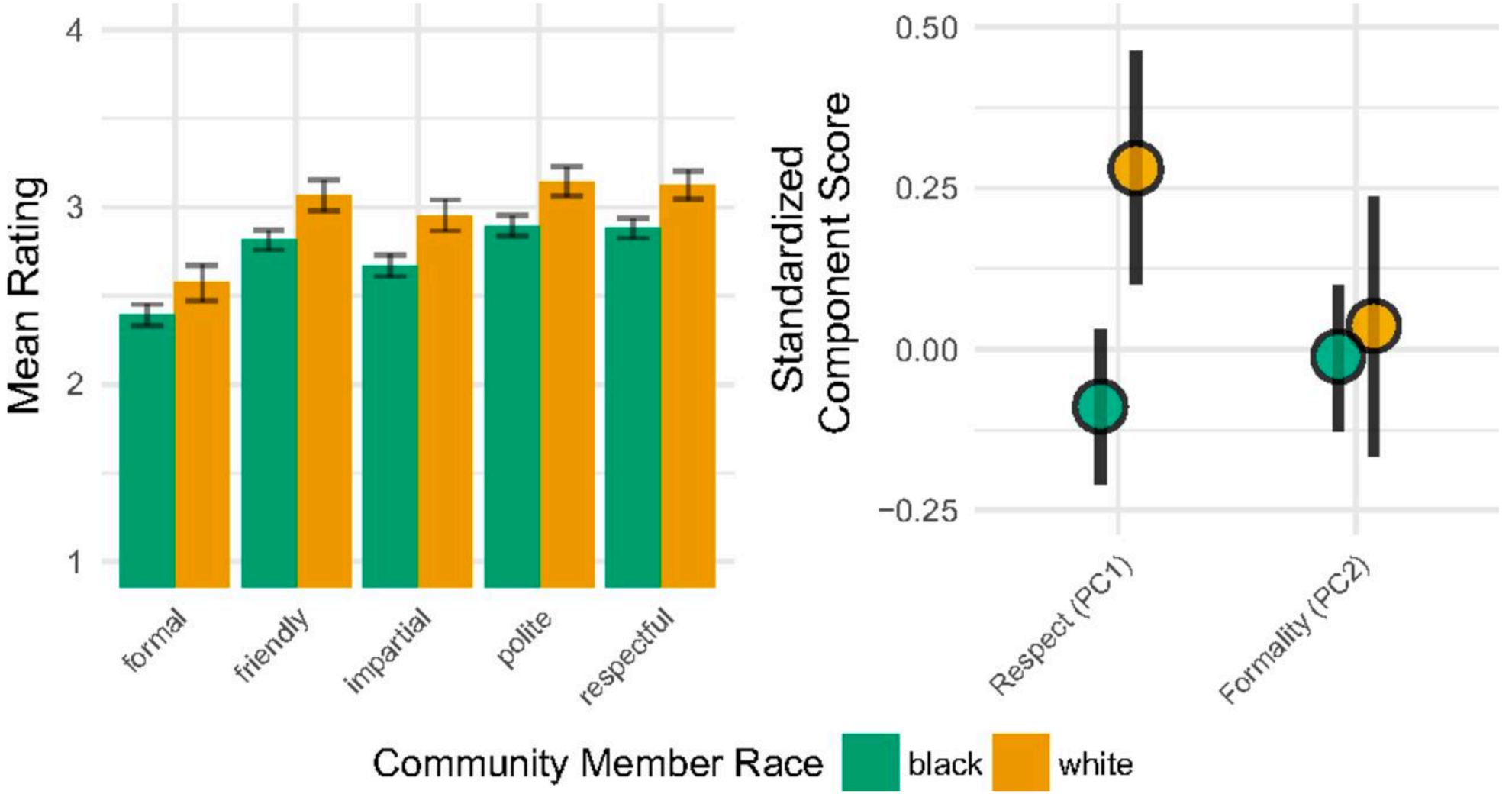
Confidence intervals

- The CI is statistically determined with some parametric assumptions
- The observed value is assumed to be a point estimate of the mean





Underwood, Ted, David Bamman, and Sabrina Lee. "The transformation of gender in English-language fiction." Journal of Cultural Analytics 3.2 (2018).



Voigt, Rob, et al. "Language from police body camera footage shows racial disparities in officer respect." Proceedings of the National Academy of Sciences 114.25 (2017): 6521-6526.

Ablation testing

Ablation testing

- Are a set of features really important for the model?

Ablation testing

- Are a set of features really important for the model?
- One can statistically test this:
  - Create model with all the features and evaluate
  - Create another model with test features removed and evaluate
  - Compare the difference in performance as a statistic and check for its statistical significance

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Recipe for hypothesis testing

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$
- Calculate some statistic x whose conditional distribution is given by $P(X|H_0)$

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$
- Calculate some statistic x whose conditional distribution is given by $P(X|H_0)$
- Calculate p-value

Recipe for hypothesis testing

- Assume significance level α; state H_0 and $H_{alternative}$
- Calculate some statistic x whose conditional distribution is given by $P(X|H_0)$
- Calculate p-value
- If p-value falls in rejection region then null hypothesis can be rejected; else null hypothesis cannot be rejected

Hypothesis testing

Hypothesis testing

- What should be $P(X|H_0)$?

Hypothesis testing

- What should be $P(X|H_0)$?
- In many situations, we can use parametric distributions to characterize $P(X|H_0)$
  - Binomial (probability of success p, #trials n)
  - Normal (mean μ and standard deviation σ)

Parametric tests

Parametric tests

- For these tests, probabilities can be calculated by plugging it in an equation

Parametric tests

- For these tests, probabilities can be calculated by plugging it in an equation
- E.g. Assume we observe 75 heads from 100 trials of a fair coin, then calculate Pr(X=75|p=0.5,n=100) using a binomial distribution's parametric form

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Central limit theorem

Central limit theorem

- Often our statistic is calculated by averaging over multiple independent instances (e.g., accuracy)

Central limit theorem

- Often our statistic is calculated by averaging over multiple independent instances (e.g., accuracy)
- According to CLT, the average of independent and identically distributed random variables tends to be a normal distribution, even if the random variables are not normally distributed

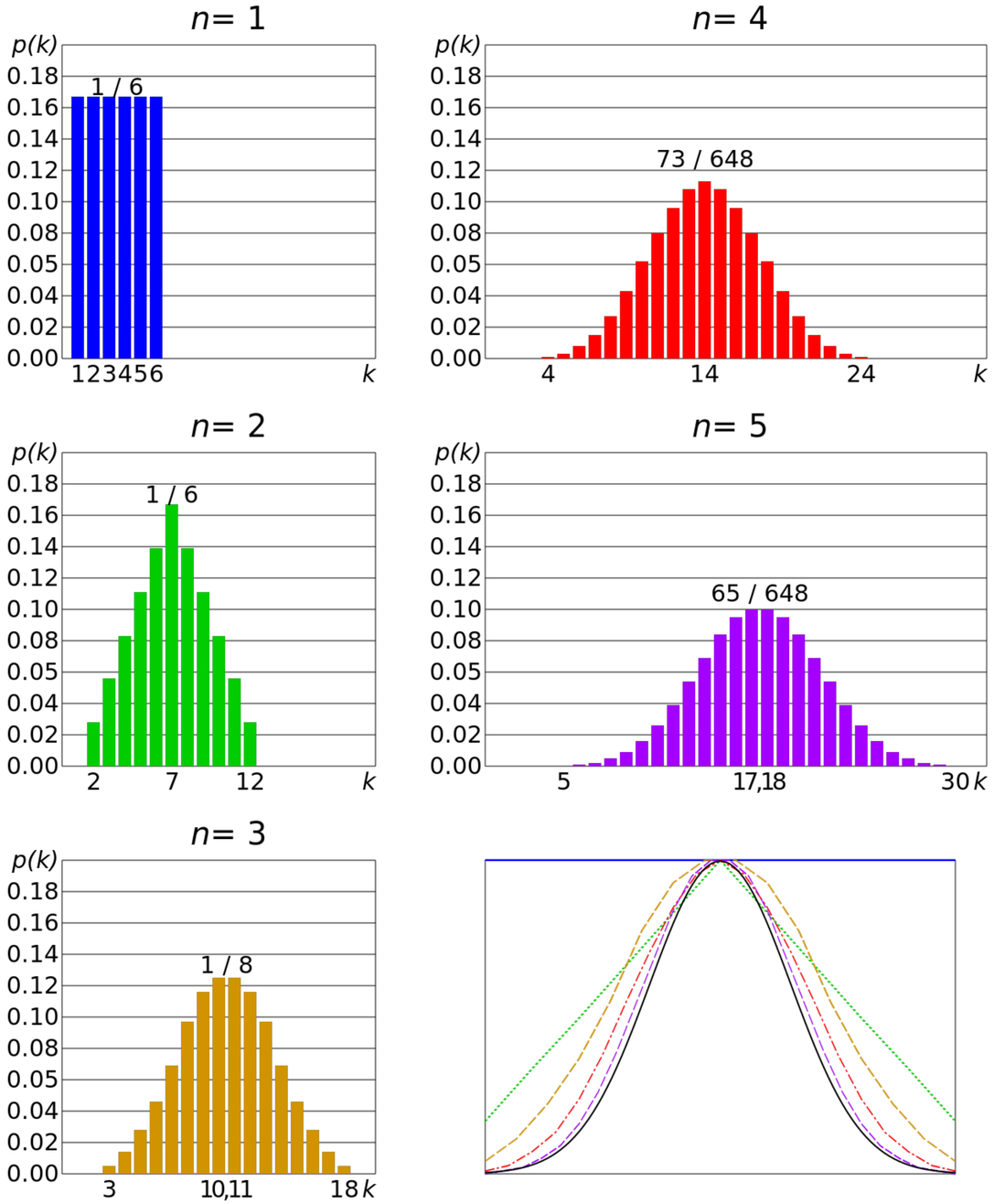


Image credit: Wikipedia

CLT allows us to parametrize the distribution of many statistics that are like sample averages

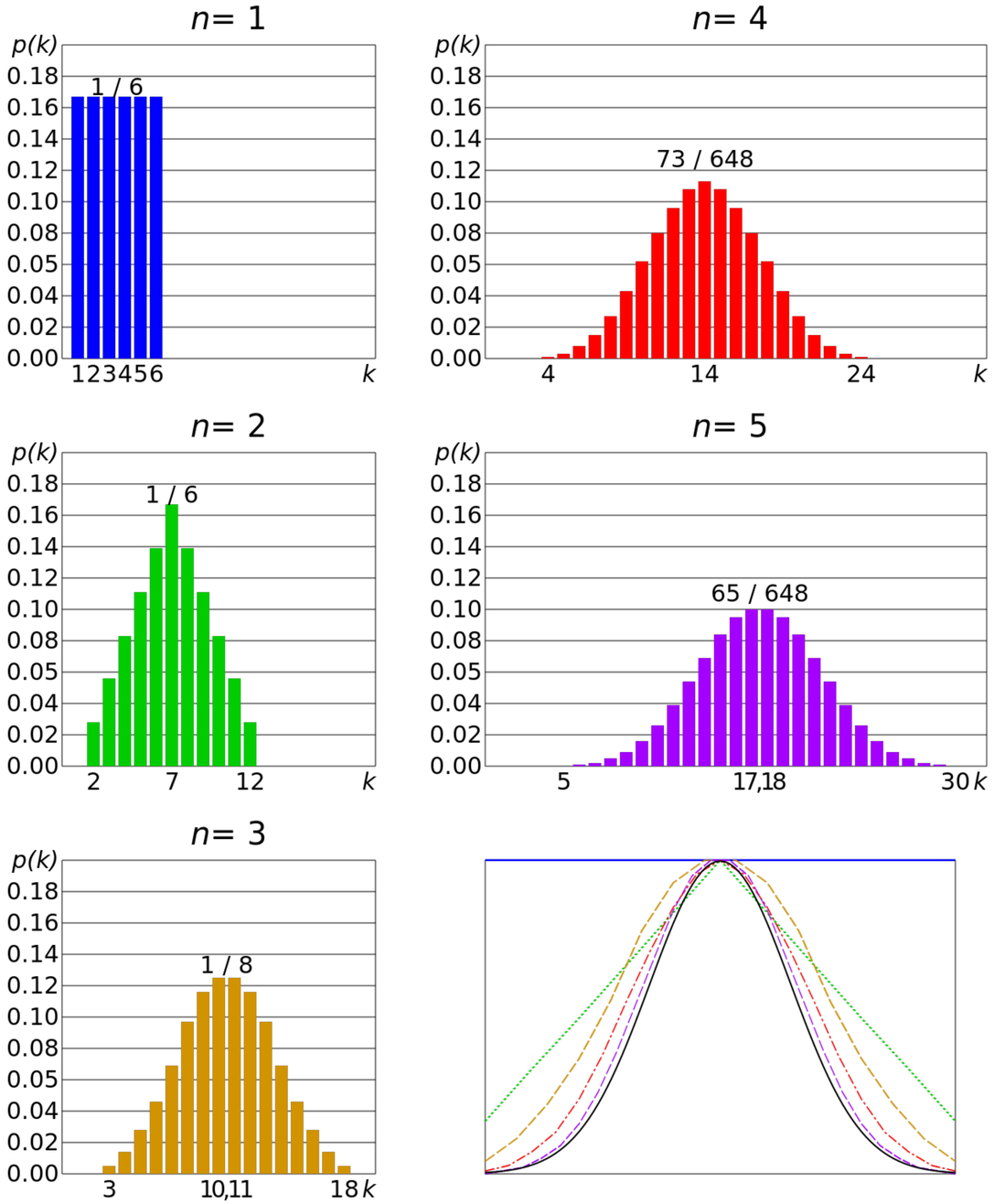


Image credit: Wikipedia

IS CLT ALWAYS APPLICABLE?

IS CLT ALWAYS APPLICABLE?

- Some statistics (e.g., median, reciprocal rank, f-score), are not simple averages.

IS CLT ALWAYS APPLICABLE?

- Some statistics (e.g., median, reciprocal rank, f-score), are not simple averages.
- What should be $P(X|H_0)$ in those cases?

Non parametric tests

Non parametric tests

- Can we construct the null distribution of the sample statistic?

Non parametric tests

- Can we construct the null distribution of the sample statistic?
- Two broadly applicable methods are:
  - Permutation tests
  - Bootstrap

Permutation tests

- Suppose we want to calculate any difference statistic between two categories
  - E.g. Median difference of % female characters between author genders
- H_0 posits no difference (i.e labels do not matter)
- Idea! If labels don't matter then repeatedly randomizing the label assignments to examples should yield us the null distribution of the statistic.

Book name
Gone with the wind
A tale of two cities
War and peace
. . .
. . .
Harry Potter I

Book name	%female characters	Author gender
Gone with the wind	33.4	F
A tale of two cities	45.6	M
War and peace	12.3	M
…	…	…
…	…	…
Harry Potter I	64.1	F

Book name	%female characters	Author gender	Permutation 1
Gone with the wind	33.4	F	M
A tale of two cities	45.6	M	M
War and peace	12.3	M	F
…	…	…	…
…	…	…	…
Harry Potter I	64.1	F	F

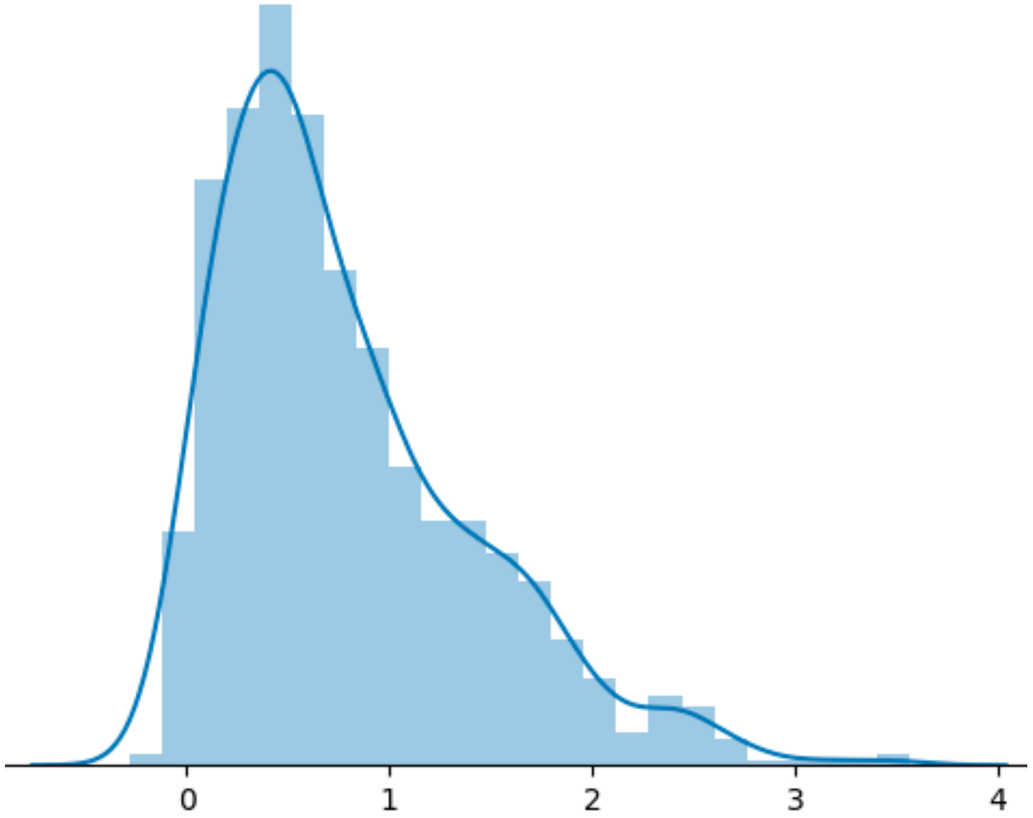
Book name	%female characters	Author gender	Permutation 1	Permutation 2
Gone with the wind	33.4	F	M	F
A tale of two cities	45.6	M	M	M
War and peace	12.3	M	F	F
…	…	…	…	…
…	…	…	…	…
Harry Potter I	64.1	F	F	M

Book name	%female characters	Author gender	Permutation 1	Permutation 2	…	Permutation n
Gone with the wind	33.4	F	M	F	…	M
A tale of two cities	45.6	M	M	M	…	F
War and peace	12.3	M	F	F	…	M
…	…	…	…	…	…	…
…	…	…	…	…	…	…
Harry Potter I	64.1	F	F	M	…	F

Book name	%female characters	Author gender	Permutation 1	Permutation 2	…	Permutation n
Gone with the wind	33.4	F	M	F	…	M
A tale of two cities	45.6	M	M	M	…	F
War and peace	12.3	M	F	F	…	M
…	…	…	…	…	…	…
…	…	…	…	…	…	…
Harry Potter I	64.1	F	F	M	…	F
	Median difference	3.9	-1.4	2.1	…	1.6

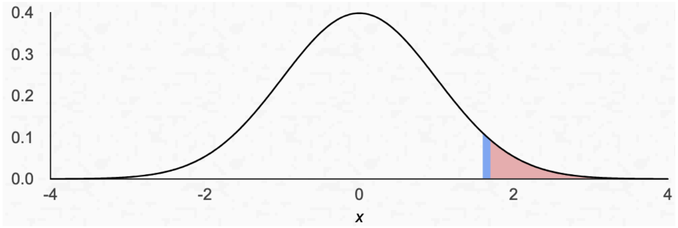
Book name	%female characters	Author gender	Permutation 1	Permutation 2	…	Permutation n
Gone with the wind	33.4	F	M	F	…	M
A tale of two cities	45.6	M	M	M	…	F
War and peace	12.3	M	F	F	…	M
…	…	…	…	…	…	…
…	…	…	…	…	…	…
Harry Potter I	64.1	F	F	M	…	F
	Median difference	3.9	-1.4	2.1	…	1.6

Is the observed statistic unusual compared to the statistics from all the permutations?

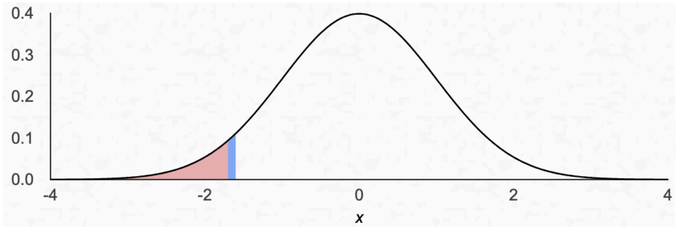


Is 3.9 unusual for this constructed null distribution?

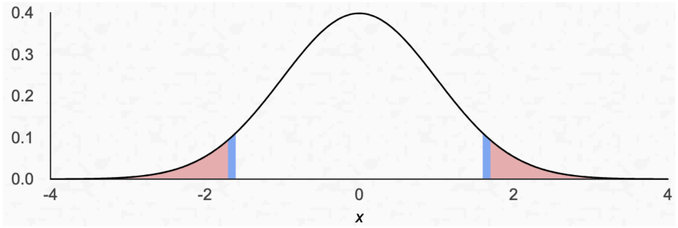
P VALUE



$$\text{p-value (x)} = P(X \geq x \mid H_0) = 1 - P(X \leq x \mid H_0)$$



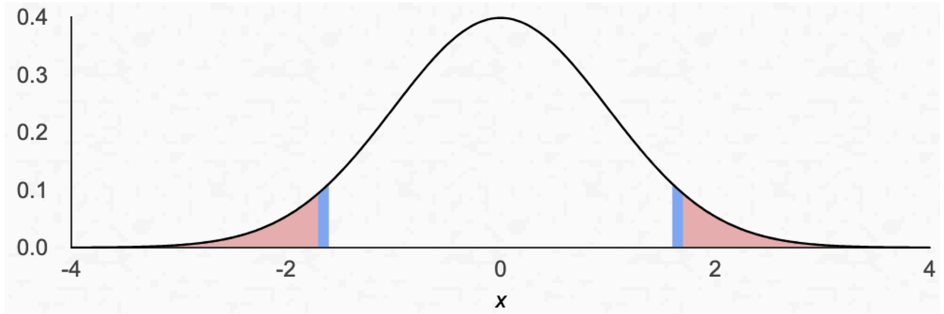
$$\text{p-value (x)} = 1 - P(X \leq x \mid H_0)$$



$$\text{p-value (x)} = 2 \times P(X \leq - \mid x \mid \mid H_0)$$

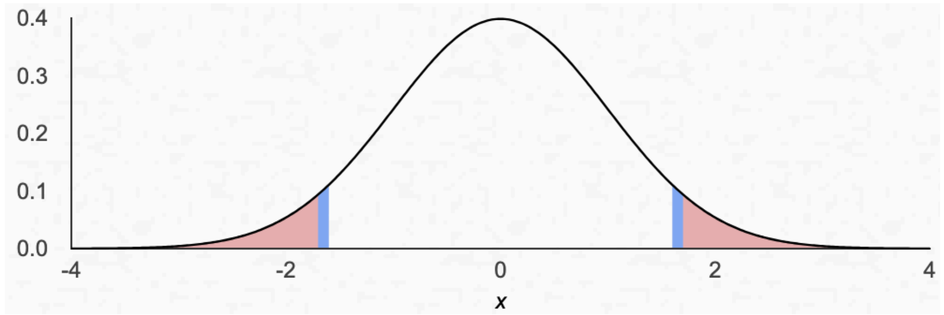
Empirical p value

$$p_{\text{empirical}} = \frac{1}{N} \sum_{i}^{N} I[abs(m_{\text{obs}}) < abs(m_{\text{perm}}^{(i)})]$$



Empirical p value

$$p_{\text{empirical}} = \frac{1}{N} \sum_{i}^{N} I[abs(m_{\text{obs}}) < abs(m_{\text{perm}}^{(i)})]$$



- N = number of permutations
- m_{obs} is the observed value of the statistic
- m_{perm} is the value of the statistic under permutation
- $I[\cdot]$ is an indicator function

Permutation tests

Permutation tests

- Permutation tests have broad application

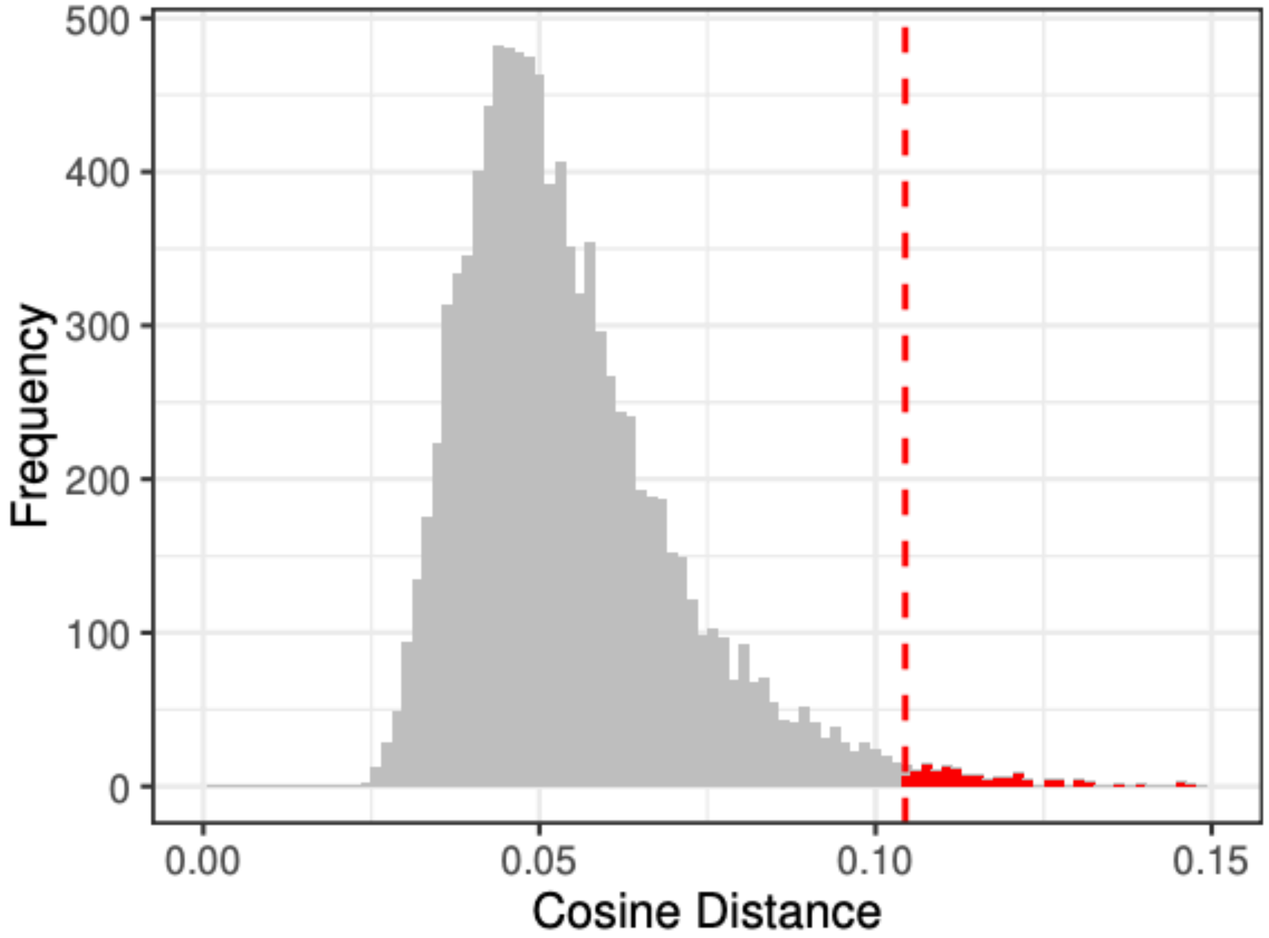
Permutation tests

- Permutation tests have broad application
- You don’t make any assumptions about the shape and form of the null distribution for the statistic

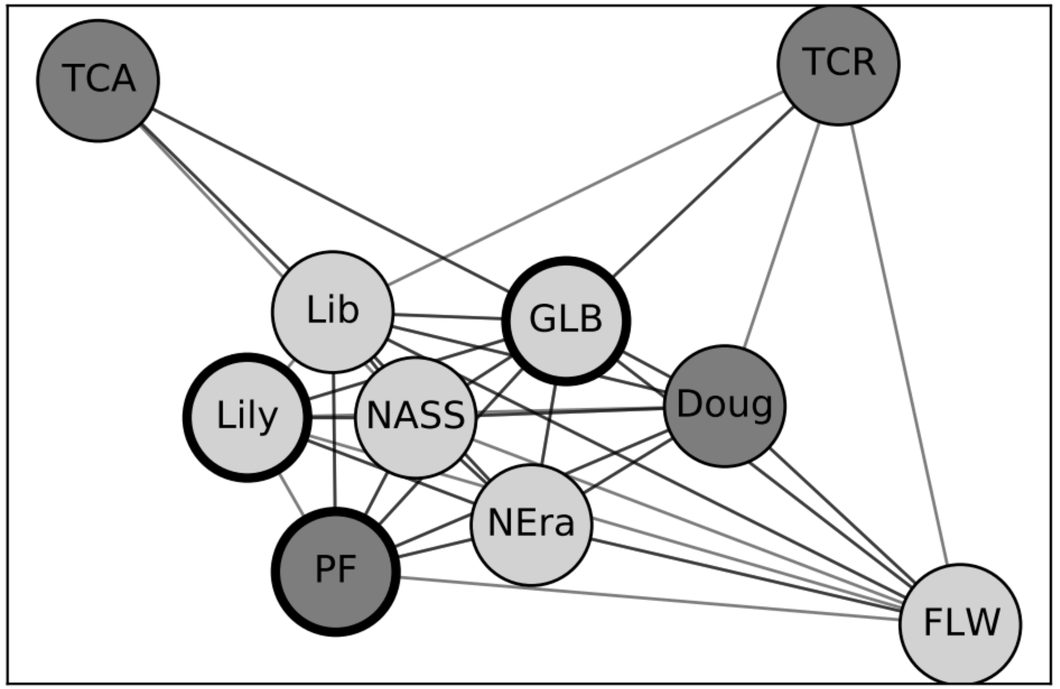
Permutation tests

- Permutation tests have broad application
- You don’t make any assumptions about the shape and form of the null distribution for the statistic
- You can use pretty much any statistic!

- Is there a significant change in meaning of some word (e.g., shovel)?



Yang Liu, Alan Medlar, and Dorota Glowacka. 2021. Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings. In Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.



Doug	Douglass Newspapers
FLW	Frank Leslie's Weekly
GLB	Godey's Lady's Book
Lib	The Liberator
Lily	The Lily
NASS	National Anti-Slavery Standard
NEra	The National Era
PF	The Provincial Freeman
TCA	The Colored American
TCR	The Christian Recorder

- Which newspaper is a consistent leader of some other newspaper?

Soni, Sandeep, Lauren F. Klein, and Jacob Eisenstein. "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers." *Journal of Cultural Analytics* 6.1 (2021).

Bootstrap

Bootstrap

- In permutation test, the data is not changed – just the assignment of labels

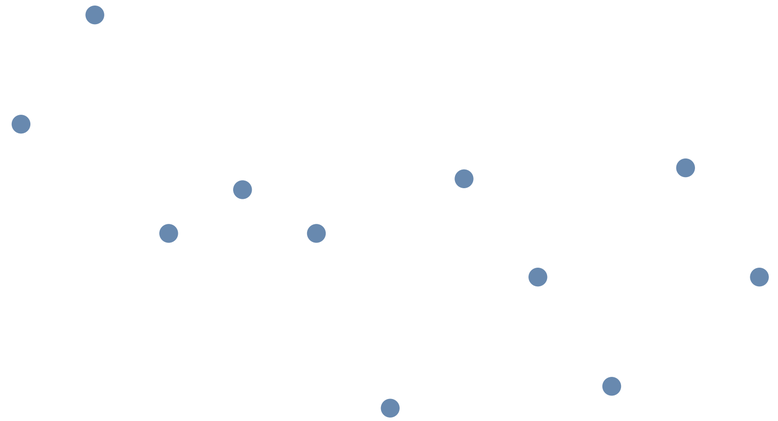
Bootstrap

- In permutation test, the data is not changed – just the assignment of labels
- The variability in the data can be quantified by constructing hypothetical datasets that follow the same distribution

Bootstrap

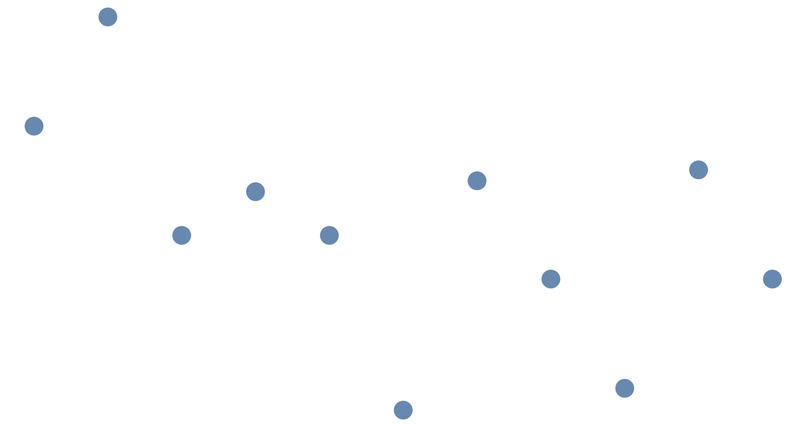
- In permutation test, the data is not changed – just the assignment of labels
- The variability in the data can be quantified by constructing hypothetical datasets that follow the same distribution
- This is the idea of bootstrapping!

Bootstrap



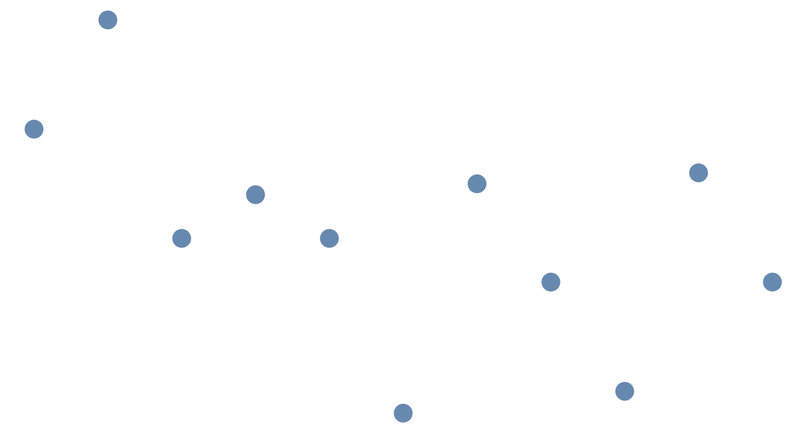
Bootstrap

- Our sample is assumed to be iid, and represents the population distribution

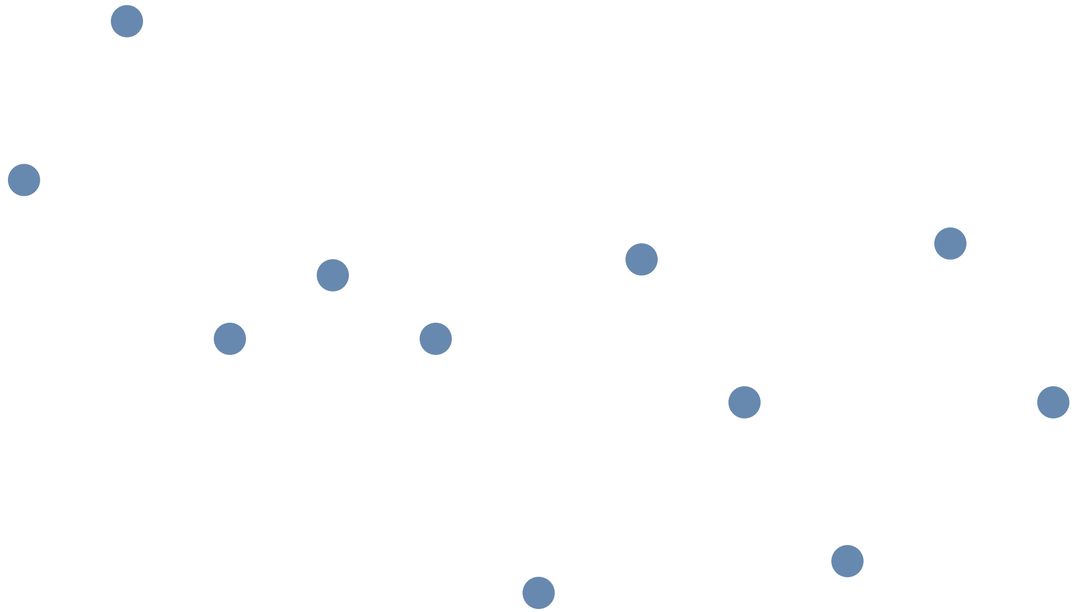


Bootstrap

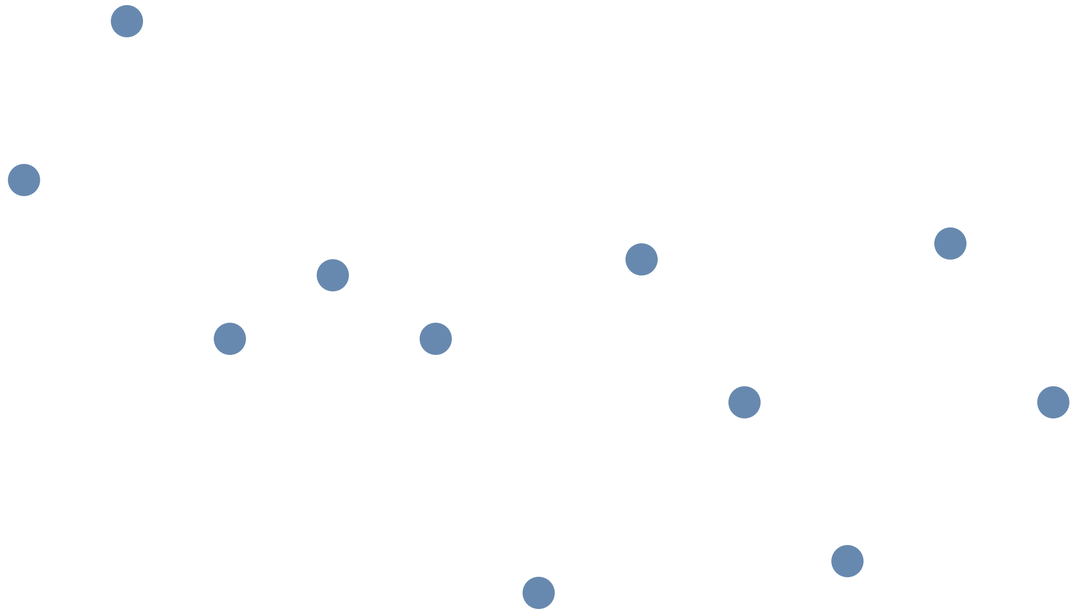
- Our sample is assumed to be iid, and represents the population distribution
- Repeatedly sample from our sample to generate many alternative datasets



Distribution of our sample

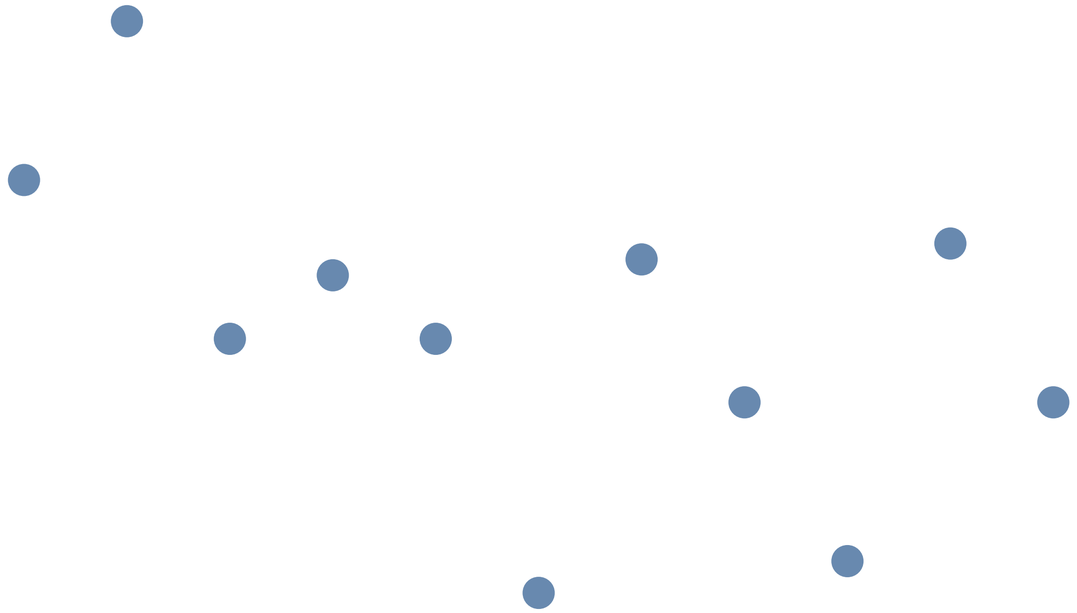


Distribution of our sample



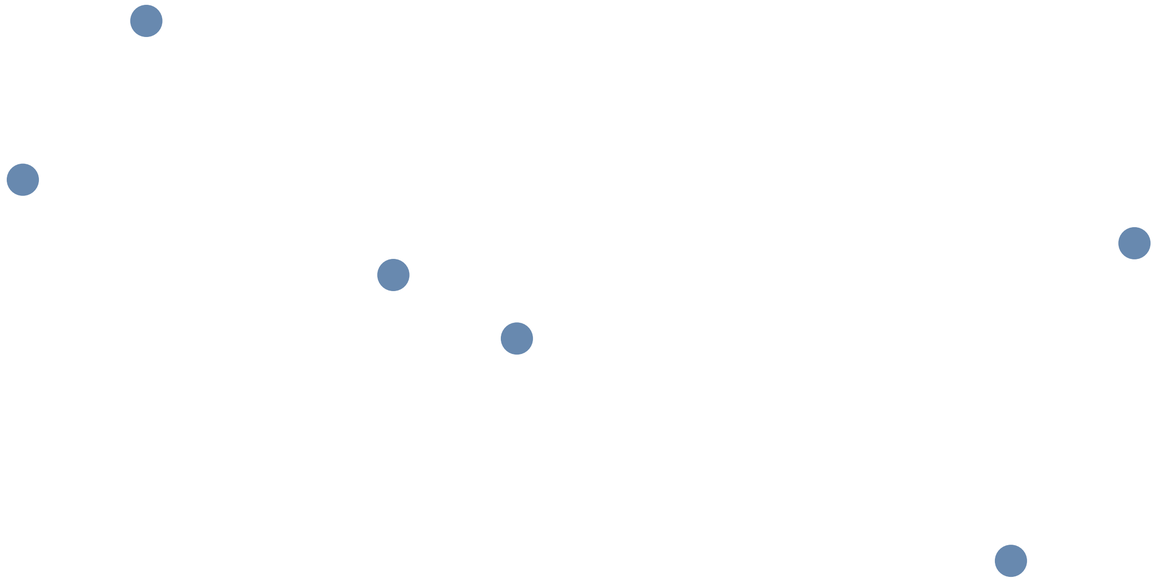
- We can calculate some metric on this sample such as accuracy

Distribution of our sample

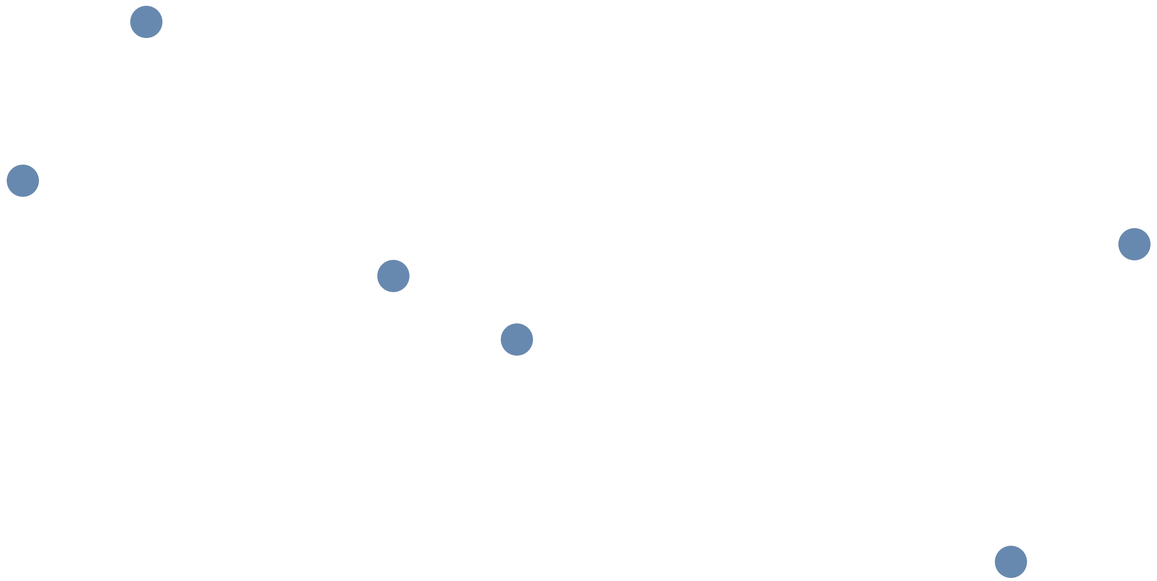


- We can calculate some metric on this sample such as accuracy
- acc_{obs}

Bootstrap 1

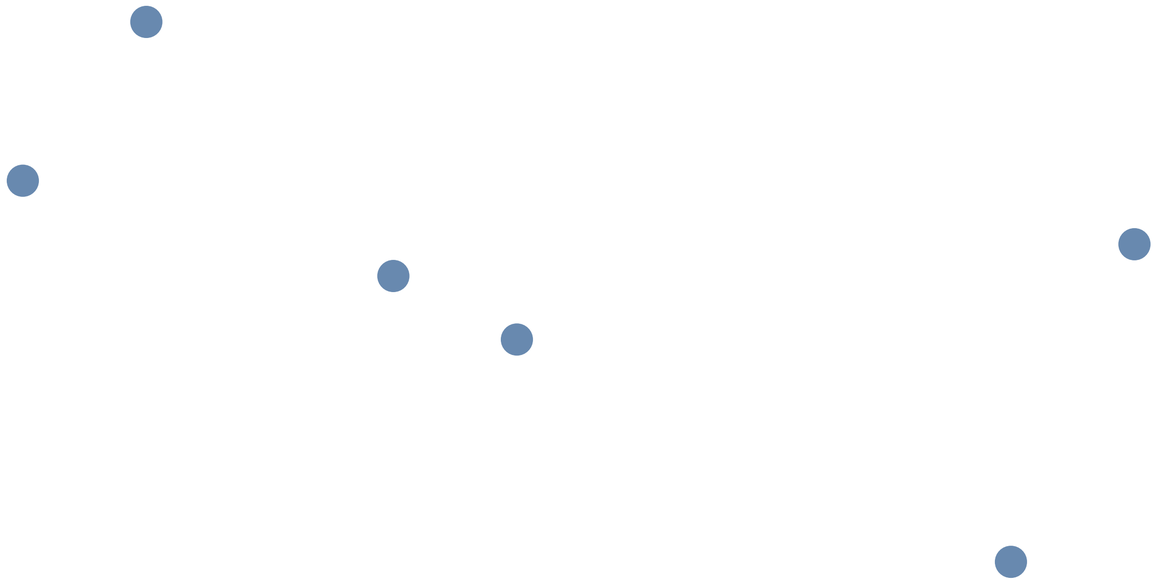


Bootstrap 1



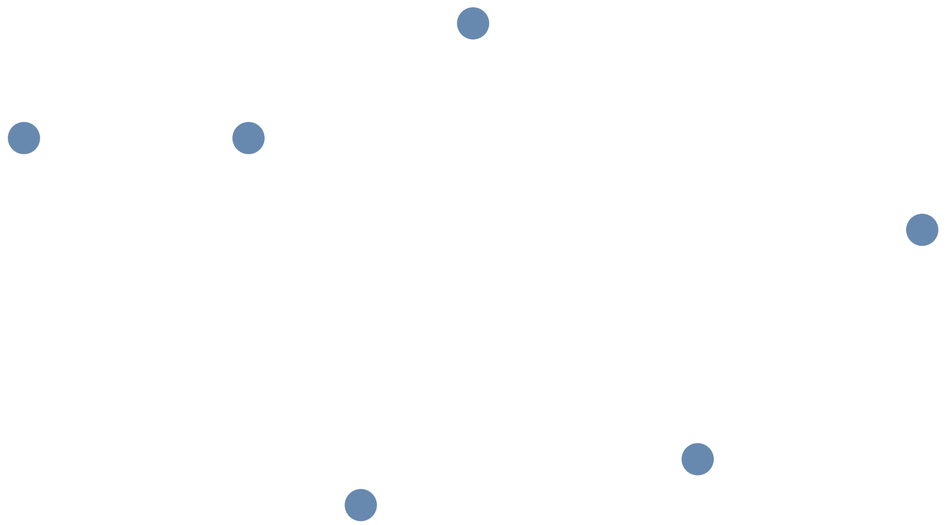
- Calculate accuracy on this bootstrapped sample

Bootstrap 1

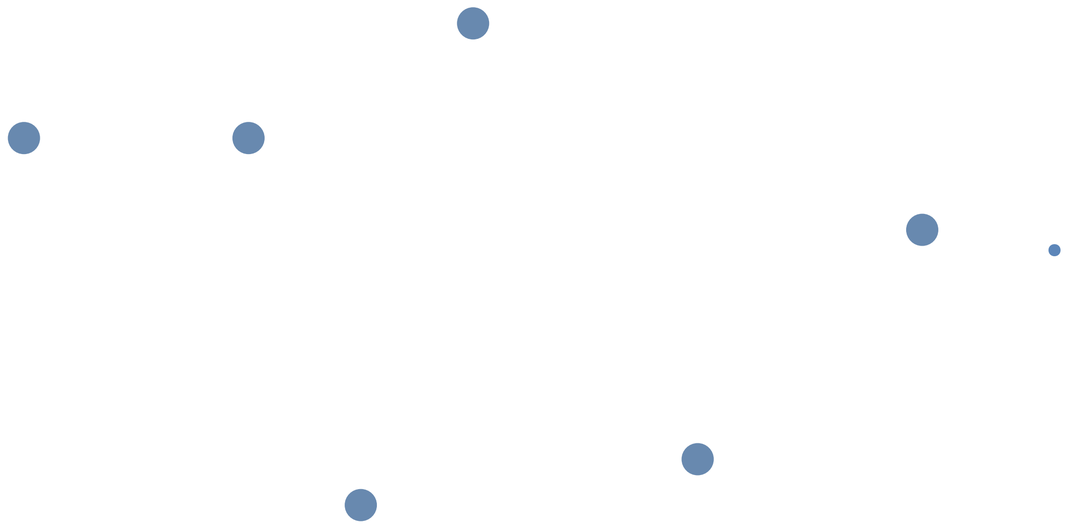


- Calculate accuracy on this bootstrapped sample
- acc_{boot}

Bootstrap 2

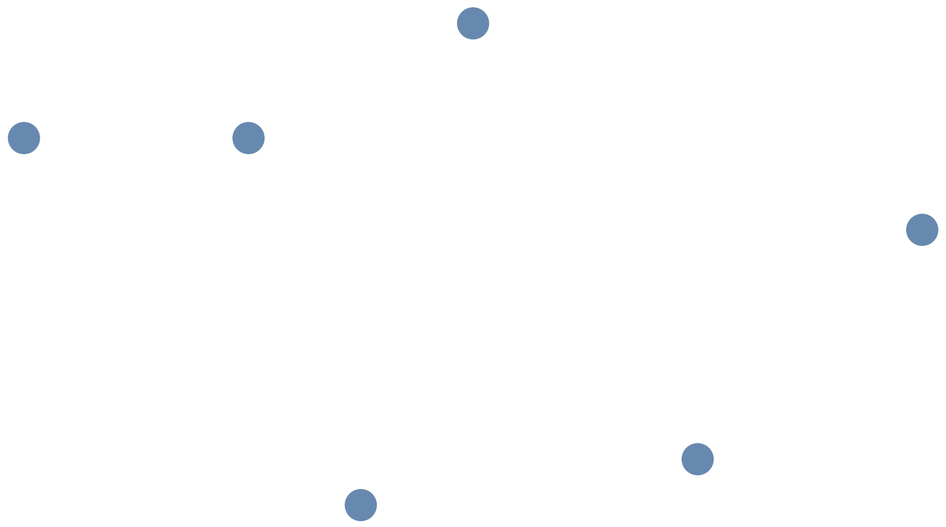


Bootstrap 2



- Again calculate the accuracy

Bootstrap 2



- Again calculate the accuracy
- acc_{boot}

And so on

Bootstrapped Confidence intervals

Bootstrapped Confidence intervals

- From b boostrap samples, we can get a size b array as estimates for accuracy

Bootstrapped Confidence intervals

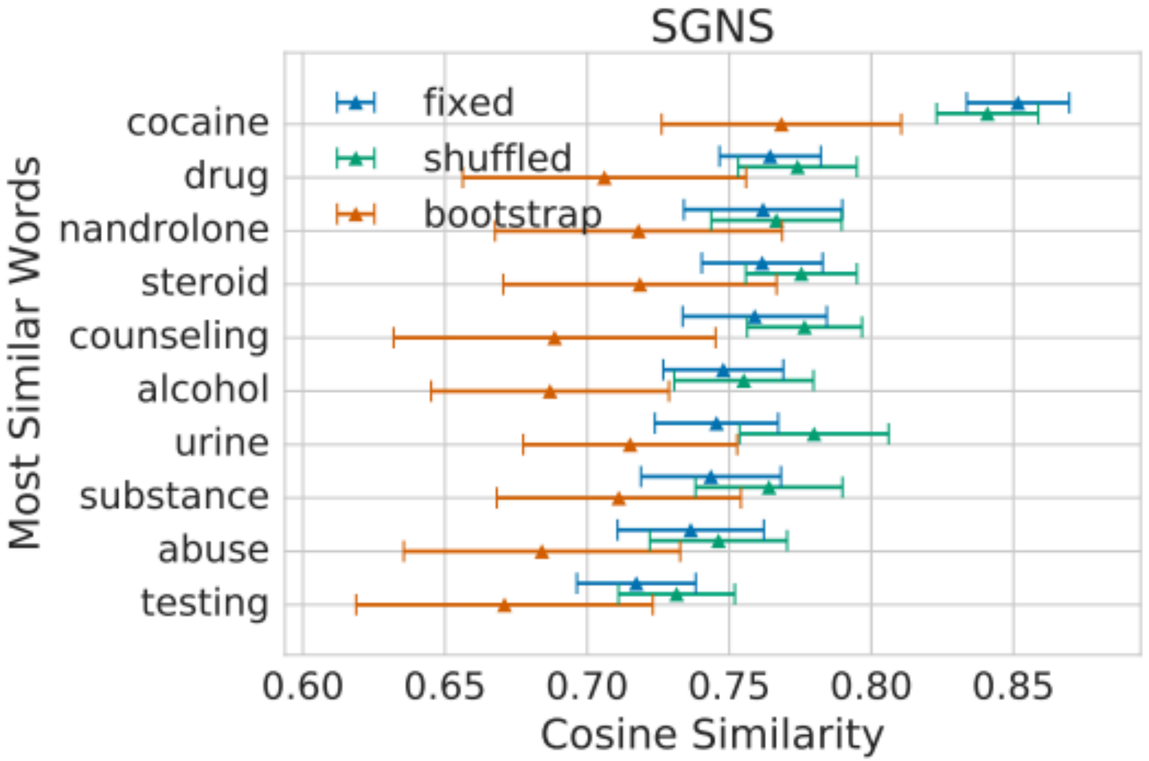
- From b boostrap samples, we can get a size b array as estimates for accuracy
- We can rank these estimates and find the middle 95% to define a range in which the accuracy estimate falls

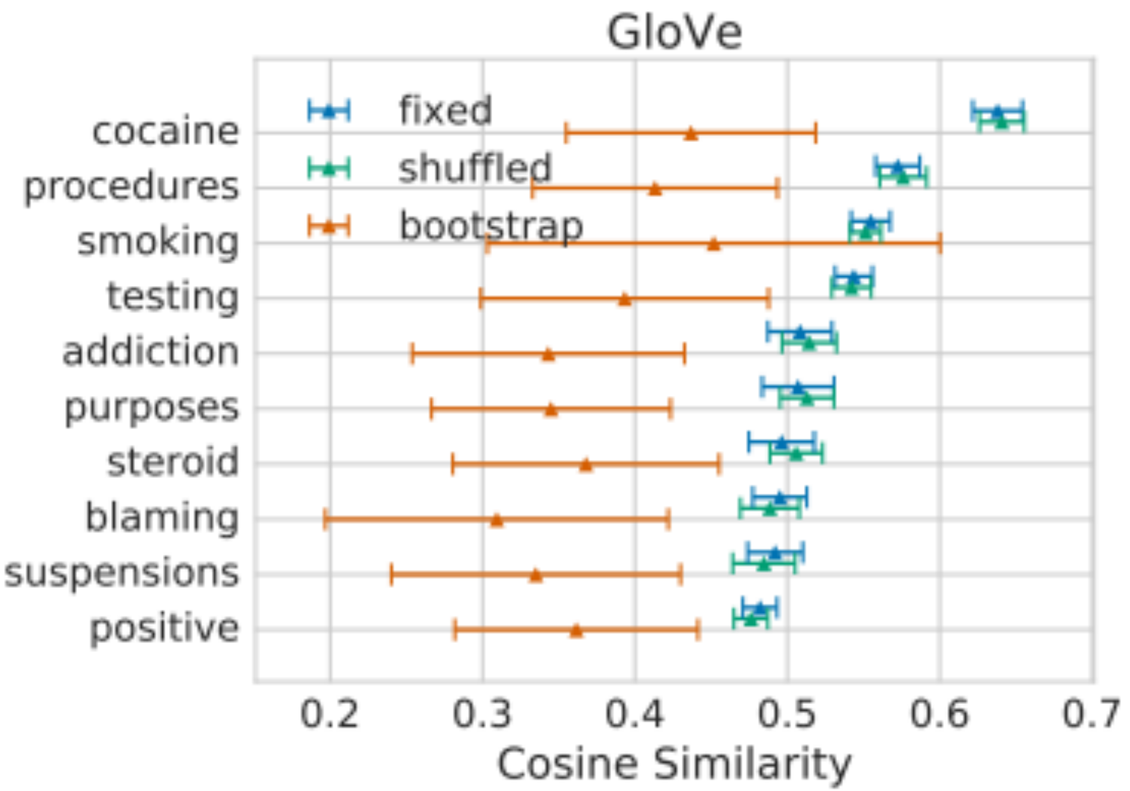
Bootstrapped Confidence intervals

- From b boostrap samples, we can get a size b array as estimates for accuracy
- We can rank these estimates and find the middle 95% to define a range in which the accuracy estimate falls
- 95% CI is [2.5,97.5] percentile

Bootstrapped Confidence intervals

- From b boostrap samples, we can get a size b array as estimates for accuracy
- We can rank these estimates and find the middle 95% to define a range in which the accuracy estimate falls
- 95% CI is [2.5,97.5] percentile
- For large b (e.g., b=1000), this gives a tight bound for the CI





Which words are most similar to `marijuana`?

Maria Antoniak and David Mimno. 2018. Evaluating the Stability of Embedding-based Word Similarities. Transactions of the Association for Computational Linguistics, 6:107–119.

Summary

Summary

- It's often a good idea to give a confidence interval for your estimated metric

Summary

- It's often a good idea to give a confidence interval for your estimated metric
- Statistical tests are useful to verify claims

Summary

- It's often a good idea to give a confidence interval for your estimated metric
- Statistical tests are useful to verify claims
- Use ablation testing to assess the importance of model components

In class

- Parametric test
- Non parametric test