Non Parametric Hypothesis Testing

Sandeep Soni

10/04/2023

Question for the day

"How to statistically test a claim using non-parametric methods?"

Hypothesis Testing

Hypothesis Testing

- We assume that the sample data is drawn from a null distribution

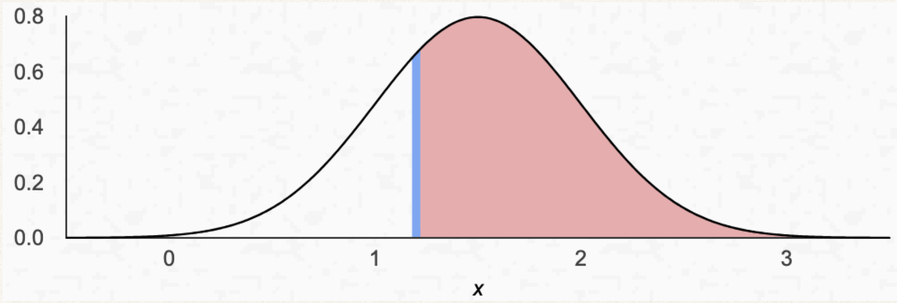
Hypothesis Testing

- We assume that the sample data is drawn from a null distribution
- When can we say that a sample statistic is so extreme that it was unlikely to have been generated from the null distribution?

Hypothesis testing

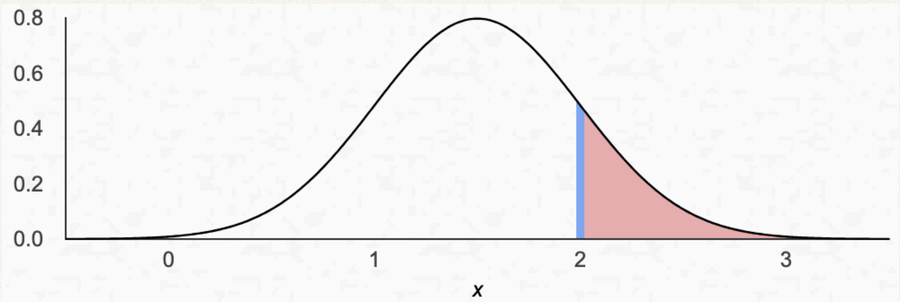
Which observation would you say is surprising if the null distribution holds?

x=1.25



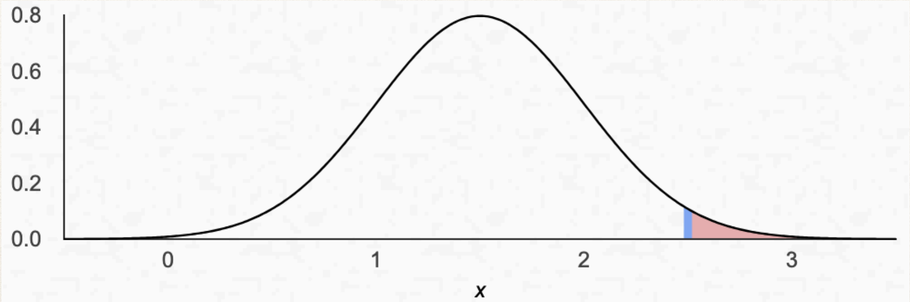
p=0.69

x=2



p=0.16

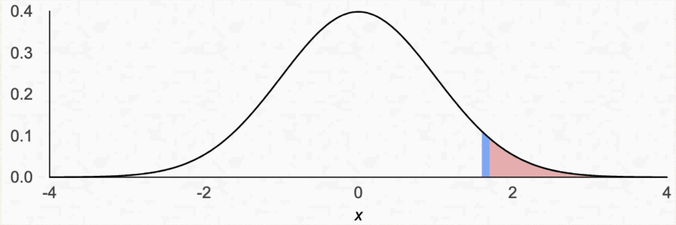
x=2.5



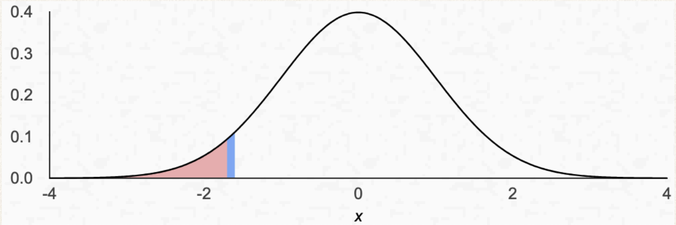
p=0.02

P VALUES

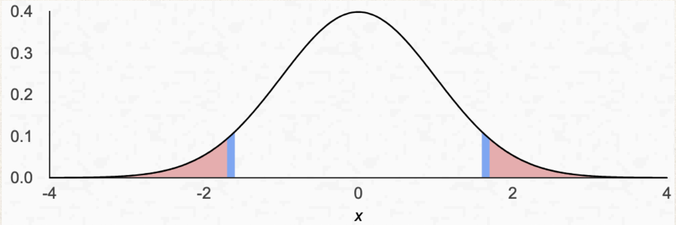
A p value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.



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) = $2 \times P(X \leq -|x| \mid H_0)$

Recipe for hypothesis testing

Recipe for hypothesis testing

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

- In these tests, we can calculate the probabilities by plugging it in an equation

Parametric tests

- In these tests, we can calculate the probabilities by plugging it in an equation
- For example, if we assume 100 trials of a fair coin, and if we observe 75 heads, then calculate P(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

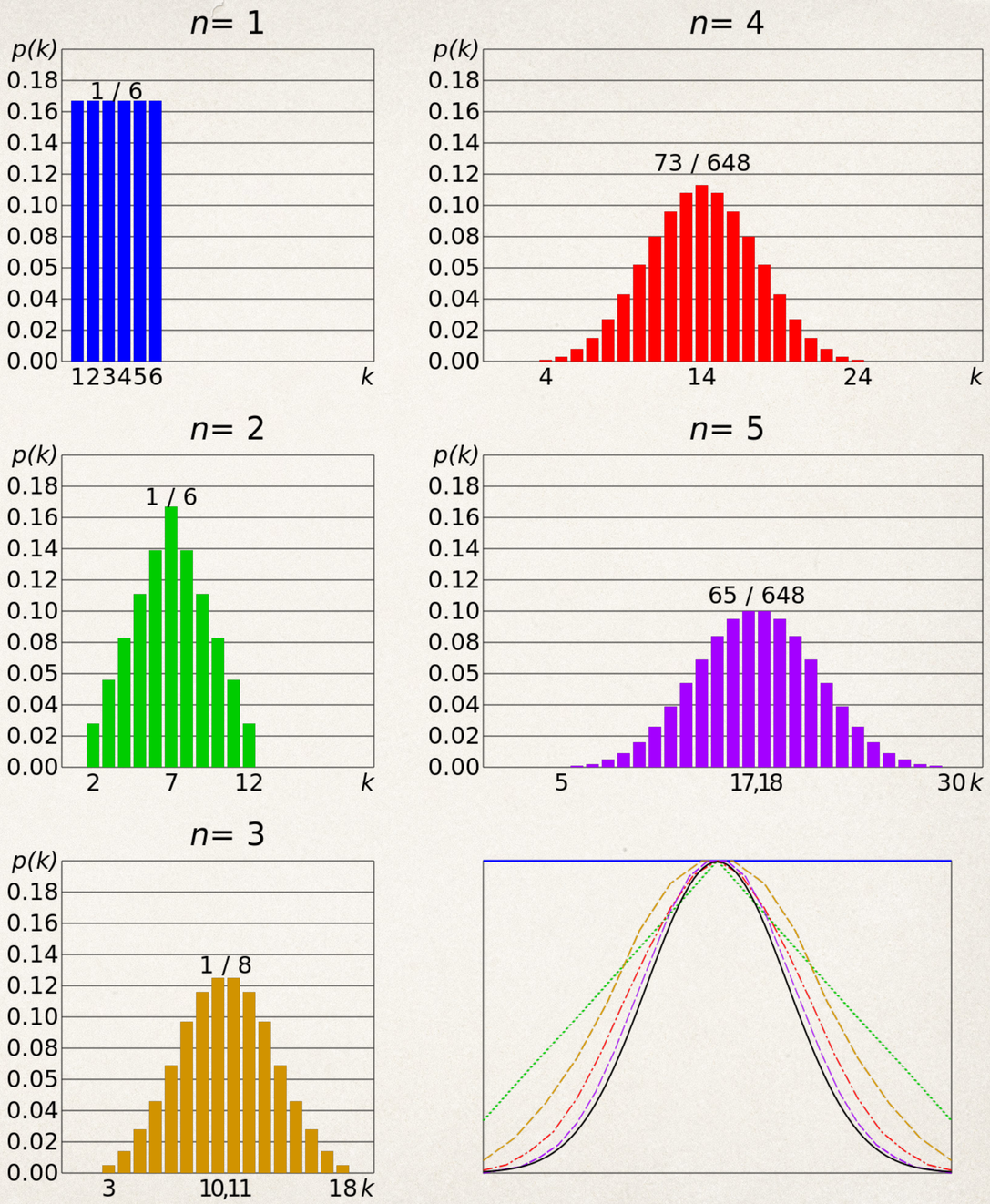


Image credit: Wikipedia

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

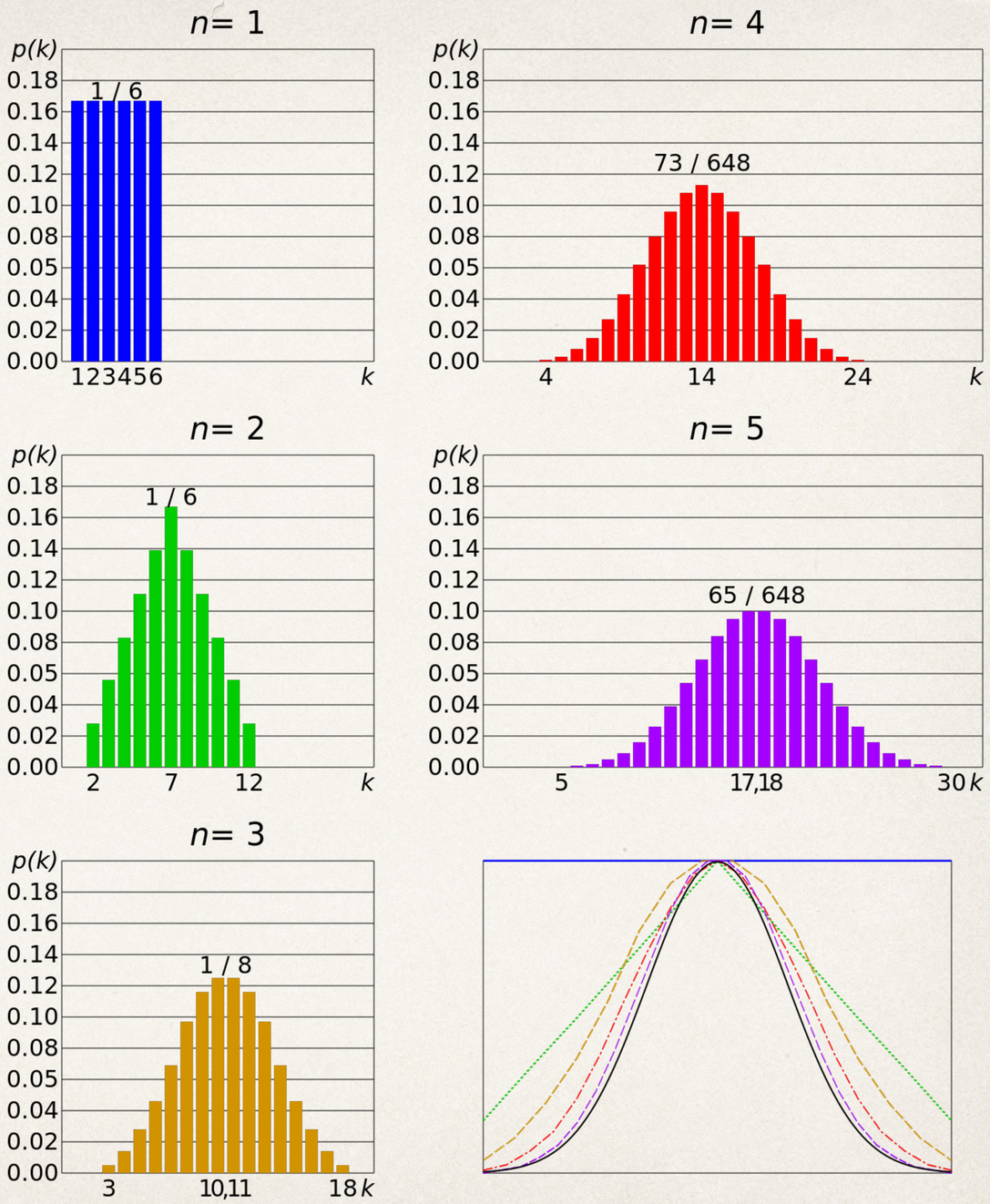


Image credit: Wikipedia

IS CLT ALWAYS APPLICABLE?

IS CLT ALWAYS APPLICABLE?

- For some statistics (e.g., median, reciprocal rank, f-score), we're not doing simple averaging

IS CLT ALWAYS APPLICABLE?

- For some statistics (e.g., median, reciprocal rank, f-score), we're not doing simple averaging
- 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 some difference between two categories
- Null hypothesis will be that there is no difference (i.e labels should not matter)
- If labels don’t matter then repeatedly randomizing the label assignments to examples should yield us the null distribution

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

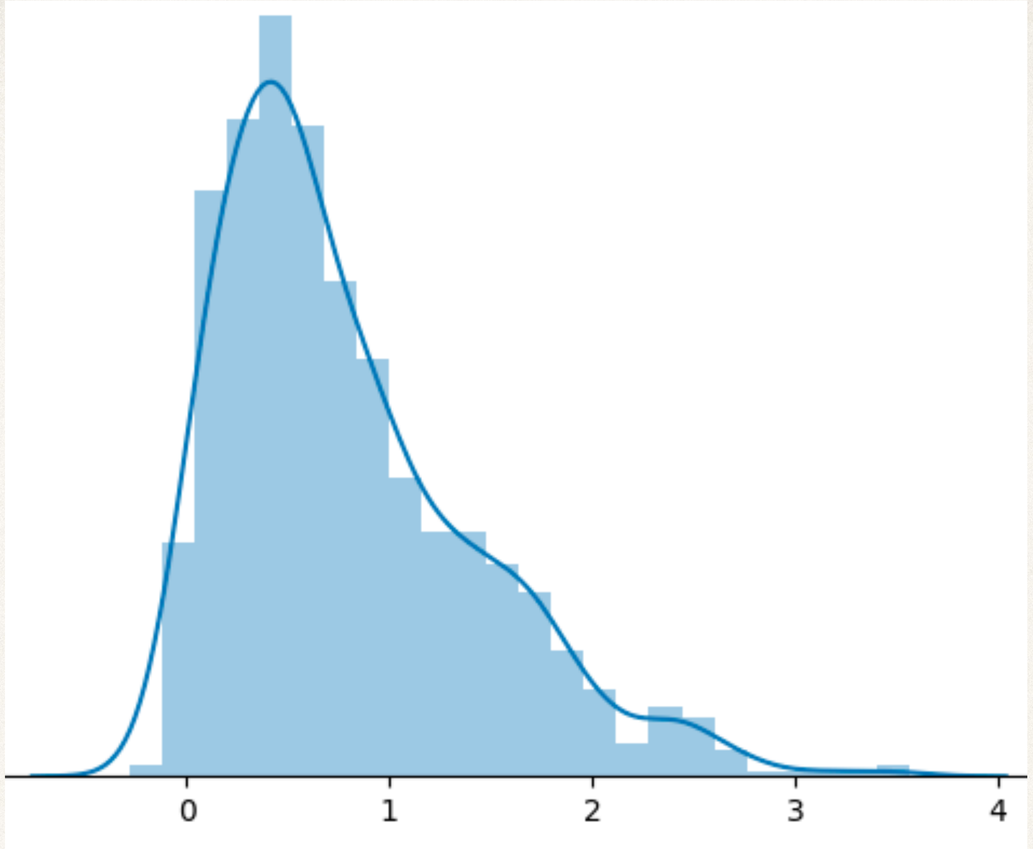
Book name	%female characters	Author gender	Permutation 1	Permutation 2
Gone with the wind	33.4	F	M	F
A tale of two cities	45.6	M	M	M
War and peace	12.3	M	F	F
…	…	…	…	…
…	…	…	…	…
Harry Potter I	64.1	F	F	M

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

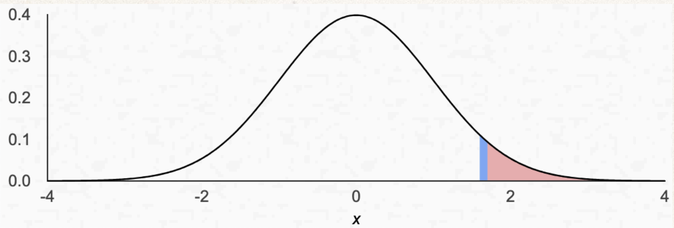
Book name	%female characters	Author gender	Permutation 1	Permutation 2	…	Permutation n
Gone with the wind	33.4	F	M	F	…	M
A tale of two cities	45.6	M	M	M	…	F
War and peace	12.3	M	F	F	…	M
…	…	…	…	…	…	…
…	…	…	…	…	…	…
Harry Potter I	64.1	F	F	M	…	F
	Median difference	3.9	-1.4	2.1	…	1.6

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

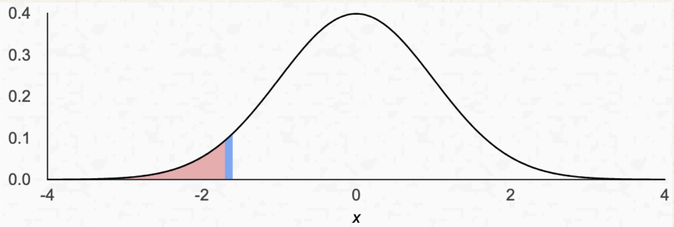


Is 3.9 unusual for this constructed null distribution?

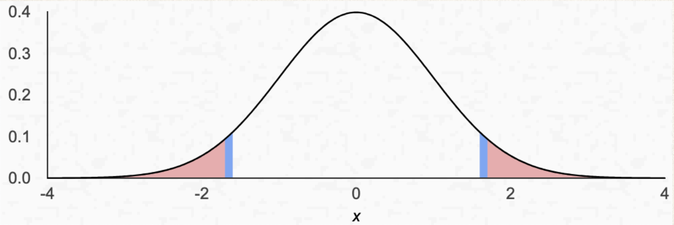
P VALUE



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



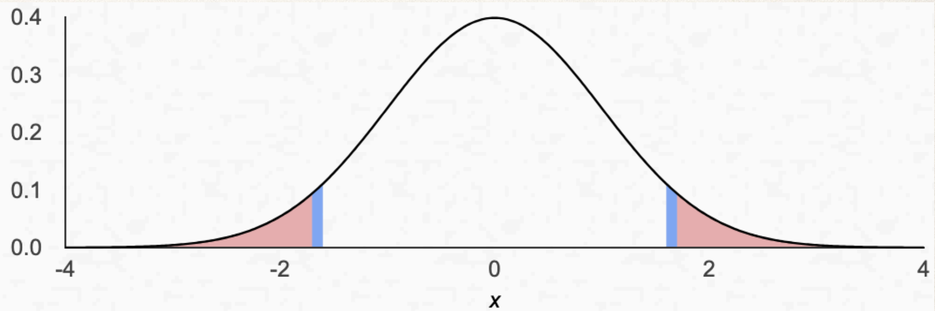
$$\text{p-value (x)} = 1 - P(X \leq x \mid H_0)$$



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

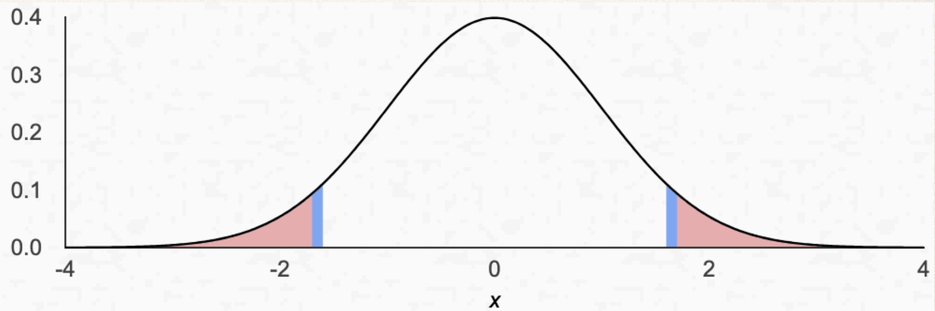
Empirical p value

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



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

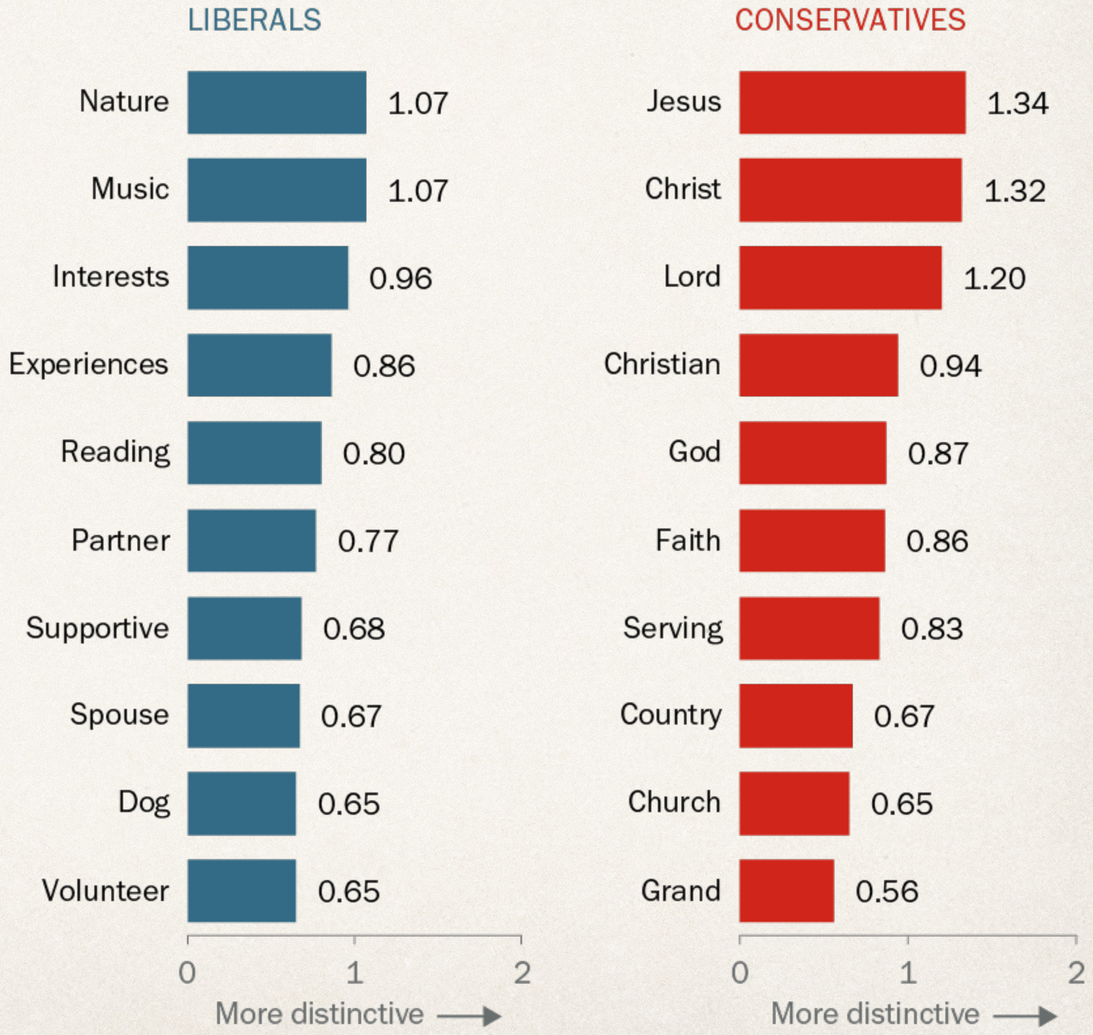
- Permutation tests have broad application
- You don’t make any assumptions about the shape and form of the null distribution for the statistic
- You can use pretty much any statistic!

Example

- How can we say PMI("Nature", *liberal*) is significantly greater than PMI("Nature", *conservative*)?

Pointwise mutual information helps identify words that are used distinctively by one group, not just used more

Most dinstinctive terms used by each group, by PMI score



Test significance of PMI

Test significance of PMI

- Let $t = \dfrac{\texttt{PMI("Nature"}, liberal)}{\texttt{PMI("Nature"}, conservative)}$

Test significance of PMI

- Let $t = \dfrac{\text{PMI}(\text{"Nature"}, liberal)}{\text{PMI}(\text{"Nature"}, conservative)}$

- Calculate t_{obs} from the observed survey responses (every response is labeled either liberal or conservative)

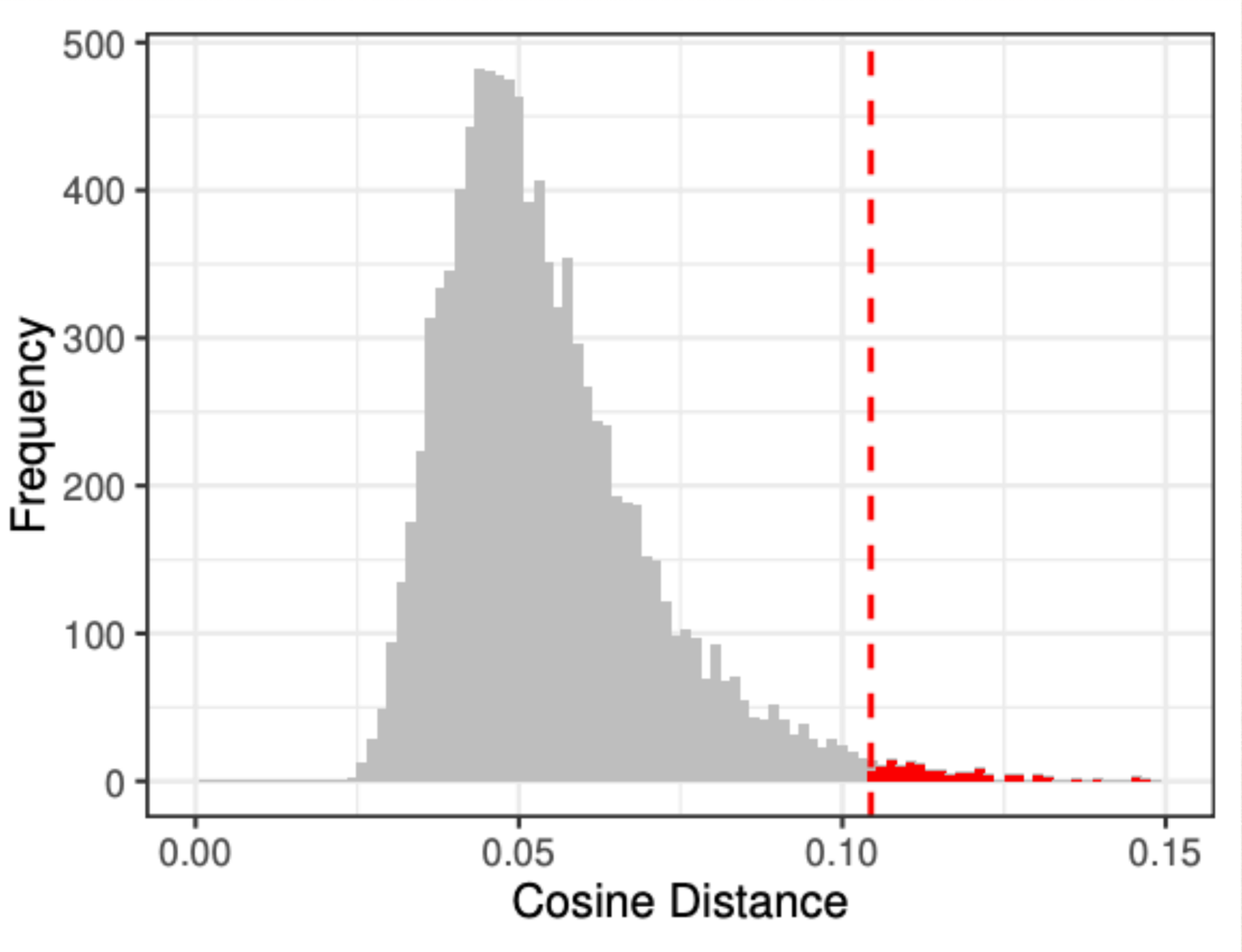
Test significance of PMI

- Let $t = \frac{\text{PMI}(\text{"Nature"}, liberal)}{\text{PMI}(\text{"Nature"}, conservative)}$
- Calculate t_{obs} from the observed survey responses (every response is labeled either liberal or conservative)
- Repeat 1 … n:
  - Shuffle the labels for all the responses
  - Recalculate t_{perm} for this permutation

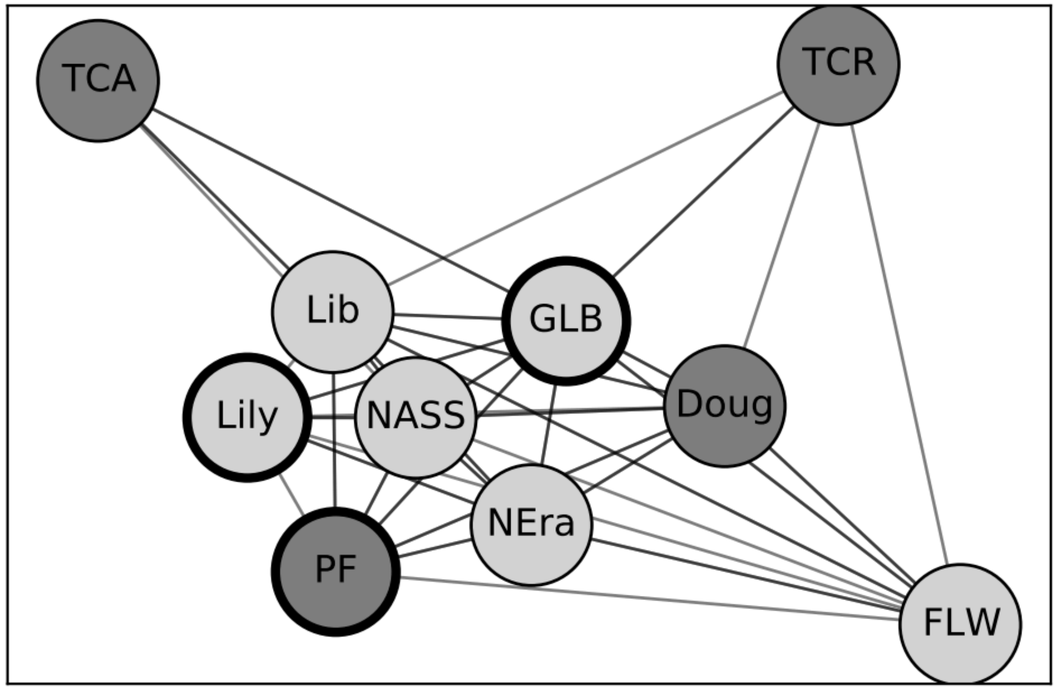
Test significance of PMI

- Let $t = \dfrac{\text{PMI}(\text{"Nature"}, \textit{liberal})}{\text{PMI}(\text{"Nature"}, \textit{conservative})}$
- Calculate t_{obs} from the observed survey responses (every response is labeled either liberal or conservative)
- Repeat 1 … n:
  - Shuffle the labels for all the responses
  - Recalculate t_{perm} for this permutation
- Calculate $p_{empirical}$ by comparing t_{obs} with the distribution of t_{perm}

- 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

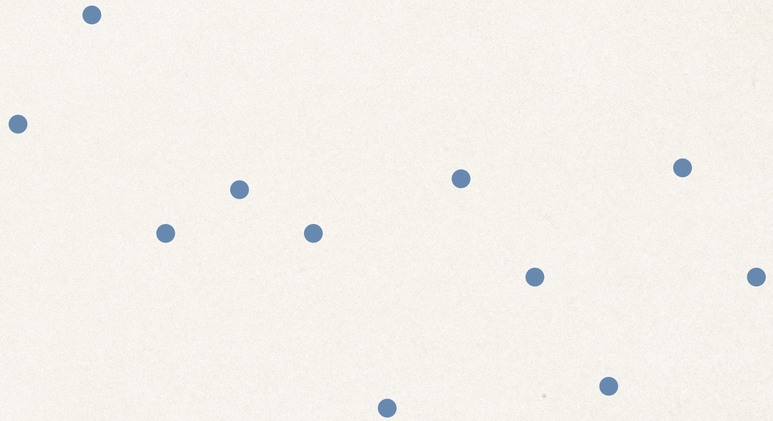
Bootstrap

- 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

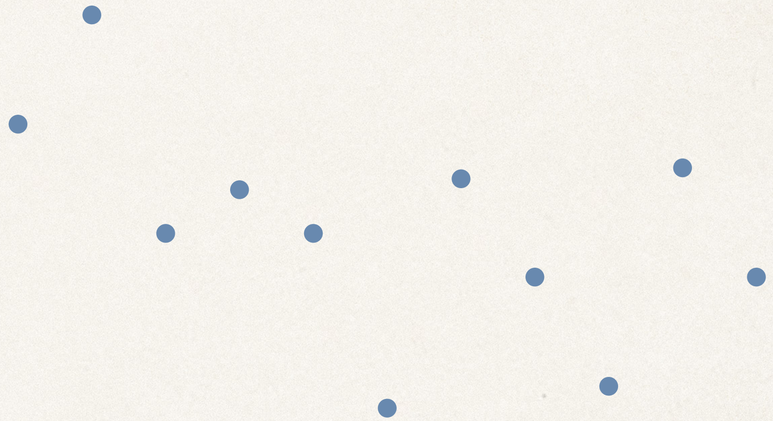
- In permutation test, the data is not changed – just the assignment of labels
- The variability in the data can be quantified by constructing hypothetical datasets that follow the same distribution
- This is the idea of bootstrapping!

Bootstrap



Bootstrap

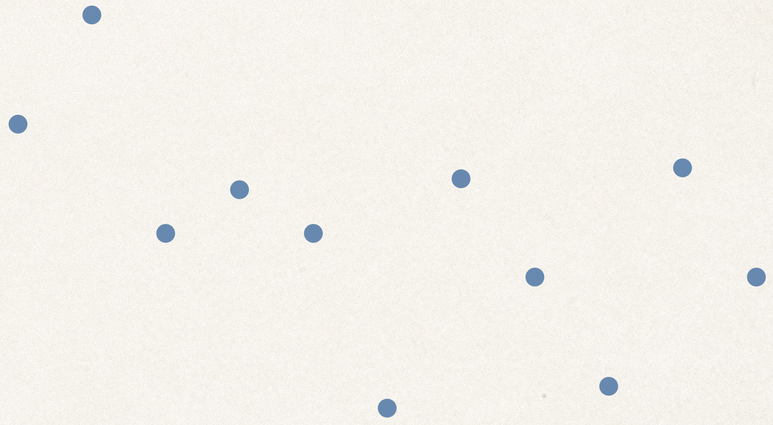
- Our sample is assumed to be iid, and represents the population distribution



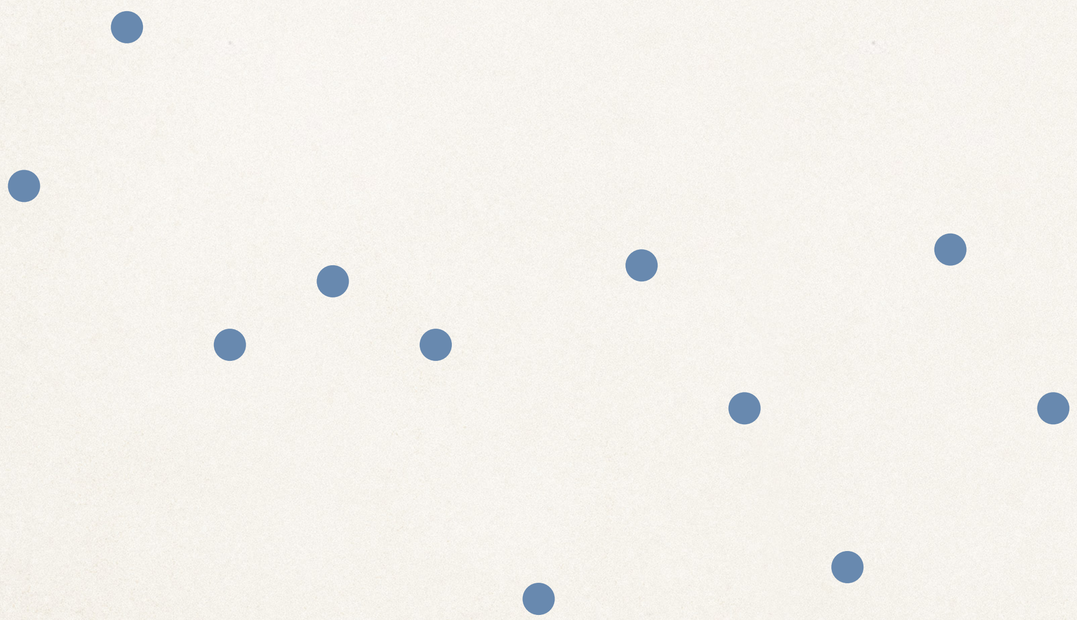
Bootstrap

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

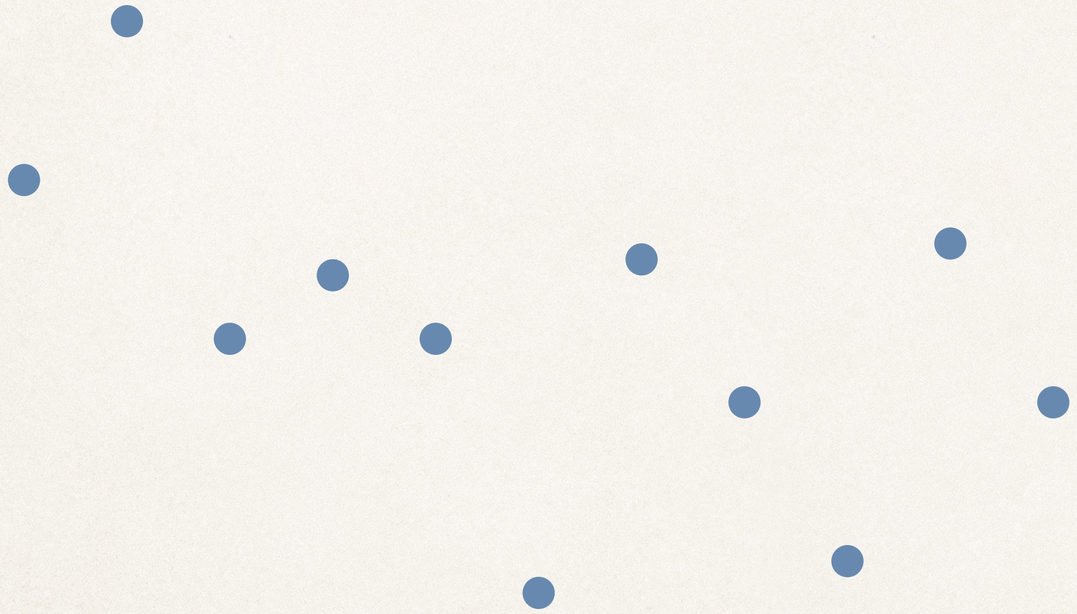
- Repeatedly sample from our sample to generate many alternative datasets



Distribution of our sample



Distribution of our sample

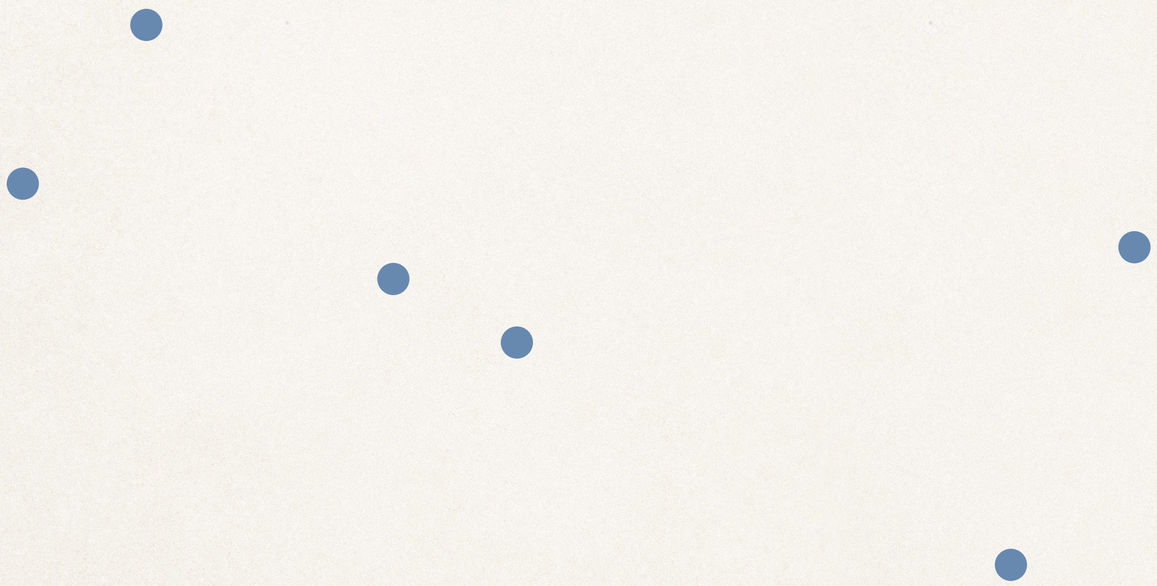


- We can calculate some metric on this sample such as accuracy

Distribution of our sample

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

Bootstrap 1



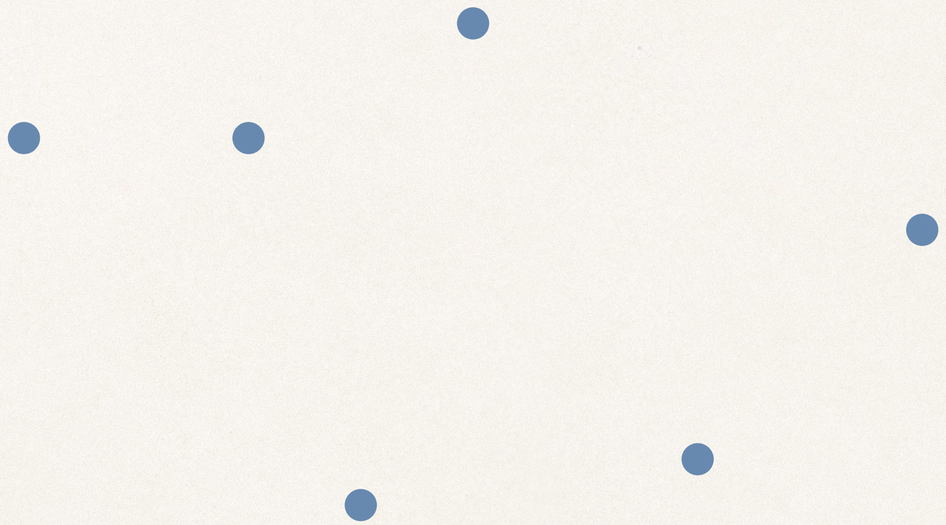
Bootstrap 1

- 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

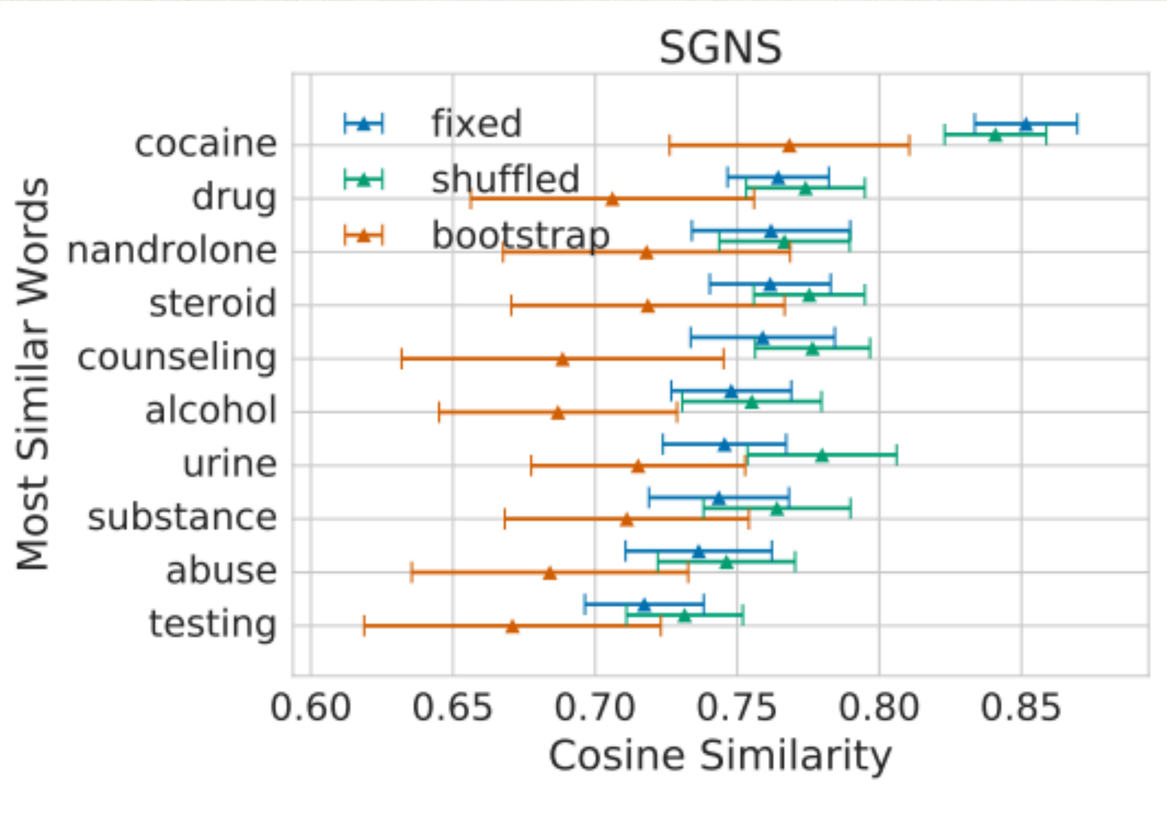
- From b boostrap samples, we can get a size b array as estimates for accuracy
- We can rank these estimates and find the middle 95% to define a range in which the accuracy estimate falls

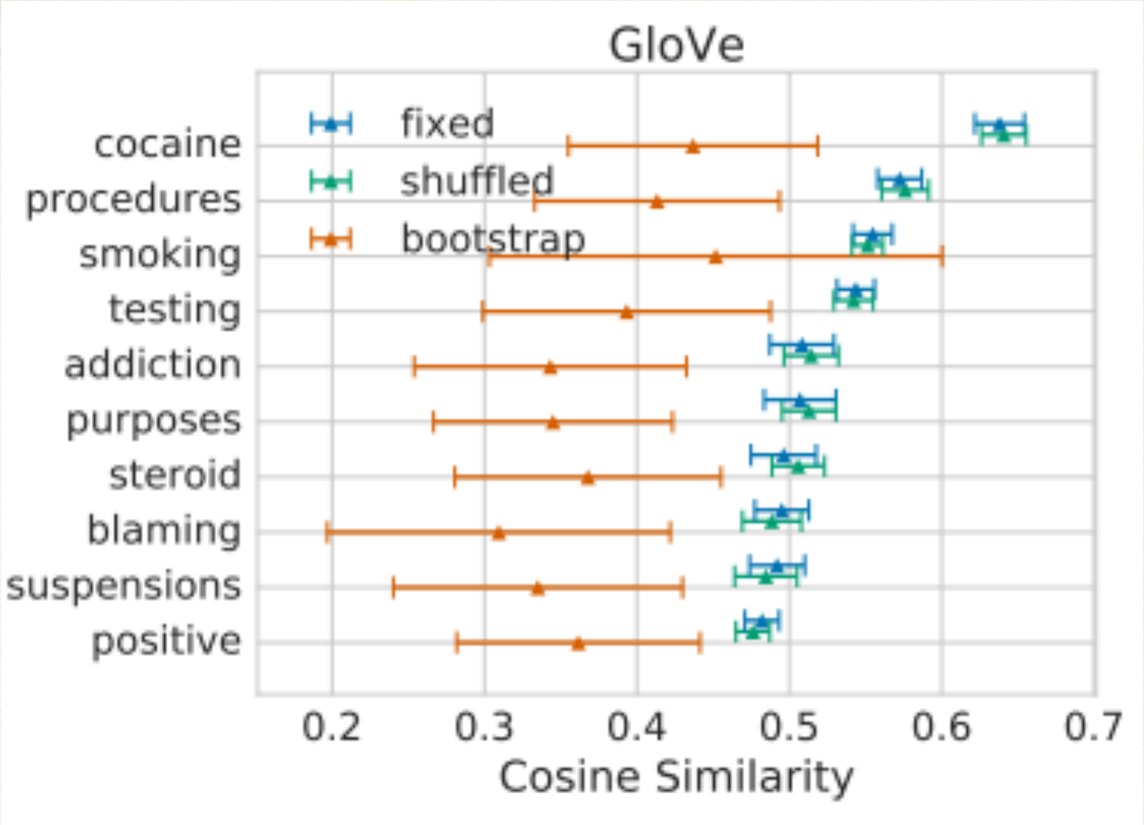
Bootstrapped Confidence intervals

- 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.

In class

- Non parametric test