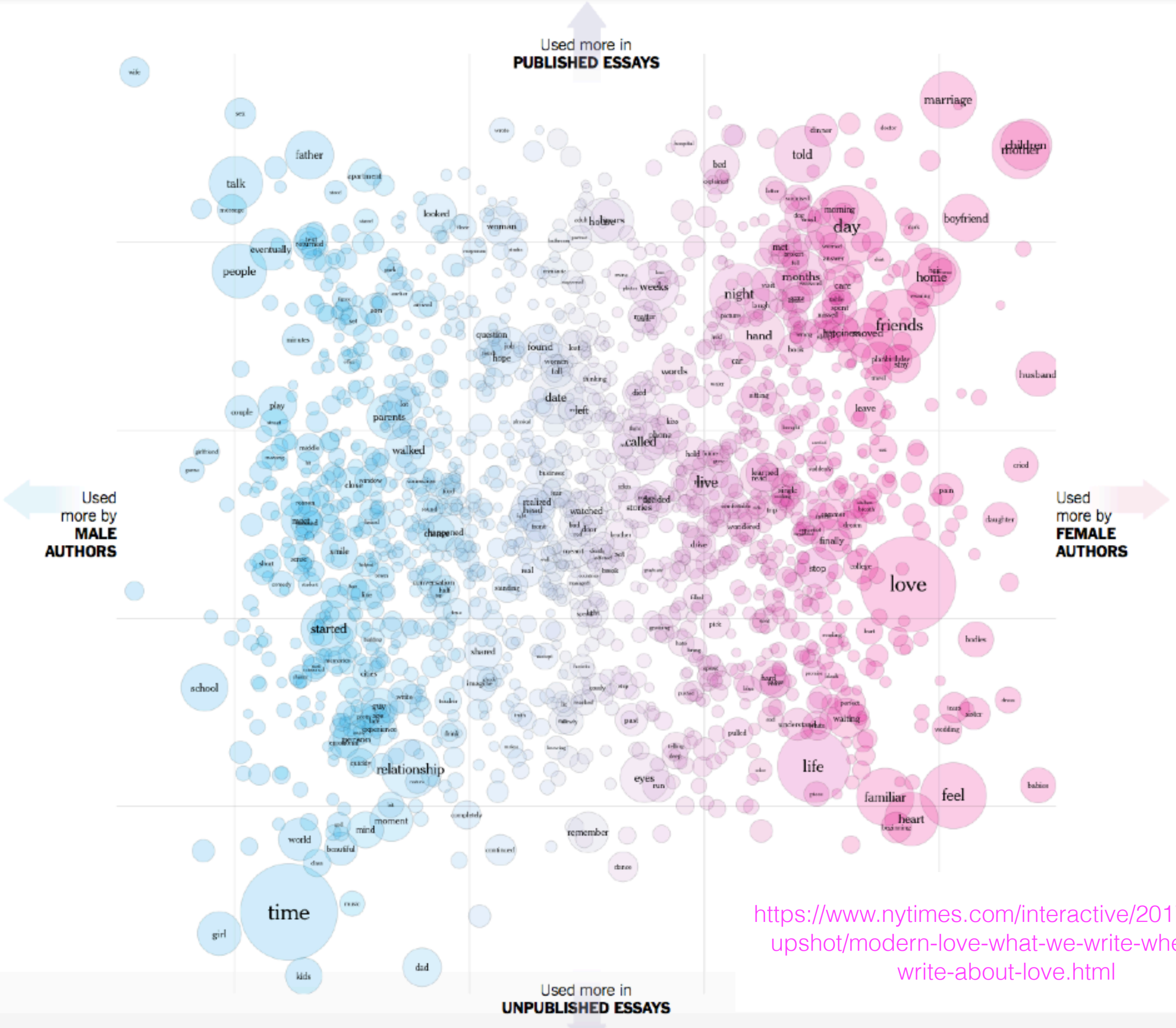
Applied Natural Language Processing

Info 256

Lecture 3: Finding Distinctive Terms (Aug 30, 2023)

David Bamman, UC Berkeley

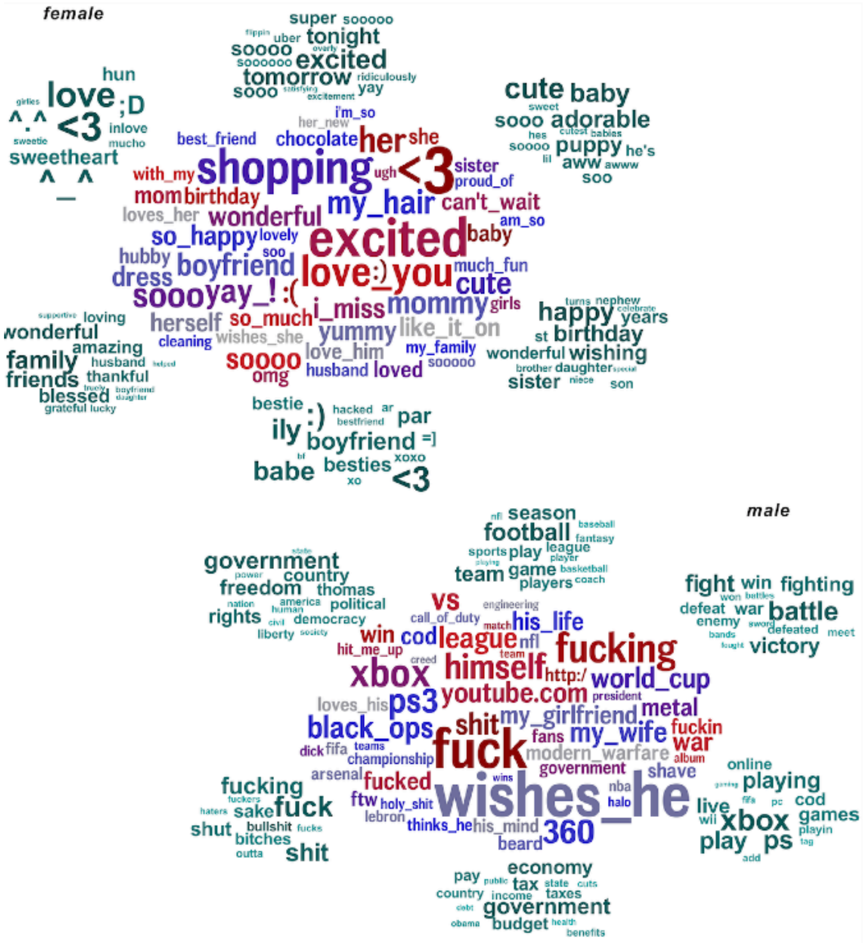


https://www.nytimes.com/interactive/2017/11/07/upshot/modern-love-what-we-write-when-we-write-about-love.html

Panel B: Phrases Used More Often by Republicans

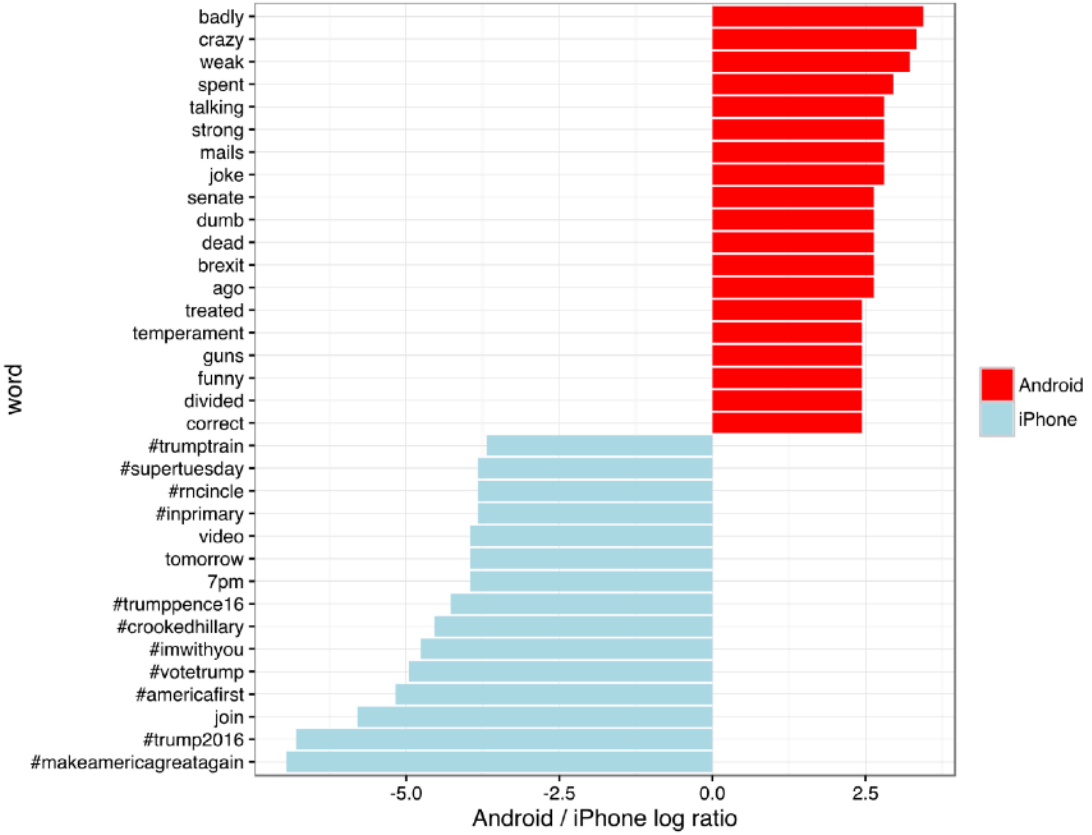
Two-Word Phrases		
stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program
Three-Word Phrases		
embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security

Gentzkow and Shapiro (2006), "What Drives Media Slant? Evidence from U.S. Daily Newspapers," Econometrica



Schwartz et al. (2013), "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach"

Which are the words most likely to be from Android and most likely from iPhone?



http://varianceexplained.org/r/trump-tweets/

Distinctive terms

- Finding distinctive terms is useful:
    - As a data exploration exercise to understand larger trends in individual word differences).
    - As a pre-processing step of feature selection.
- When the two datasets are A and ¬A, these terms also provide insight into what A is about.
- Many methods for finding these terms! (Developed in NLP, corpus linguistics, political science, etc.)

Difference in proportions

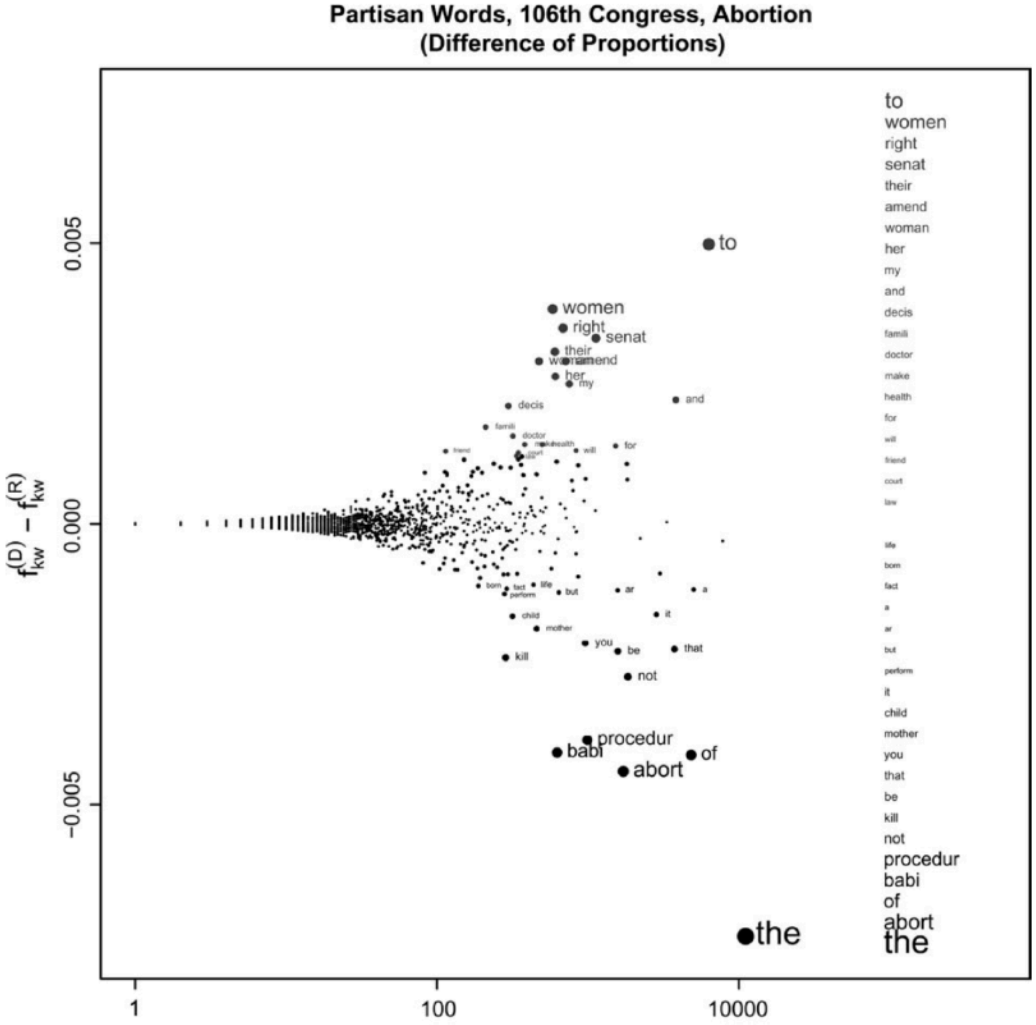
For word w written by author with label k (e.g., {democrat, republican}), define the frequency to be the normalized count of that word

$$f_{w,k} = \frac{C(w,k)}{\sum_{w'} C(w',k)}$$

count of word w in group k

count of all words in group k

$$f_{w,k=dem} - f_{w,k=repub}$$



Monroe et al. (2009), “Fightin’ Words”

Difference in proportions

- The difference in proportions is a conceptually simple measure and easily interpretable.
- Drawback: tends to emphasize words with high frequency (where even comparatively small differences in word usage between groups is amplified).
- Also, no measure whether a difference is statistically meaningful. We have uncertainty about the what the true proportion is for any group.

χ^2

- χ^2 (chi-square) is a statistical test of dependence—-here, dependence between the two variables of word identity and corpus identity.

- For assessing the difference in two datasets, this test assumes a 2x2 contingency table:

	word	¬word
corpus 1	7	104023
corpus 2	104	251093

X^2

Does the word *robot* occur significantly more frequently in science fiction?

	robot	¬robot	
sci-fi	104	1004	= 10.3%
¬sci-fi	2	13402	= 0.015%

X^2

For each cell in contingency table, sum the squared difference between observed value in cell and the expected value assuming independence.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

	robot	¬robot	sum	frequency
sci-fi	104	1004	1108	0.076
¬sci-fi	2	13402	13404	0.924
sum	106	14406		
frequency	0.007	0.993		

Assuming independence:

$$P(\text{robot}, \text{scifi}) = P(\text{robot}) \times P(\text{scifi})$$
$$= 0.007 \times 0.076 = 0.00053$$

Among 14512 words, we would expect to see 7.69 occurrences of *robot* in sci-fi texts.

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

$P(\text{scifi})$	0.076
$P(\neg\text{scifi})$	0.924

$P(\text{robot})$	$P(\neg\text{robot})$
0.007	0.993

χ^2

- What χ^2 is asking is: how different are the observed counts from the counts we would expect given complete independence?

	robot	¬robot
sci-fi	104	1004
¬sci-fi	2	13402

	robot	¬robot
sci-fi	7.69	1095.2
¬sci-fi	93.9	13315.2

X2

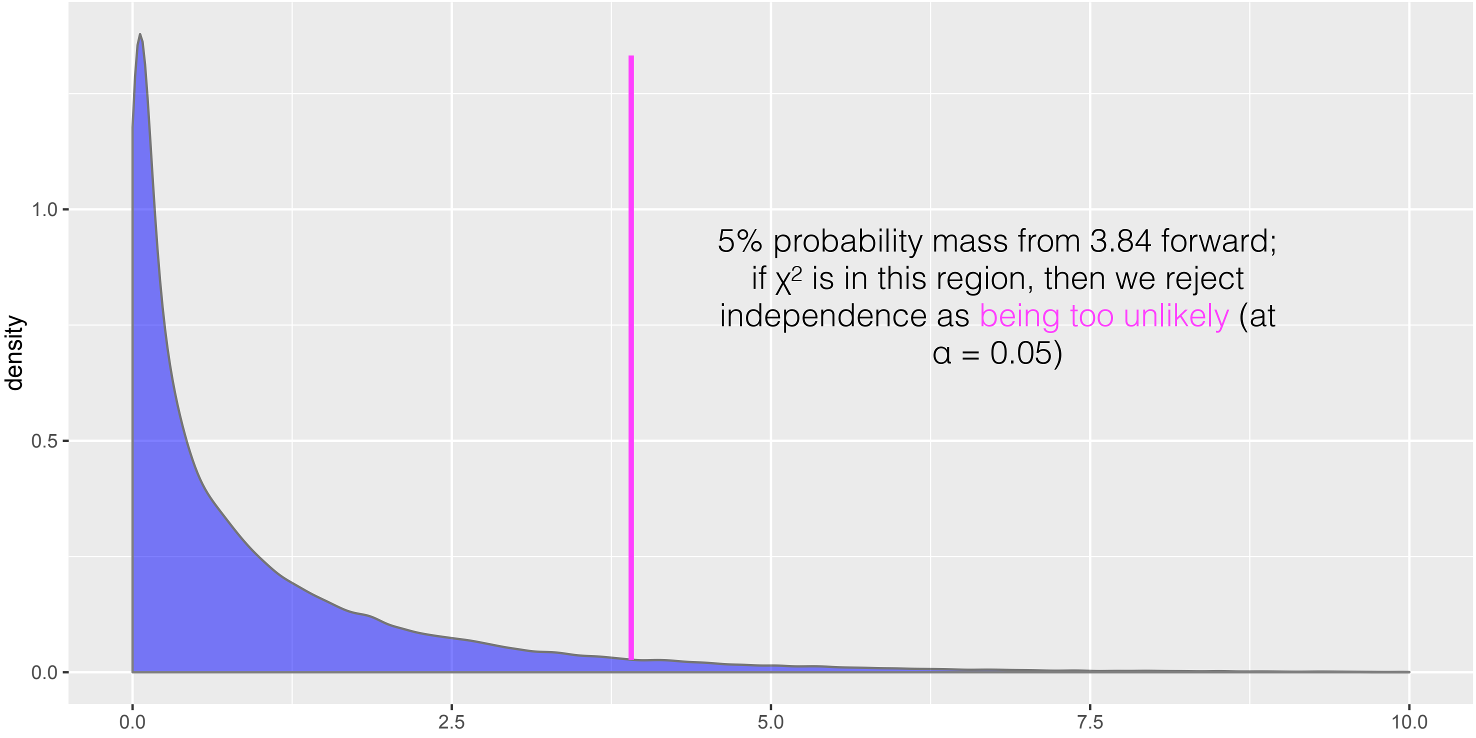
- With algebraic manipulation, simpler form for 2x2 table O (cf. Manning and Schütze 1999)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

χ^2

- The χ^2 value is a statistic of dependence with a probability governed by a χ^2 distribution; if this value has low enough probability in that measure, we can reject the null hypothesis of the independence between the two variables.

χ^2


5% probability mass from 3.84 forward;
if χ² is in this region, then we reject
independence as being too unlikely (at
α = 0.05)
density
1.0
0.5
0.0
0.0
2.5
5.0
7.5
10.0

X^2

- Chi-square is ubiquitous in corpus linguistics (and in NLP as a measure of collocations).

- A few caveats for its use:

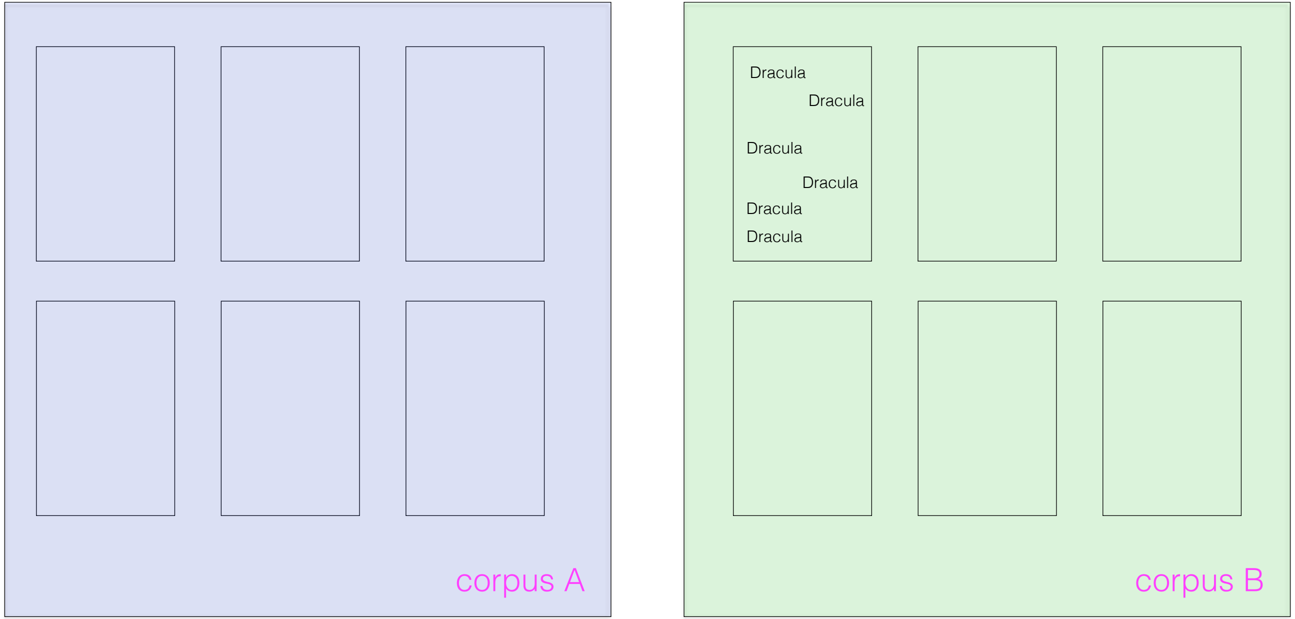
    - Each cell should have an *expected* count of at least 5

    - Each observation is independent

X^2

- A drawback, however, is due to the burstiness of language: the tendency for the same words to clump together in texts.

- Chi-square is testing for independence of two variables (word identity and corpus identity), but it assumes each mention of the word is independent from the others.



- Is Dracula really a word that distinguishes these two corpora?
- It distinguishes one text, but otherwise doesn't appear in the corpus at all.

Mann-Whitney rank sums test

- Mann-Whitney is a test of the difference in some quantity of interest in two datasets. Null hypothesis: if you select a random sample from group A and another from group B, just as likely that A will be greater than B as less than B.

A	A	A	A	A	A	A	A
1	2	1	4	3	2	0	1

B	B	B	B	B	B
8	4	9	7	6	10

Mann-Whitney

A	A	A	A	A	A	A	A
1	2	1	4	3	2	0	1

B	B	B	B	B	B	B
8	4	9	7	2	10	5

	A	A	A	A	A	A	B	A	B	B	B	B	B	B
	0	1	1	1	2	2	2	3	4	5	7	8	9	10
ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Mann-Whitney

A	A	A	A	A	A	A	A
1	2	1	4	3	2	0	1

B	B	B	B	B	B	B
8	4	9	7	2	10	5

	A	A	A	A	A	A	B	A	B	B	B	B	B	B
	0	1	1	1	2	2	2	3	4	5	7	8	9	10
ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14

$$R_1 = 7+9+10+11+12+13+14 = 76$$

Mann-Whitney

R_1 = 7+9+10+11+12+13+14 = 76

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

n_1 = sample size for dataset from which R_1 is derived (e.g., number of chunks)

- Once we have this U value, we can ask whether it's significantly different from the average value we would expect if there's no difference between the two groups at all. (We can do so by converting U to a z-score using a Normal approximation and checking significance).

A	A	A	A	A	A	A	A
1	2	1	4	3	2	0	1

B	B	B	B	B	B	B
8	4	9	7	6	10	5

- In corpus linguistics, each measurement is the count of a word in a fixed-sized chunk of text (e.g., 500 words).

- This lets us accommodate a more realistic assumption about the burstiness of language.

A	A	A	A	A	A	A	A
0	0	0	417	0	0	0	0

B	B	B	B	B	B
0	0	0	0	0	0

417 mentions of Dracula in one book

Not a significant difference in ranks

A	B	A	A	B	B	A	B	A	B	A	B	A	A
0	0	0	0	0	0	0	0	0	0	0	0	0	417
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Log-odds ratios with priors

- The odds of a word is another informative measure:

$$\frac{\text{probability of event occuring in corpus X}}{\text{probability of event not occuring in corpus X}}$$

$$\frac{\left(\frac{\text{count of word } w \text{ in corpus X}}{\text{\# words in X}}\right)}{\left(\frac{\text{count of all words} \neg w \text{ in corpus X}}{\text{\# words in X}}\right)} = \frac{\text{count of word } w \text{ in corpus X}}{\text{count of all words} \neg w \text{ in corpus X}}$$

$$\frac{\text{\# mentions of "recall"}}{\text{\# mentions of} \neg\text{"recall"}} \qquad \frac{14}{1{,}000{,}000} = 0.000014$$

Democrat
n=1000014

Log-odds ratios with priors

- Two get a measure of difference, we can compare two odds in different corpora

Democrat n=1000014	Republican n=2000042
$\frac{14}{1,000,000} = 0.000014$	$\frac{42}{2,000,000} = 0.000021$

$$\frac{\text{\# mentions of "recall"}}{\text{\# mentions of} \neg \text{"recall"}}$$

- The odds ratio gives us one way of combining these into a single score

$$\frac{\frac{14}{1,000,000}}{\frac{42}{2,000,000}} = 0.667$$

Log-odds ratios with priors

- But this is bounded by (0, ∞) and not easy to interpret with respect to the boundary (1) separating a word being likelier in corpus than another.

$$\frac{\frac{14}{1,000,000}}{\frac{42}{2,000,000}} = 0.667$$

- We can work with the log instead, which transforms this into the space (-∞, ∞), with 0 as a boundary

$$\log\left(\frac{\frac{14}{1,000,000}}{\frac{42}{2,000,000}}\right) = -0.4054$$

Log-odds ratios with priors

- What if we have 0 counts?

$$\log\left(\frac{\frac{14}{1{,}000{,}000}}{\frac{0}{2{,}000{,}000}}\right) = 🔥$$

- We can add pseudocounts! e.g., assume vocabulary size of 10,000 words, 100 here = 10,000 * 0.01 to account for total pseudocount mass added, and we remove 0.01 from the denominators since the denominator is the count of ¬word.

$$\log\left(\frac{\frac{14+0.01}{1{,}000{,}000+100-0.01}}{\frac{0+0.01}{2{,}000{,}000+100-0.01}}\right) = 7.94$$

Log-odds ratios with priors

$$\log\left(\frac{\frac{14+0.01}{1{,}000{,}000+100-0.01}}{\frac{42+0.01}{2{,}000{,}000+100-0.01}}\right) = -0.4050$$

$$= \log\left(\frac{14+0.01}{1{,}000{,}000+100-0.01}\right) - \log\left(\frac{42+0.01}{2{,}000{,}000+100-0.01}\right)$$

How confident are we about these estimates?

Log-odds ratios with priors

- Transform them into z-scores by dividing them by the standard deviation.

$$\approx \frac{\log\left(\frac{14+0.01}{1{,}000{,}000+100-0.01}\right)-\log\left(\frac{42+0.01}{2{,}000{,}000+100-0.01}\right)}{\sqrt{\frac{1}{14+0.01}+\frac{1}{42+0.01}}}$$

The larger the term counts (e.g., 14, 42), the more confident we can be that the difference is meaningful

Other methods

- There are many other methods for learning distinguishing words between two corpus; major classes:
    - Model-based methods that assume parametric forms + Bayesian priors (for smoothing) [Monroe et al. 2009]
    - Methods using classification to learn informative features that separate classes.

Activity

- Hypothesize terms that will be different between 2020 Democrat and Republican platforms.
- Execute chi-square to find terms that are different
- Compare to Mann-Whitney for this data.