

# A/B Testing

## A

## B



CONTROL



VARIATION

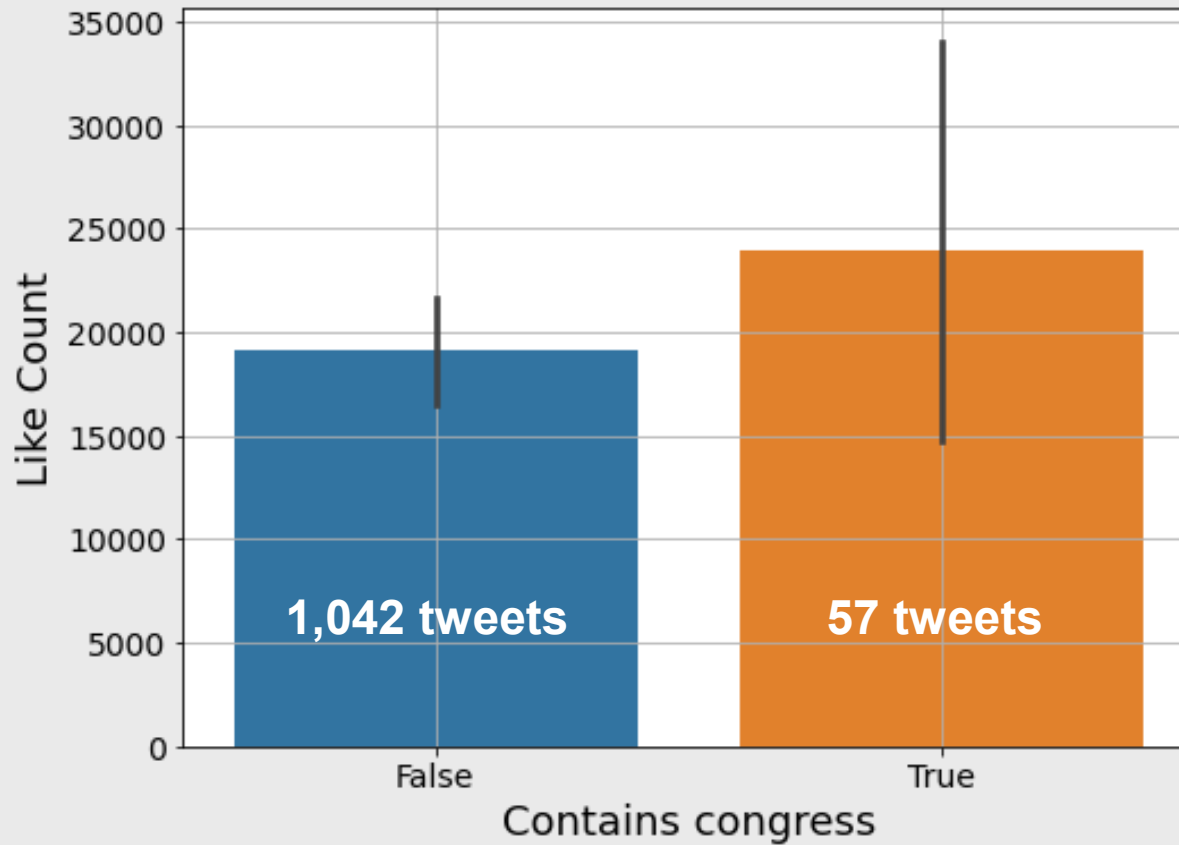
# Comparing Tweets

- **Tweet 1: 2,290 likes**, If we only get 2 recon bills per year, and BBB was supposed to be the 2nd recon of 2021, does pushing “roll it over” or does the Senate clock restart in 22, BBB is erased as the 2nd 2021 bill, & Dems now only have 2 swings left instead of 3? Likely the latter but not confirmed
- **Tweet 2: 52,388 likes**, There is no reason members of Congress should hold and trade individual stock when we write major policy and have access to sensitive information. There are many ways members can invest w/o creating actual or appeared conflict of interest, like thrift savings plans or index funds

# Comparing Tweets

- **Tweet 1: 2,290 likes**, If we only get 2 recon bills per year, and BBB was supposed to be the 2nd recon of 2021, does pushing “roll it over” or does the Senate clock restart in 22, BBB is erased as the 2nd 2021 bill, & Dems now only have 2 swings left instead of 3? Likely the latter but not confirmed
- **Tweet 2: 52,388 likes**, There is no reason members of **Congress** should hold and trade individual stock when we write major policy and have access to sensitive information. There are many ways members can invest w/o creating actual or appeared conflict of interest, like thrift savings plans or index funds

# Comparing More Tweets

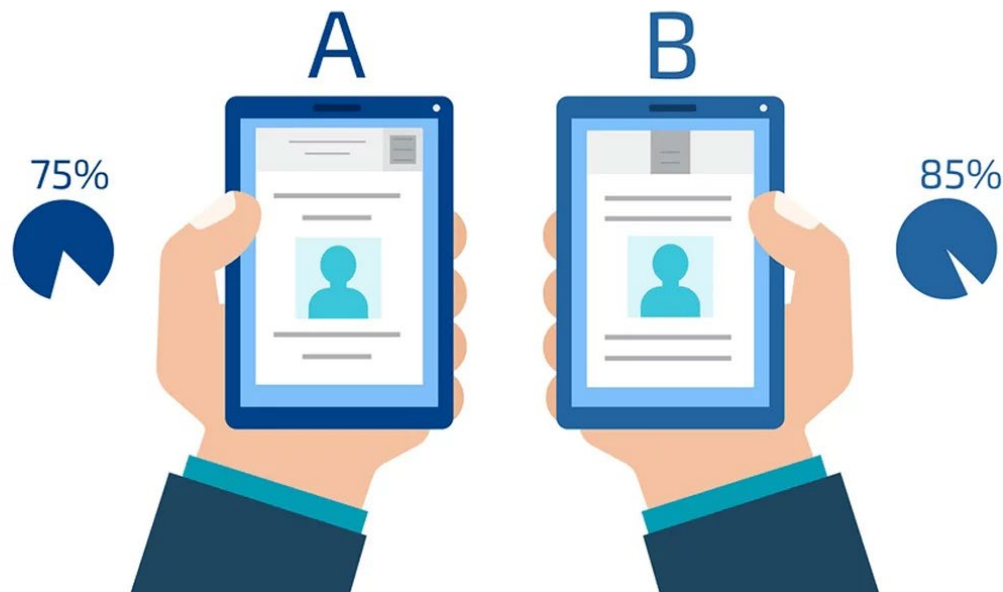


# Comparing Tweets

- Does the word “Congress” increase a tweet’s like count?
- We can use **A/B testing** to answer this question quantitatively

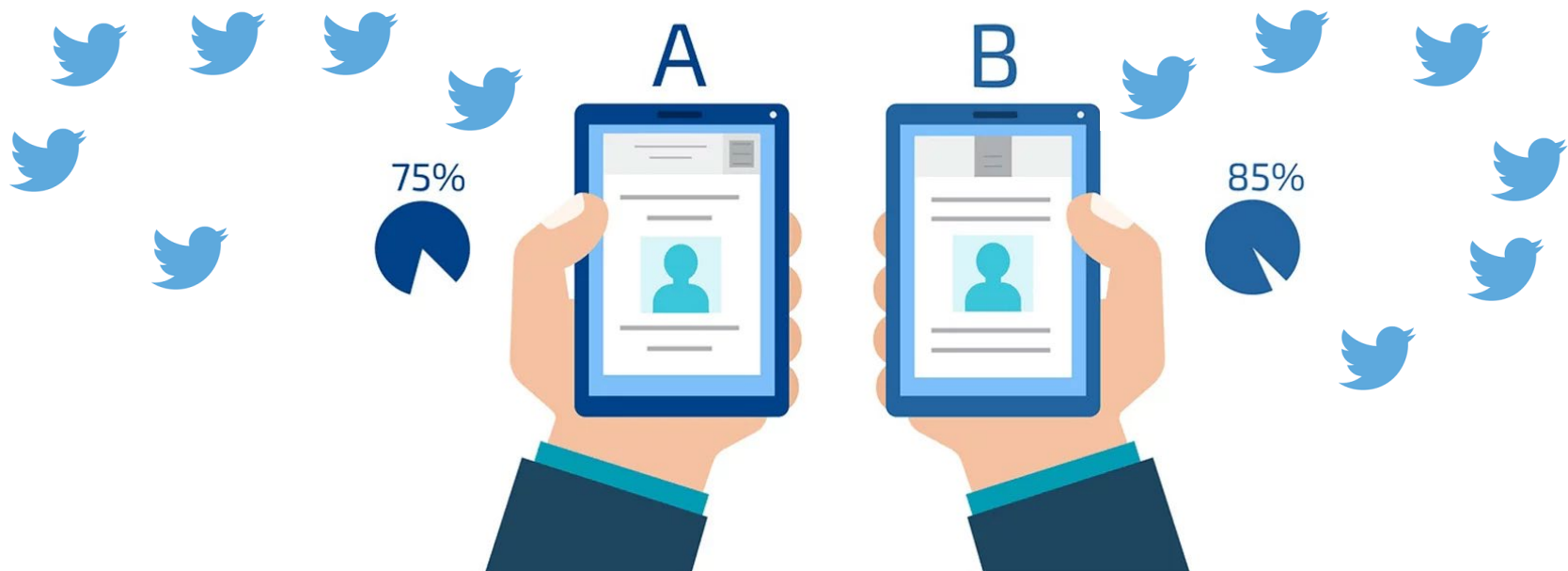
# A/B Testing

- Compare two versions of something to see which is better (version A vs version B)



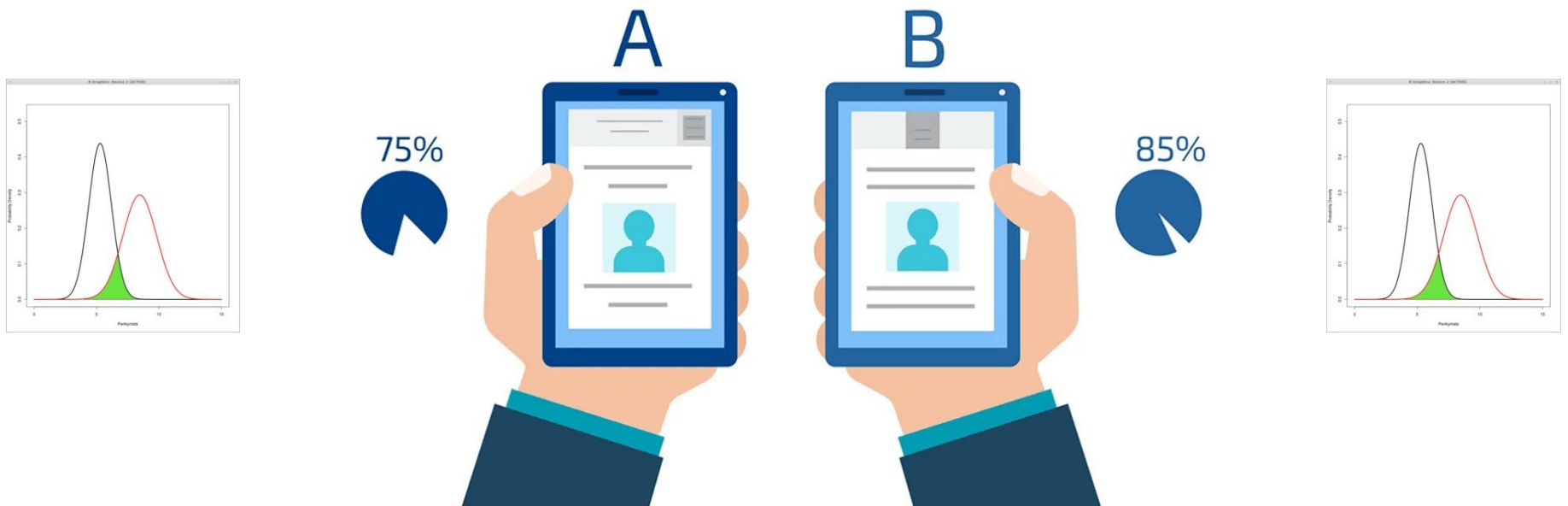
# A/B Testing

- Run an experiment to generate data for each version



# A/B Testing

- Apply a **statistical test** to data to determine which version is better





# A/B Testing in Social Media

- We can A/B test for many features' effect on many responses

Features
Tweet contains a keyword
Image contains a feature (color, animal, location, etc)
Tweet/image belongs to a cluster
Tweet sentiment
User
Time of day, day of week

Responses
Engagement (likes, retweets, etc.)
Tweet sentiment
Tweet contains a keyword
Image contains a feature

# Statistical Tests

- **Z-test**
- **T-test**
- **Fisher exact test**

# Null Hypothesis

- Statistical tests check if the “null hypothesis” is true
- **Null hypothesis** – there is no difference in the versions (i.e. no effect)
- We will denote the null hypothesis as **H0**

# Test Statistic

- Each statistical test is based on a test statistic
- **Test statistic** – a function of the data whose distribution is known under the null hypothesis
- If the test statistic is too big (or too small), we reject the null hypothesis

# P-Value

- How big is too big for a test statistic under  $H_0$ ?
- **P-value** – probability of the test statistic being greater than the observed value under  $H_0$
- Small p-value means it is unlikely you would see what you see under  $H_0$ 
  - So you can reject the null hypothesis

# Statistical Tests

- **Z-test**
- **T-test**
- **Fisher exact test**

# Z-Test Assumptions

- Assume **we know the underlying st. dev.** for version A and B
- $H_0$ : version A and version B have the **same mean**

# Z-Test

- **Number of samples:**  $n_A, n_B$
- **Underlying st. devs.:**  $\sigma_A, \sigma_B$
- **Sample means:**  $\hat{\mu}_A, \hat{\mu}_B$

- **Z-statistic:** 
$$z = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

- **Z-statistic is normally distributed with mean 0 and st. dev. 1**

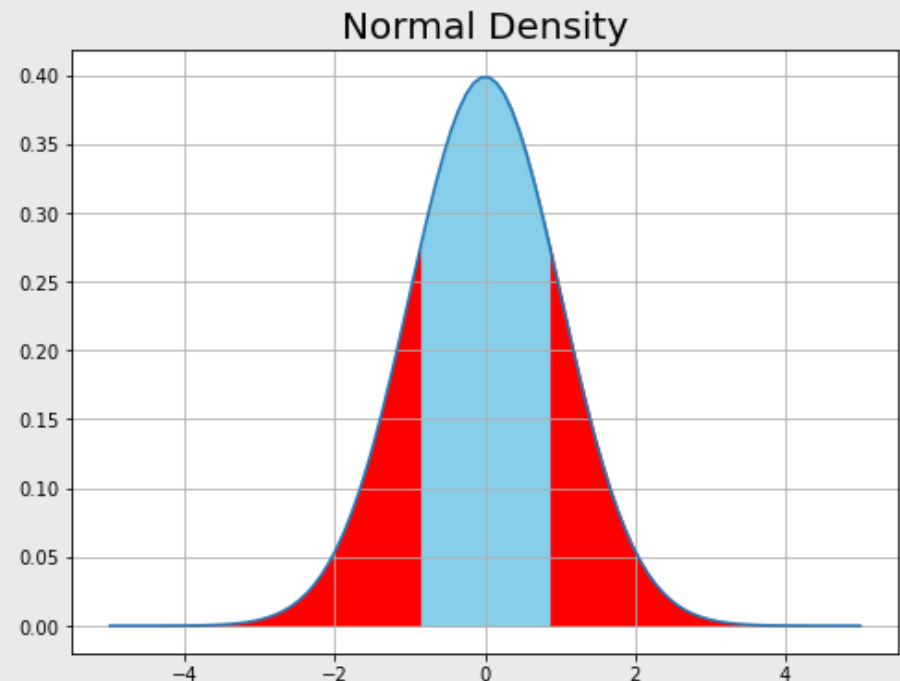
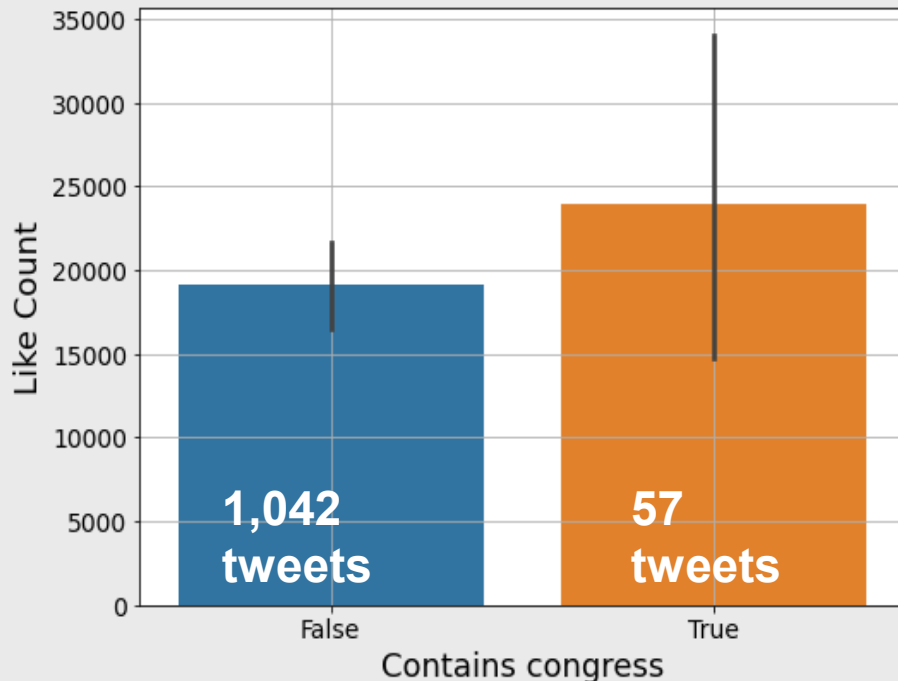


# Z-Test Limitations

- In practice Z-test is limited because usually the **underlying standard deviations are not known** 😞
- Other tests exist which have less restrictive assumptions 😊

# Example: Z-Test

- Calculate z-statistic: **z-stat = -0.970**
- P-value from normal density: **p-value = 0.332**
- Can't reject  $H_0$  at 1% level



# Statistical Tests

- Z-test
- T-test
- Fisher exact test

# T-Test Assumptions

- Assume we **don't know the underlying st. dev.** of version A and B
- $H_0$ : version A and version B have the **same mean**

# T-Test

- **T statistic:**

$$t = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$$

# T-Test

- **T statistic:**

$$t = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$$

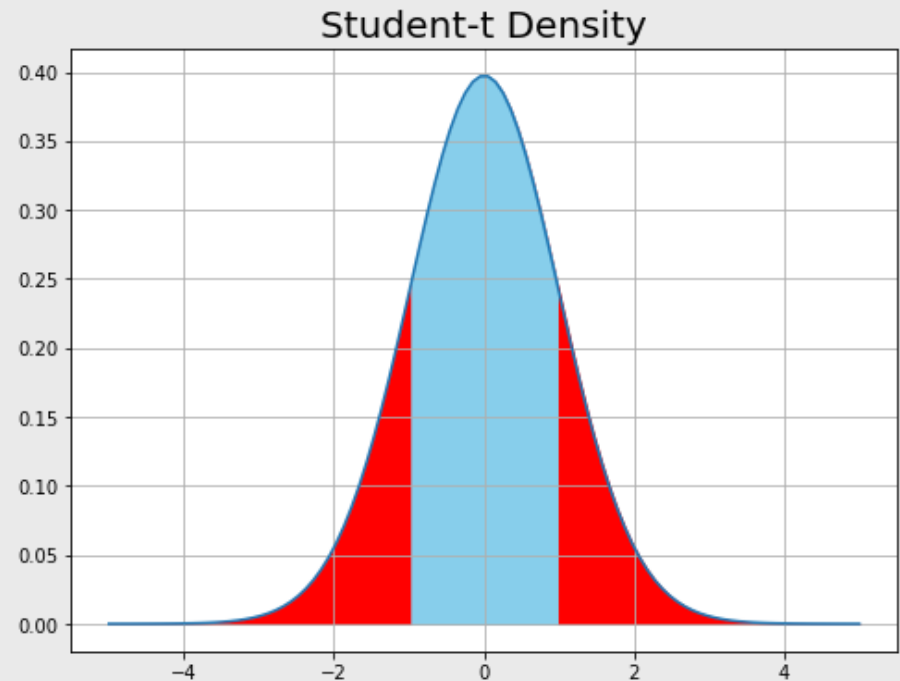
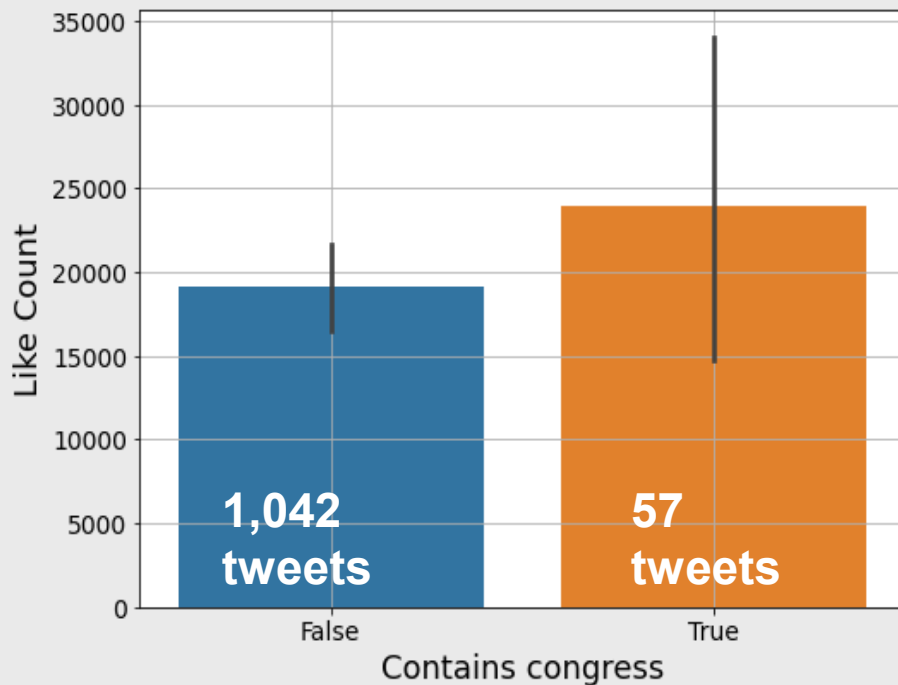
- **T-statistic follows a Student-t distribution with d.f. degrees of freedom**

- **Degrees of freedom:**

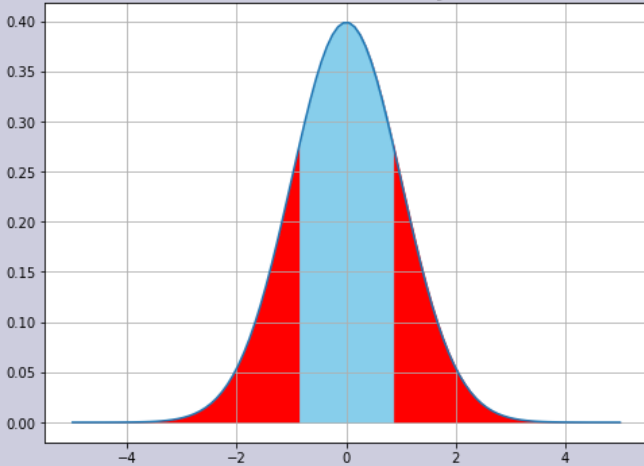
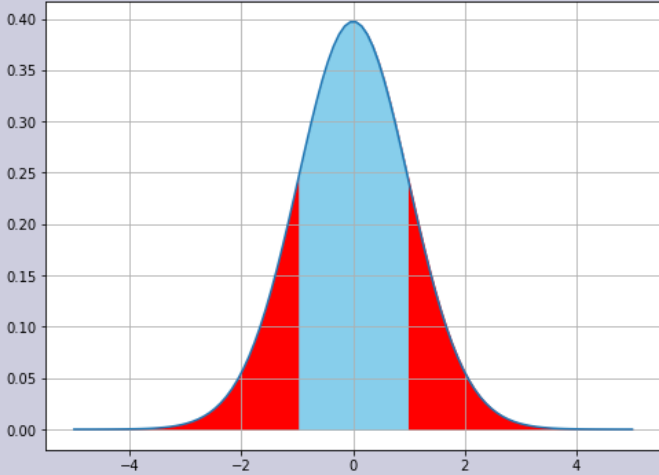
$$d.f. = \frac{\left(\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}\right)^2}{\frac{\left(\frac{\hat{\sigma}_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{\hat{\sigma}_B^2}{n_B}\right)^2}{n_B - 1}}$$

# Example: T-Test

- Calculate t-statistic: **t-stat = -0.970**
- P-value from student-t density: **p-value = 0.336**
- Can't reject H0 at 1% level



# Z-Test vs T-Test

	Z-Test	T-Test
Test	Difference in mean	Difference in mean
Assumption	Known st. dev. – z-stat has normal distribution	Unknown st. dev. – t-stat has Student-t distribution
Statistic	-0.970	-0.970
P-value	0.332	0.336
Density	<p>Normal Density</p> 	<p>Student-t Density</p> 



# Statistical Tests

- Z-test
- T-test
- Fisher exact test

# Fisher Exact Test Assumptions

- Data is in the form of 2 x 2 contingency table

	Tweets of A	Tweets of B	Total
Contains "X"	a	b	a+b
Does not contain "X"	c	d	c+d
	a+c	b+d	n =a+b+c+d

- H0: version A and version B have the **same frequencies in contingency table**

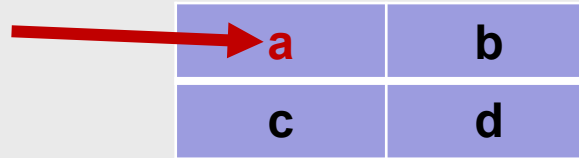
# Example: Fisher Exact Test

- $p_A$  = frequency of “Congress” in AOC tweets =  $57/(57+1042) = 0.052$
- $p_B$  = frequency of “Congress” in ElonMusk tweets =  $0/(0+100) = 0$
- $H_0: p_A = p_B$

	Tweets of AOC	Tweets of ElonMusk
Contains “Congress”	57	0
Does not contain “Congress”	1042	100

# Fisher Exact Test

- Test statistic:  $a$



<b>a</b>	b
c	d

- Test statistic follows a **hypergeometric distribution**
- To get p-value, add up probabilities of tables with same row and column sums, but more extreme differences in groups

# Example: Fisher Exact Test

- P-value from hypergeometric density:

$$p\text{-val} = 0.012$$

- Can't reject  $H_0$  at 1% level

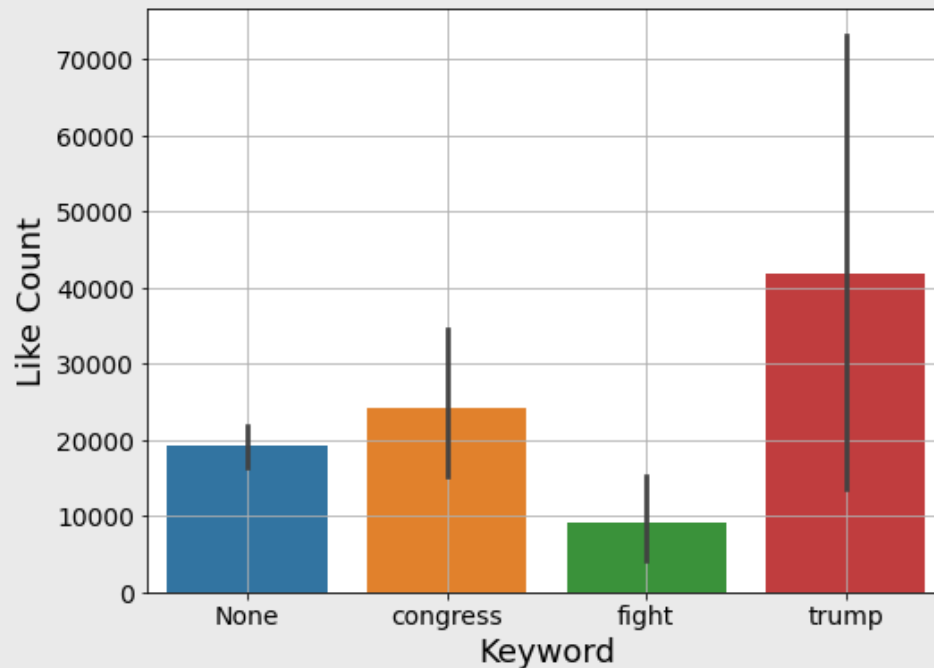
	Tweets of AOC	Tweets of ElonMusk
Contains "Congress"	57	0
Does not contain "Congress"	1042	100

# Other Tests

Test Name	Null Hypothesis	Assumptions
Mann-Whitney U Test	$P(A > B) = 0.5$	
Wilcoxon Signed Rank Test	$\text{median}(A) = \text{median}(B)$	Paired observations
Pearson's Chi-Squared Test	$\text{distribution}(A) = \text{distribution}(B)$	Data is categorical
Kolmogorov–Smirnov test	$\text{distribution}(A) = \text{distribution}(B)$	

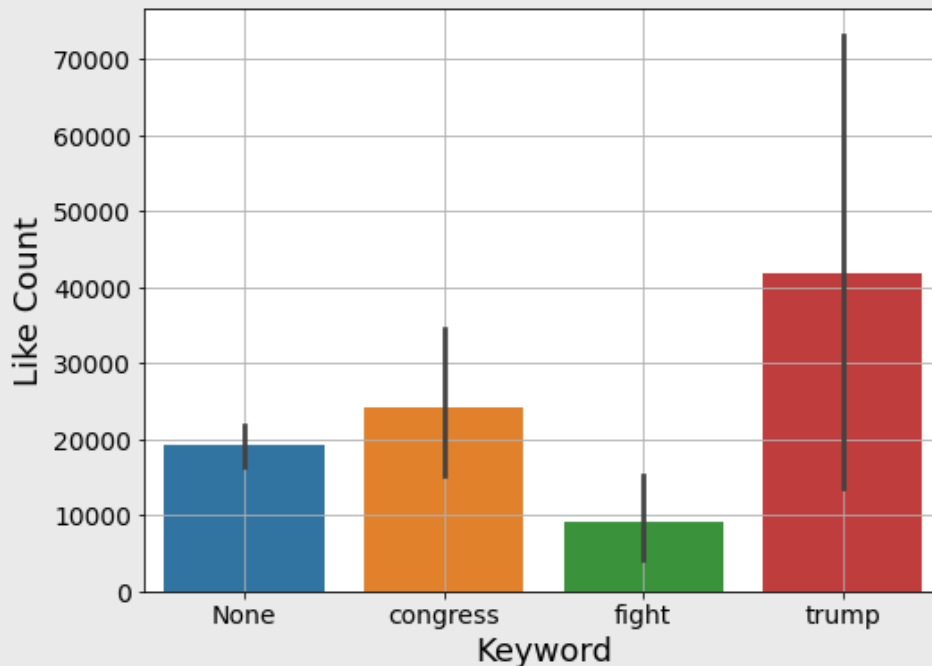
# Multiple Hypotheses

- Let's compare the like count for tweets containing a keyword, but for multiple keywords
- We are checking multiple hypotheses: one for each keyword (keyword vs no keyword)



# Multiple Hypotheses

- We can do a t-test for each keyword
- Which keywords show significance at 1%?
- Is there a problem here?



Keyword	p-value
Congress	0.3356
Trump	0.1950
Fight	0.0033



# Multiple Hypotheses

- Assume  $H_0$  is true for each keyword (keyword does not change like count)
- How many keywords will have p-values significant at a 1% level?

# Multiple Hypotheses

- Assume  $H_0$  is true for each keyword (keyword does not change like count)
- How many keywords will have p-values significant at a 1% level?

**Answer = 1%!**

# P-Hacking

- When testing multiple hypotheses, by luck you will get some very small p-values, even if there is no effect
- You will mistakenly think some things are significant - False positives ☹️
- How do we fix this problem?

# Family-Wise Error Rate

- **Family-wise error rate (FWER)** = the probability of making one or more false positives when testing multiple hypotheses
- When we say a 1% significance, we mean the FWER is 1%
- This means the threshold for each test must be **LOWER** than 1%

# Bonferroni Correction

- Goal: test **m** hypotheses at FWER  **$\alpha$**
- Each hypothesis needs a p-value **lower than  $\alpha/m$**
- Bonferroni correction is very simple, and very conservative
  - Might miss some true positives

# Bonferroni Correction

1. Any hypothesis where p-value is less than fixed threshold rejects null hypothesis

$$\alpha = 0.01, m = 5$$

Hypothesis	P-value	P-value threshold	Decision
1	0.0001	0.0020	Reject H0
2	0.0010	0.0020	Reject H0
3	0.0030	0.0020	Don't reject H0
4	0.0060	0.0020	Don't reject H0
5	0.0090	0.0020	Don't reject H0

# Sidak Correction

- Goal: test **m** hypotheses at FWER  **$\alpha$**
- Each hypothesis needs a p-value **lower than  $1-(1-\alpha)^{1/m}$** 
  - This threshold is chosen so the probability that at least one null hypothesis is rejected =  **$\alpha$**
- Sidak correction is similar to Bonferroni correction

# Sidak Correction

1. Any hypothesis where p-value is less than fixed threshold rejects null hypothesis

$$\alpha = 0.01, m = 5$$

Hypothesis	P-value	Bonferroni P-value threshold	Sidak P-value threshold
1	0.0001	0.002000	0.002008
2	0.0010	0.002000	0.002008
3	0.0030	0.002000	0.002008
4	0.0060	0.002000	0.002008
5	0.0090	0.002000	0.002008



# **Issues with Fixed Threshold Methods**

- **If we use a fixed threshold for all p-values, we are being unfair to the smaller p-values**
- **We don't need a super low threshold for all p-values**
- **Solution: use a sequential procedure where the threshold changes for larger p-values**

# Holm–Bonferroni Method

1. Rank p-values in increasing order:

$$p_1 \leq p_2 \leq \dots p_{m-1} \leq p_m$$

2. Choose your acceptable FWER:

$$\text{FWER} = \alpha$$

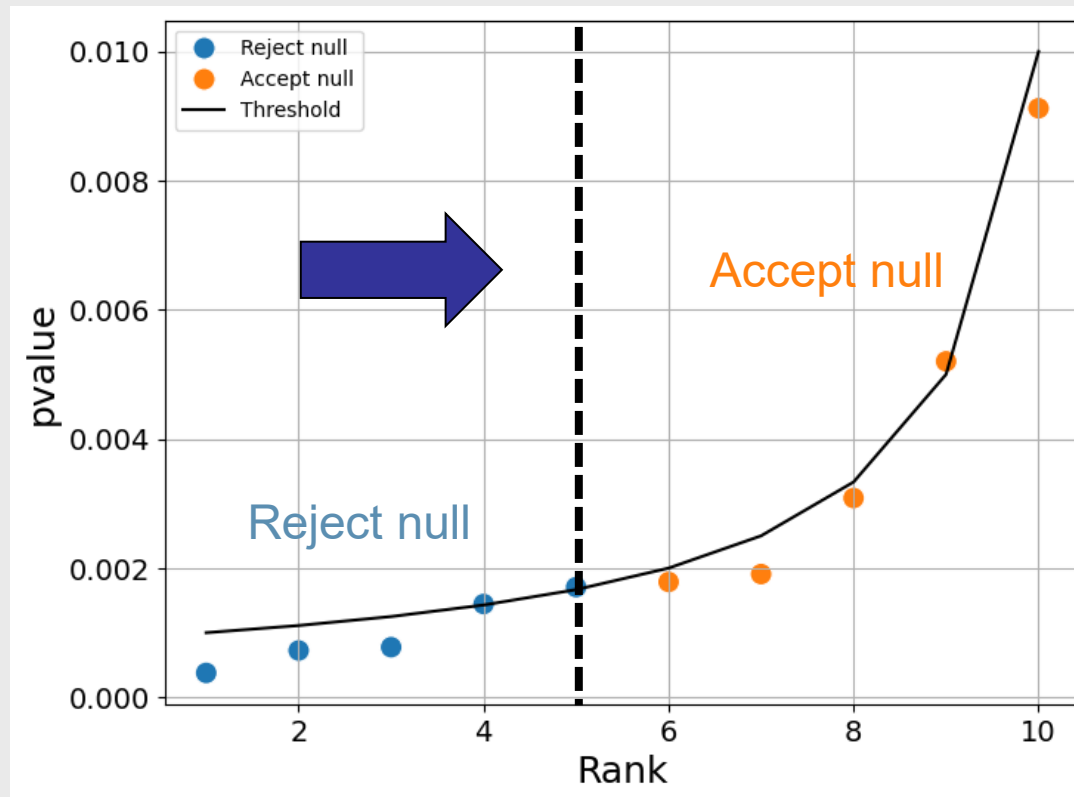
3. Threshold for p-value k:

$$\frac{\alpha}{m - k + 1}$$

4. Find the smallest p-value that is greater than its threshold
5. Reject null for all smaller p-values
6. Accept null for larger p-values (even if their p-value is less than their threshold)

# Holm-Bonferroni Method

- Start at smallest p-value and go up until a p-value is above the threshold line
- All larger p-values are not significant (even if below their threshold)



# **FWER vs FDR**

# FWER vs FDR

- **Family-Wise Error Rate (FWER)**
  - Probability of at least one false positive (reject null erroneously)
  - Very strict criterion
  - Bonferroni, Bonferroni-Holm, and Sidak control FWER

# **FWER vs FDR**

- **Family-Wise Error Rate (FWER)**
  - Probability of at least one false positive (reject null erroneously)
  - Very strict criterion
  - Bonferroni, Bonferroni-Holm, and Sidak control FWER
- **False discovery rate (FDR)**
  - Fraction of hypotheses that are false positives
  - FDR is more attractive than FWER when we care more about not missing a true discovery
  - Motivated by fields like genomics where you test 10,000 different hypotheses at once

# FWER vs FDR

- **Family-Wise Error Rate (FWER)**
  - Probability of at least one false positive (reject null erroneously)
  - Very strict criterion
  - Bonferroni, Bonferroni-Holm, and Sidak control FWER
- **False discovery rate (FDR)**
  - Fraction of hypotheses that are false positives
  - FDR is more attractive than FWER when we care more about not missing a true discovery
  - Motivated by fields like genomics where you test 10,000 different hypotheses at once
- **Benjamini-Hochberg procedure controls FDR instead of FWER**
  - More useful if you have a lot of hypotheses

# Benjamini-Hochberg Procedure

1. Rank p-values in increasing order:

$$p_1 \leq p_2 \leq \dots p_{m-1} \leq p_m$$

2. Choose your acceptable FDR:

$$\text{FDR} = \alpha$$

3. Threshold for p-value k:

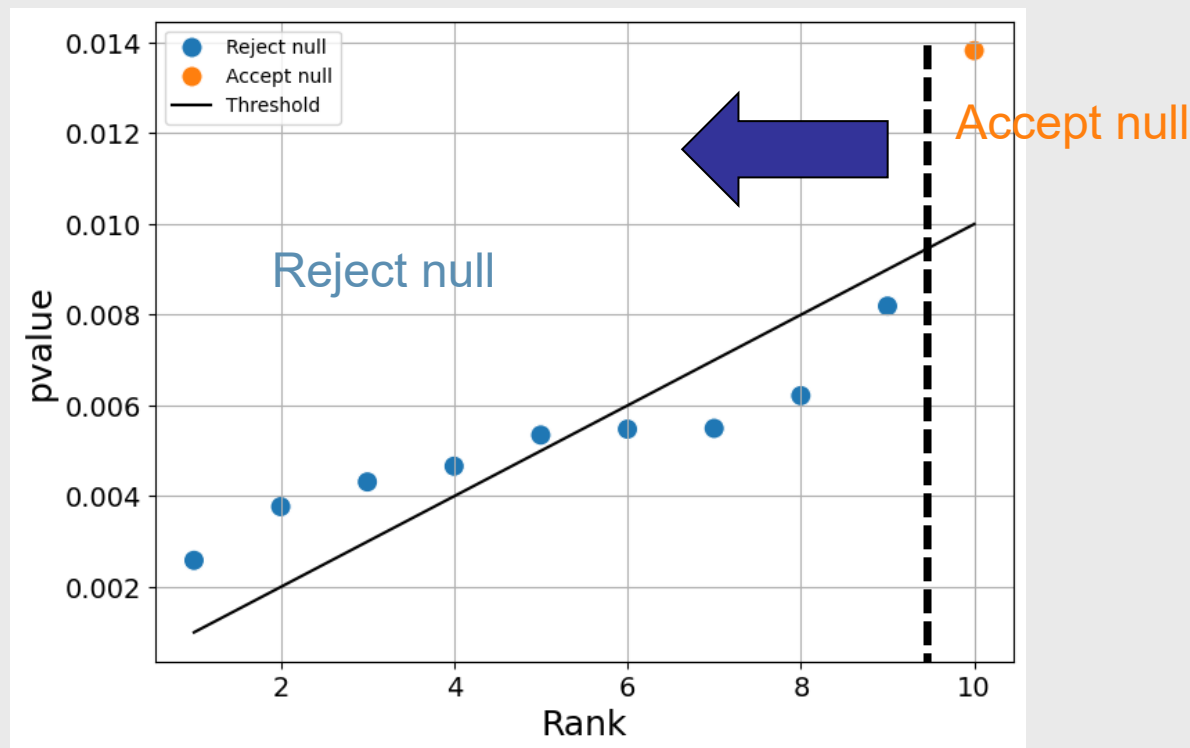
$$\frac{k}{m}\alpha$$

4. Find the largest p-value that is less than its threshold
5. Reject null for all smaller p-values (even if their p-value is bigger than their threshold)



# Benjamini-Hochberg Procedure

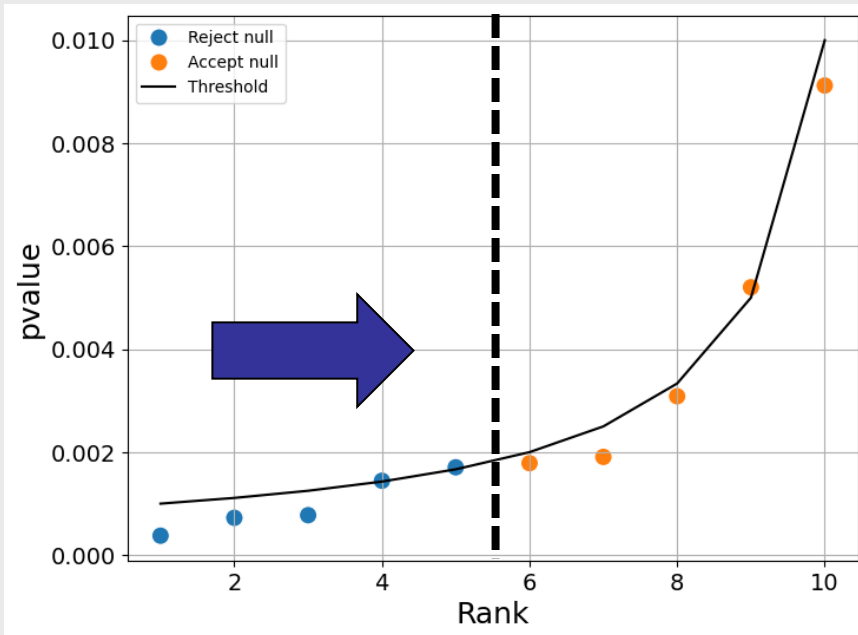
- Thresholds form a straight line when plotted versus p-value rank
- Start at largest p-value and go down until a p-value is below the threshold line
- All smaller p-values are significant (even if above their threshold)



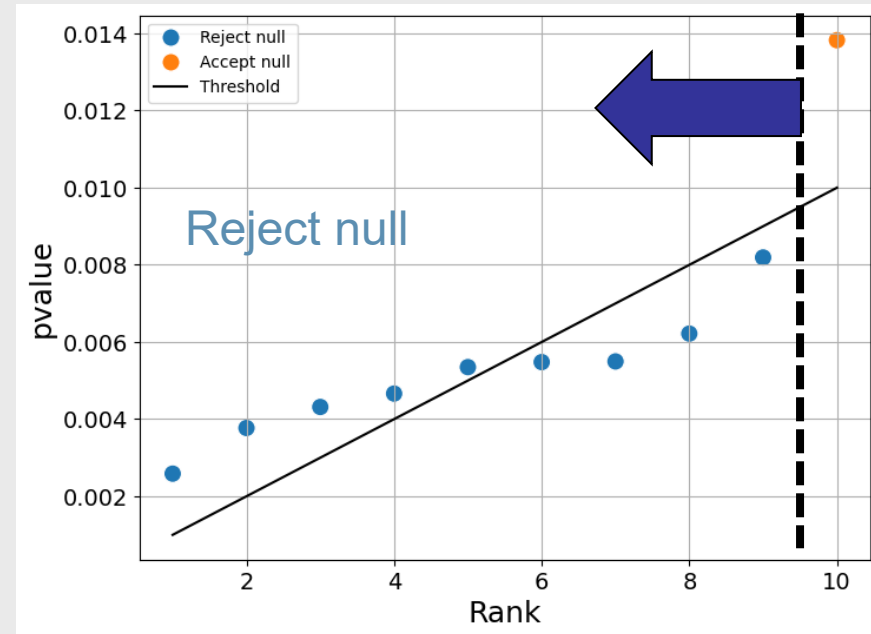
# HB vs BH

- Benjamini Hochberg procedure has many more “significant” p-values than Holm-Bonferroni

Holm-Bonferroni



Benjamini-Hochberg



# Multiple Hypotheses Tests

Test Name	P-value Threshold	Controlled Error	Method
Bonferroni Correction	$\frac{\alpha}{m}$	Family-wise error rate (FWER)	Fixed threshold
Sidak Correction	$1 - (1 - \alpha)^{1/m}$	FWER	Fixed threshold
Holm-Bonferroni Method	$\frac{\alpha}{m - k + 1}$	FWER	Step-up
Benjamini-Hochberg Procedure	$\frac{\alpha}{m}k$	False discovery rate (FDR)	Step-down

# Coding Session

- We will run A/B tests on some tweets
- Learn to use different statistical tests in Python