

Class 08

DATA1220-55, Fall 2024

Sarah E. Grabinski

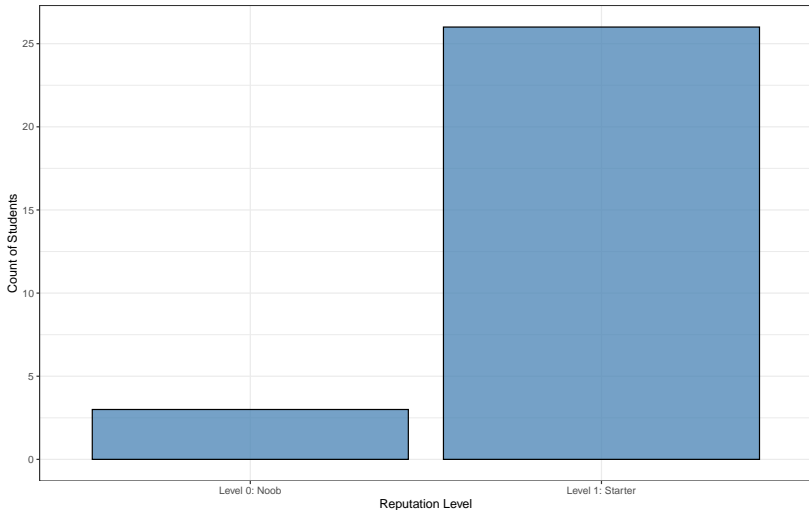
2024-09-16

Campuswire Discussion Post

- ▶ Read Section 2.2.5 in your OpenIntro Statistics text book
- ▶ Read this opinion piece on the use of pie charts to visualize proportions
- ▶ Answer the question on Campuswire for additional participation points
- ▶ It will not be available after Friday 9/20/24

Example: Bar Plot of the Count of Students by Reputation Level

Bar plot of the number of students by reputation level achieved
Reputation standings current as of 9/16/24

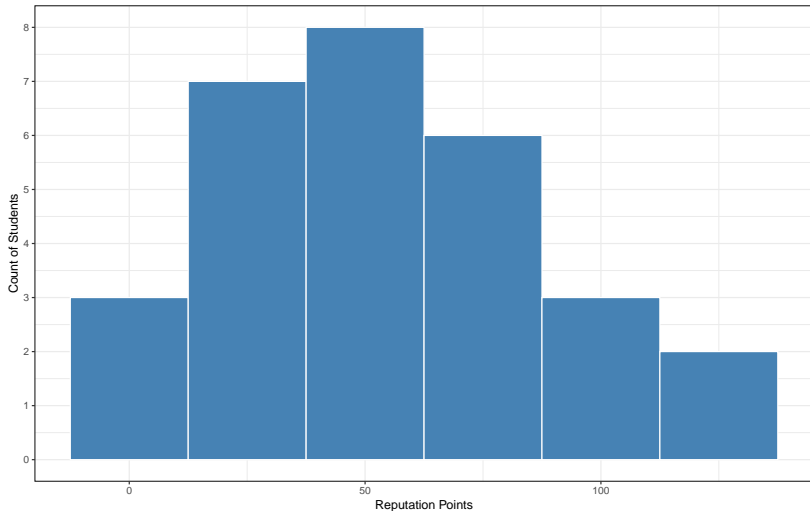


Sample size n = 29

Example: Histogram of Reputation Points

Histogram of the number of students by reputation points

Reputation standings current as of 9/16/24

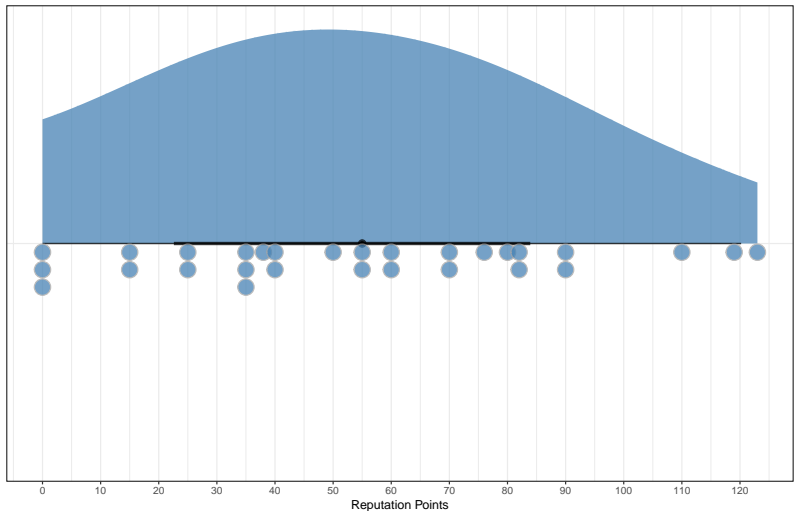


Sample size $n = 29$

Example: Rain Cloud Plot of Reputation Points

Rain Cloud Plot of Reputation Point Distribution

Reputation standings current as of 9/16/24

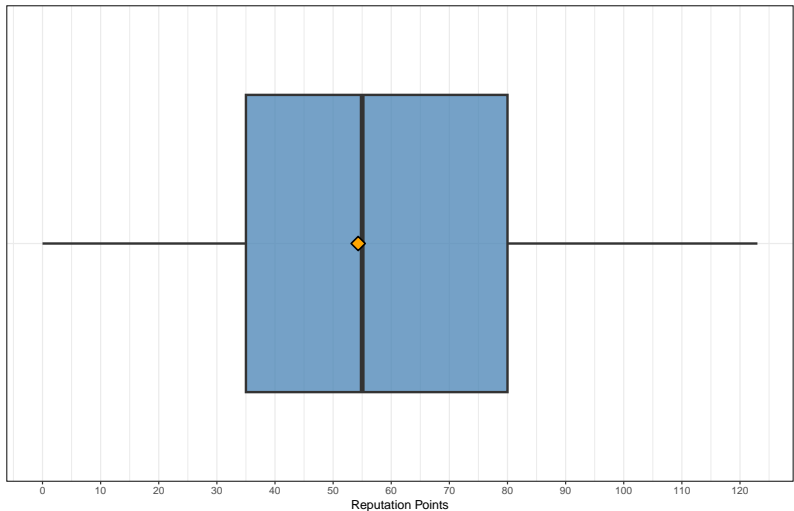


Sample size $n = 29$

Example: Boxplot of Median vs Mean in Reputation Points

Boxplot of Reputation Point Distribution

Reputation standings current as of 9/16/24



Sample size $n = 29$; Orange point = sample mean

Homework 2 Due Friday 9/20/24 by 6:00 PM

- ▶ Instructions (`homework2_instructions.pdf`), a Quarto markdown template (`homework2_template.qmd`), and an example HTML output (`homework2_example.html`) are available for download under Chapter 2 on the Modules page in Canvas.
- ▶ Upload **TWO** (2) documents to Homework 2 on the Assignments page in Canvas by **Friday 9/20/2024 by 6:00pm**: `homework2_yourlastname.qmd` and `homework2_yourlastname.html`
- ▶ Video walk-through of Homework 2 under Tutorials on the Modules page in Canvas. Make sure you're caught up on the video walk-through of homework 1.

Homework 1 Learnings (part I)

You now have experience making professional-looking HTML documents that embed statistical programming and data visualizations into traditional text.

- ▶ You started a project in RStudio
- ▶ You transferred files from your Downloads folder to your project folder
- ▶ You edited a Quarto Markdown Document (i.e. QMD, a file ending in `.qmd`)
- ▶ You rendered your Quarto Markdown Document into an HTML file (i.e. a file ending in `.html`).

Homework 1 Learnings (part II)

You have done basic statistical analysis.

- ▶ You loaded data into a project's environment in RStudio
- ▶ You inspected that data to determine the data type of the variables
- ▶ You created a codebook describing the variable
- ▶ You analyzed the relationship between 2 numerical variables
- ▶ You communicated your findings about that relationship

Homework 2 Objectives

- ▶ Effectively describe numerical distributions
- ▶ Select the appropriate summary statistics based on distribution shape
- ▶ Match numerical distributions to their summary statistics
- ▶ Calculate proportions from a contingency table

Late Policy

“This homework is due by 6:00pm on Friday, 9/20/24. No credit will be lost for assignments received by 7:00pm to account for issues with uploading. 10% of the points will be deducted from assignments received by 9:00am on Saturday, 9/21/24.

Assignments turned in after this point are only eligible for 50% credit, so it benefits you to turn in whatever you have completed by the due date.”

How can I get help with homework?

- ▶ **Read the textbook.** Many of you are asking for additional examples. Luckily, there are tons we didn't go over in the textbook.
- ▶ **Look at the homework early.** Only 1 person has looked at the homework since I posted it. Make sure you leave enough time to get help if you need it.
- ▶ **Ask a question on our Campuswire class feed.** I'm only one person, and I may not be able to give you a prompt answer. However, the 20+ other people in the class might be able to.
- ▶ **Come to office hours.** I will be available after class today (Monday 9/23/2024) and Wednesday 9/25/2024 from 2:30pm - 4:00pm. If you cannot make it, reach out to me to try and schedule an appointment.

Last Time...

- ▶ Contingency tables: counts and proportions (frequencies)
- ▶ Visualizing frequencies: bar plots, mosaic plots
- ▶ Describing numerical relationships: linear vs nonlinear, strong vs weak
- ▶ Visualizing 3+ variables

Today...

- ▶ Review Survey 1 Results
- ▶ Sarah's Objectives
- ▶ Introduce Chapter 3 on Probability

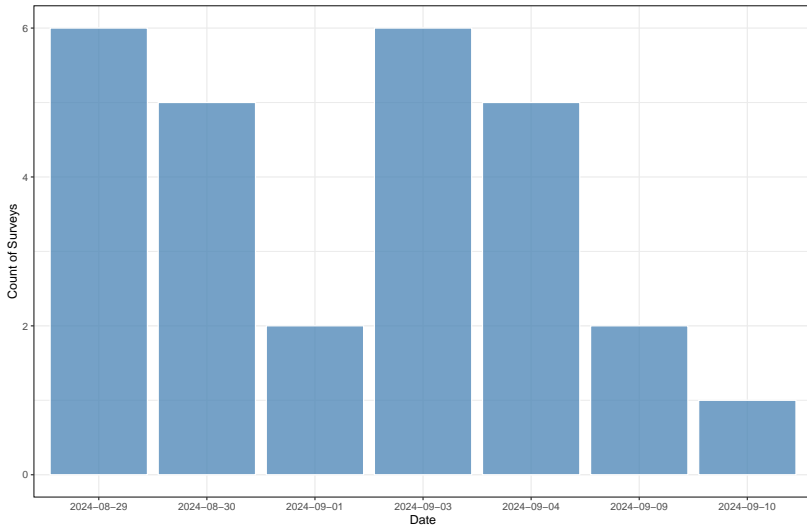
Survey 1 Results

# of Responses	# of Students	Response Rate
27	29	93.1%

Survey 1: Bar Plot of Surveys by Date

Number of Surveys Completed by Date in DATA1220-55

Responses received from 27 of 29 Students



Survey 1: Contingency Table of Graduation Year by Personal Pronouns (Counts)

Table 1: Count of Students by Graduation Year and Personal Pronouns

pronouns	2024	2025	2026	2027	2028	Total
He/Him	0	1	2	6	0	9
She/Her	1	2	3	7	3	16
Total	1	3	5	13	3	25

Survey 1: Contingency Table of Graduation Year and Personal Pronouns (Proportions by Row)

Table 2: Proportion of Graduation Years Among Students with He/Him and She/Her Pronouns ($n = 25$)

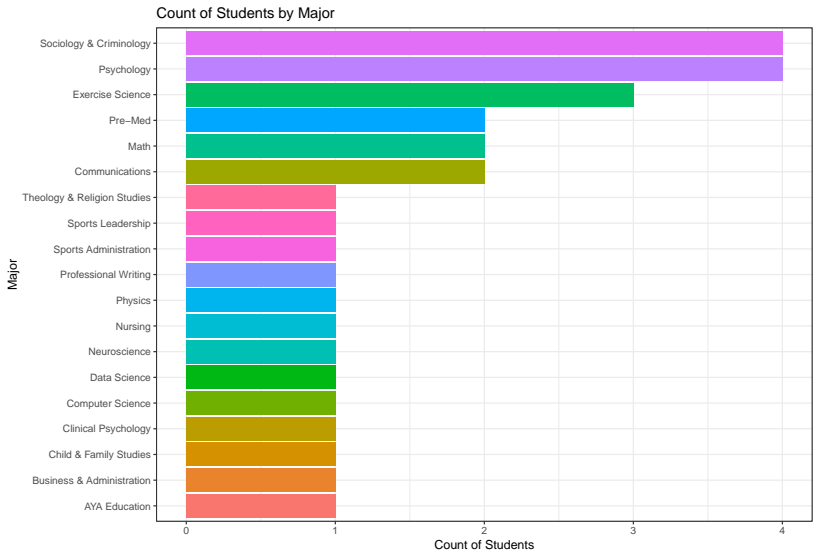
pronouns	2024	2025	2026	2027	2028	Total
He/Him	0.0%	11.1%	22.2%	66.7%	0.0%	100.0%
She/Her	6.2%	12.5%	18.8%	43.8%	18.8%	100.0%
Total	4.0%	12.0%	20.0%	52.0%	12.0%	100.0%

Survey 1: Contingency Table of Graduation Year by Personal Pronouns (Proportions by Column)

Table 3: Proportion of He/Him and She/Her Pronouns Among Students from Each Graduation Year ($n = 25$)

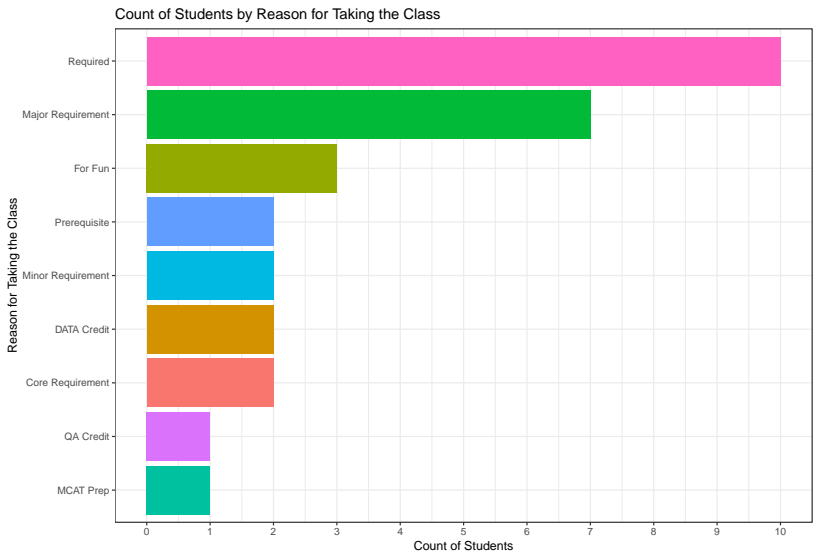
pronouns	2024	2025	2026	2027	2028	Total
He/Him	0.0%	33.3%	40.0%	46.2%	0.0%	36.0%
She/Her	100.0%	66.7%	60.0%	53.8%	100.0%	64.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Survey 1: Majors



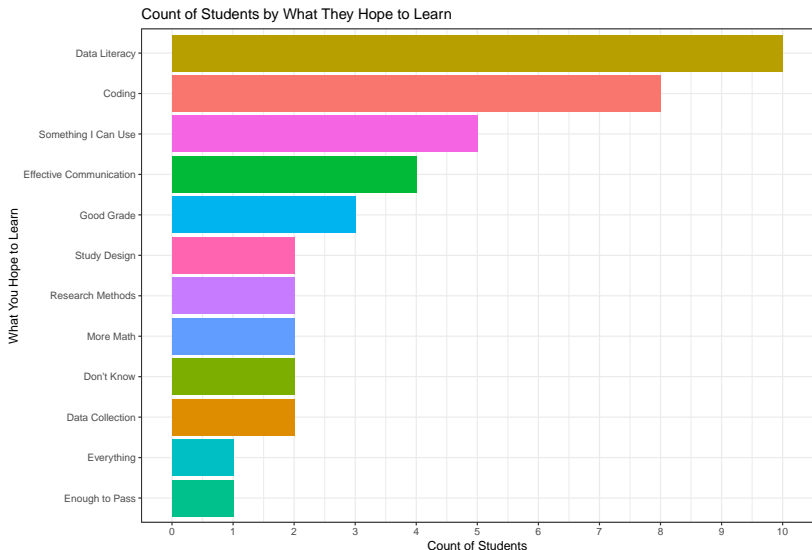
n = 27

Survey 1: Bar Plot of Motivations for Class



n = 27

Survey 1: Bar Plot of What you Hope to Learn



n = 27

Survey 1: Contingency Table of Level of Excitement and Coding Experience (Counts)

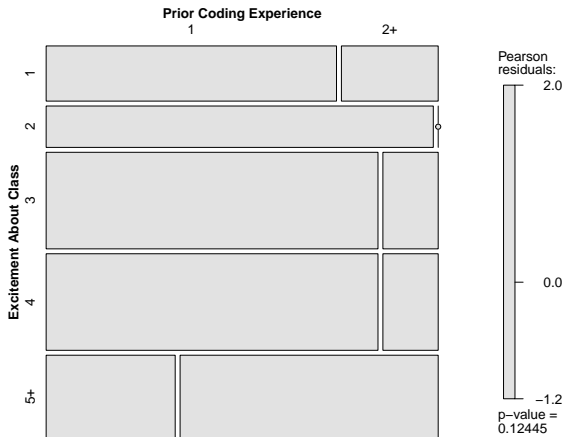
excited	1	2	3	4	Total
1	3	0	1	0	4
2	3	0	0	0	3
3	6	0	1	0	7
4	6	1	0	0	7
5	0	1	1	1	3
6	2	0	0	1	3
Total	20	2	3	2	27

Survey 1: Contingency Table of Level of Excitement and Coding Experience (With Binning)

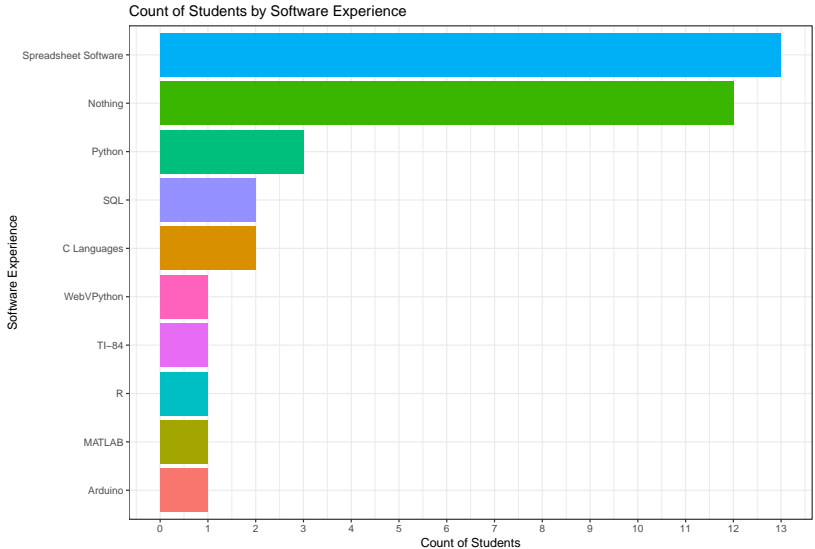
excited	1	2+	Total
1	3	1	4
2	3	0	3
3	6	1	7
4	6	1	7
5+	2	4	6
Total	20	7	27

Survey 1: Mosaic Plot of Level of Excitement and Coding Experience

Mosaic Plot of Excitement About Class
by Prior Coding Experience

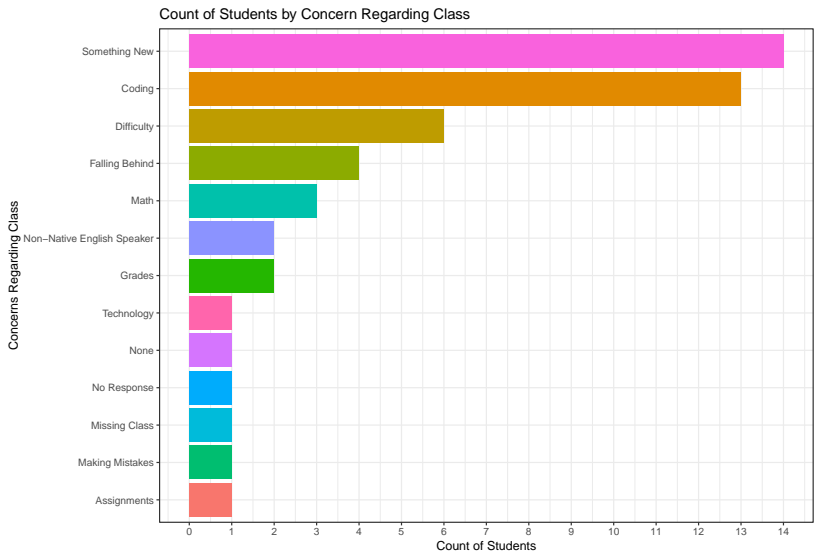


Survey 1: Bar Plot of Data Analysis Software Experience



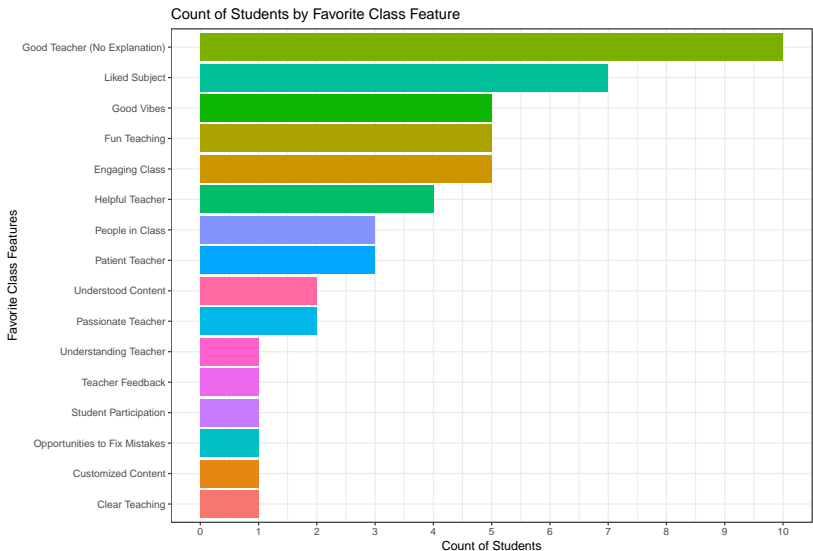
n = 27

Survey 1: Bar Plot of Concerns about Class



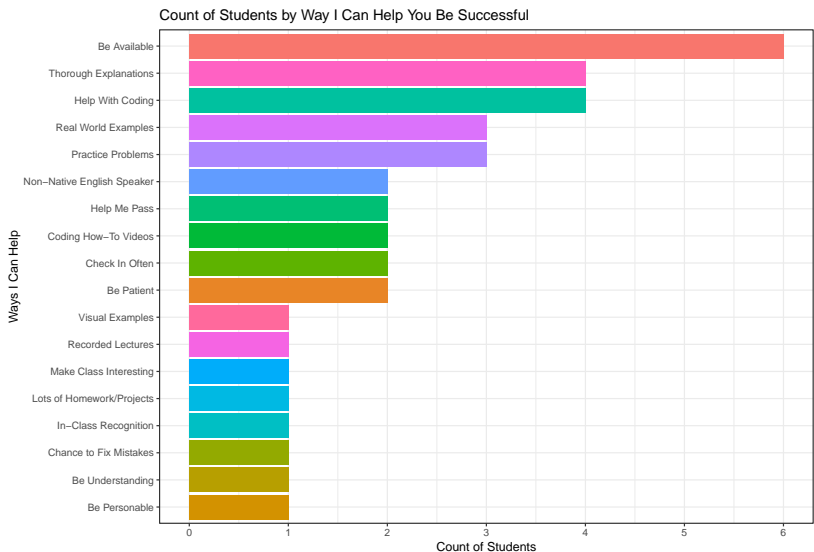
n = 27

Survey 1: Bar Plot of Favorite Class Features



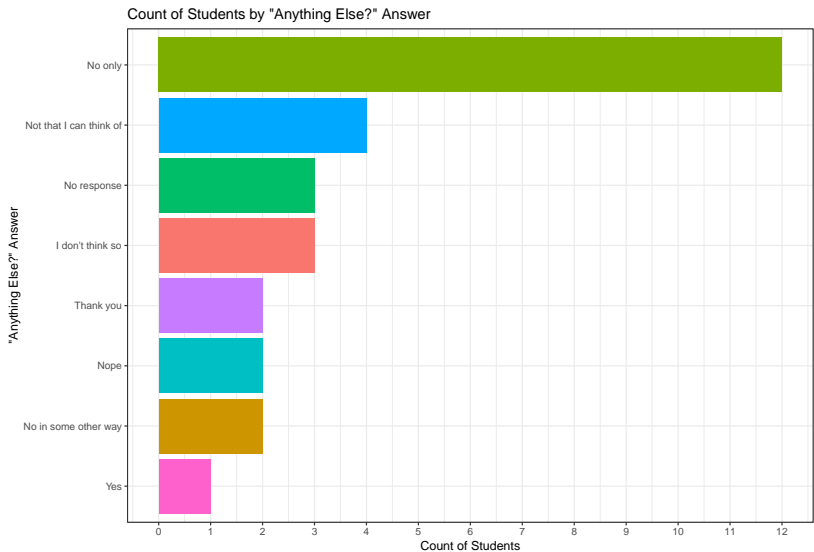
n = 27

Survey 1: Bar Plot of How I Can Help



n = 27

Survey 1: Bar Plot of "Anything Else?" Answers



n = 27

Sarah's Objectives

- ▶ ***Be available.*** I will check in with our next survey about office hour times.
- ▶ Provide ***real world examples*** and ***practice problems***.
- ▶ Help with ***coding*** and make ***how-to videos***.
- ▶ ***Be patient*** and give ***thorough explanations***.

Sarah's Other Objectives

- ▶ ***Be a good teacher.*** Unclear how.
- ▶ ***Get students to like statistics.*** Trying!
- ▶ Be ***fun, helpful, passionate,*** and ***engaging.***
- ▶ Promote ***good vibes.*** I need your help!

Your Objectives

Many people remarked that their favorite class was their favorite because of the people in it, who were engaged, had fun, and participated.

- ▶ Come prepared
- ▶ Ask and answer questions in class
- ▶ Ask and answer questions on Campuswire
- ▶ Don't be afraid to make mistakes

Chapter 3 Objectives

- ▶ Define probability, random processes, and the law of large numbers
- ▶ Describe the sample space for disjoint and non-disjoint outcomes
- ▶ Calculate probabilities using the General Addition and Multiplication Rules
- ▶ Create a probability distribution for disjoint outcomes

Defining Probability

What does the word **probability** mean to you?

Defining Probability

What does the word **probability** mean to you?

“Highly likely”

Defining Probability

What does the word **probability** mean to you?

“Highly likely”

“Probably”

Defining Probability

What does the word **probability** mean to you?

“Highly likely”

“Probably”

“About even”

Defining Probability

What does the word **probability** mean to you?

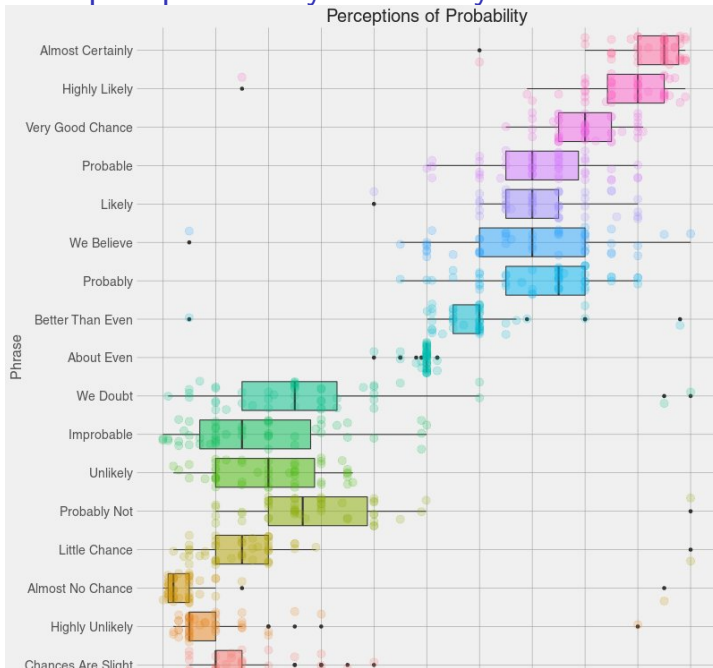
“Highly likely”

“Probably”

“About even”

“Almost no chance”

People interpret probability differently



So what is probability?

i Frequentist Definition

The proportion of times that a particular outcome would occur if we observed a random process an infinite number of times.

- ▶ A **random process** is one where you know which outcomes are possible (i.e. the **sample space**) but you don't know which outcome comes next
- ▶ Examples of a **random process**: coin toss, die roll, stock market

How do you know a process is random?

POPULAR SCIENCESCIENCE • TECHNOLOGY • ENVIRONMENT • DIY • GEAR • MERCH • NEWSLETTER🔍📱📺📺📧

Big Fall Wellness Sale

Our lowest prices of the year!

Now \$394


Shop now

TECHNOLOGY

A brief history of shuffling your songs, from Apple to Adele

Spotify made a change to one of music's most popular features. Here's what that means for how we listen to tunes.

BY [SHIRA FEDER](#) ✓ POSTED ON NOV 30, 2021 3:00 PM EST






SHIRA FEDER
Contributor, Tech

Shira Feder covers tech, science, and health. She holds a master's degree from the Craig Newmark Graduate School of Journalism and has written for The Washington...

2025 Equinox EV LT starting at \$34,995²



2024 Equinox EV LT as shown: \$43,295¹

Learn More


 **CHEVROLET**

Figure 2: Both Apple and Spotify took steps to make their “shuffle” features less random after complaints from users

A brief history of “Shuffle”

- ▶ January 11, 2005 – Apple releases the iPod Shuffle, a small device capable only of playing music randomly
 - ▶ September 7, 2005 – Apple offers “Smart Shuffle” in response to complaints, which controlled how likely songs from the same album or artist would play close together
- ▶ July 2011 – Spotify launches in the United States using the Fisher-Yates Algorithm, which is like picking tickets out of a hat until no more remain
 - ▶ February 2014 – Spotify modifies their sampling algorithm to ensure an even distribution across albums/artists

What went wrong?

- ▶ The human brain is good at finding patterns in noise, even when there are none
- ▶ If an artist is repeated “too soon”, the listener doesn't feel the order is random
- ▶ We perceive a “random” distribution as also being “uniform” and “fair”

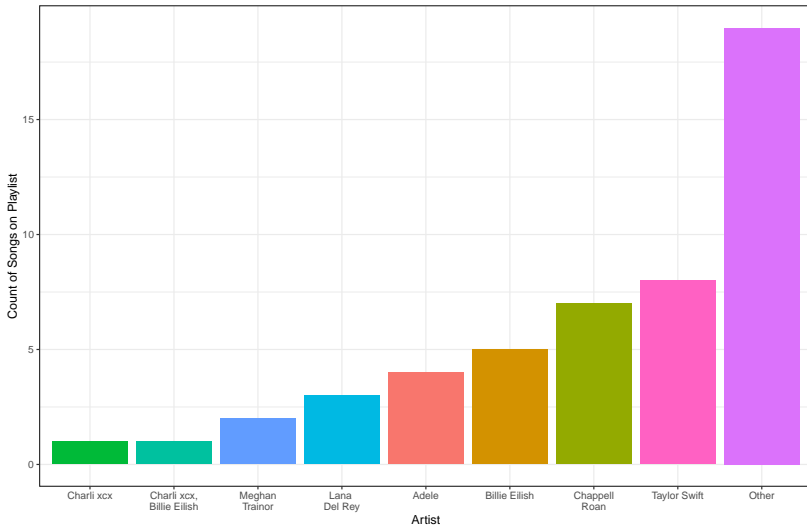
So why didn't we like a “true” random shuffle?

- ▶ Songs not evenly distributed across albums and artists on a playlist
 - ▶ Some albums/artists may play more often than others because they have more songs
 - ▶ Artists/albums with more songs also more likely to play sequentially
- ▶ A true random shuffle might play the same artist multiple times in a row
 - ▶ It's unusual but not impossible to roll a 1 on a die 6 times in a row

Example: Spotify Playlists

Number of Songs on the 'Taylor Swift Radio' playlist on Spotify by Artist

Total artists = 26



What if shuffle was truly random?

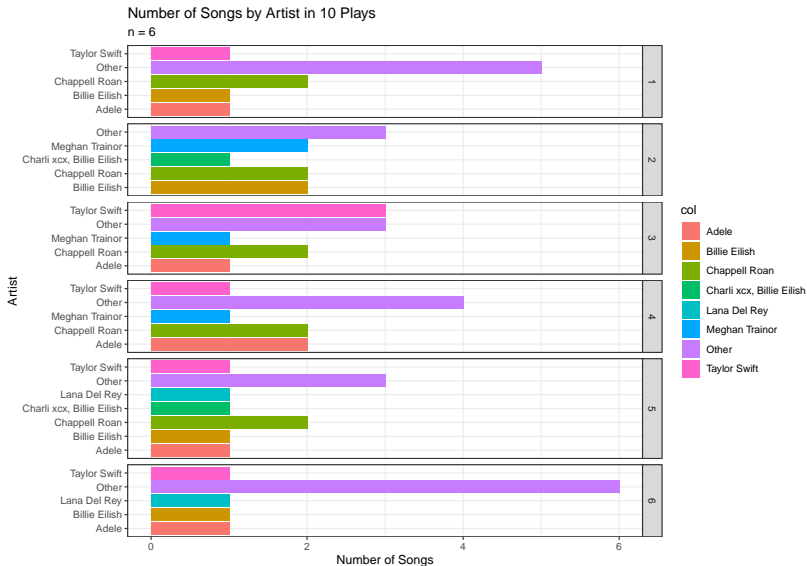
Each time the song changes, every song on the playlist is eligible to be played next

- ▶ Does not matter if the song was just played
- ▶ Does not matter who the artist is

We call this ***sampling with replacement***.

- ▶ Like drawing a playing card, looking at it, then putting it back in the deck before the next draw.
- ▶ Repetition of outcomes is possible.

How often were artists repeated during Spotify's original shuffle?



What is the probability of hearing a song by Chappell Roan?

- ▶ There is our “observed” probability that the next song is by Chappell Roan
 - ▶ Sample proportion (\hat{p}_n)
- ▶ There is some “true” real-world probability that the next song is by Chappell Roan
 - ▶ Population proportion (p)

Defining the sample space

The **sample space** is the total collection of possible outcomes for a **random process**.

- ▶ Die rolls: 1, 2, 3, 4, 5, 6
- ▶ Coin flips: heads, tails
- ▶ Stock market: up, down, no change

Here, the **sample space** is the songs on the playlist ($n = 50$) and the artists who perform them ($n = 26$).

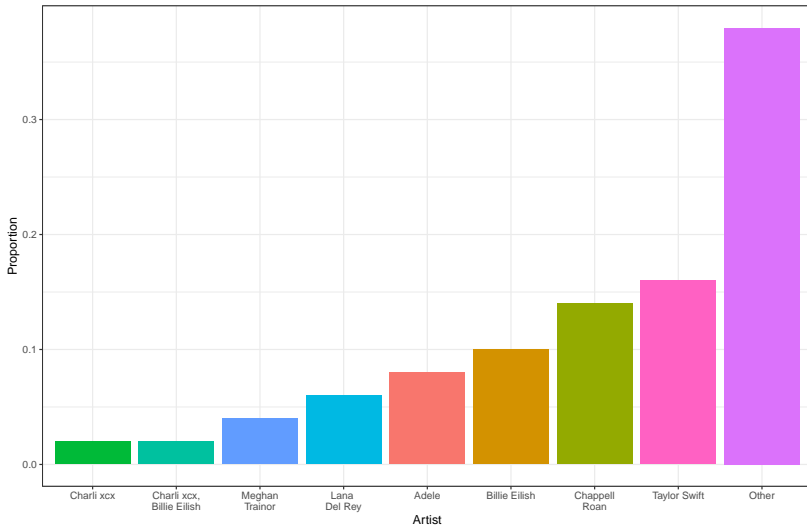
Calculating probabilities

- ▶ Probabilities are proportions, or the number of observations with a particular value divided by the total number of observations (n).
- ▶ Proportions range from 0 (no observations in data) to 1 (all observations in data)
- ▶ Also may be a percentage, ranging from 0% to 100% (multiply proportion by 100)

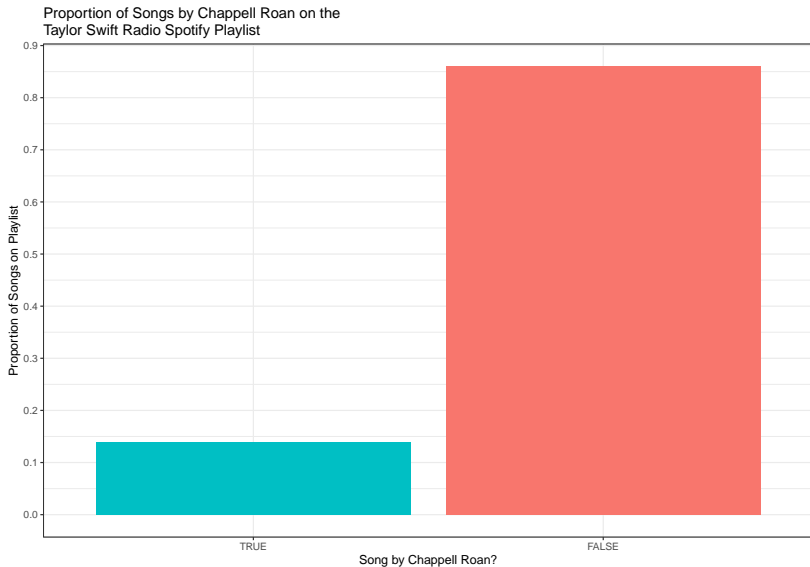
Proportion of Songs by Artist

Proportion of Songs by Artist on Taylor Swift Radio

n = 50



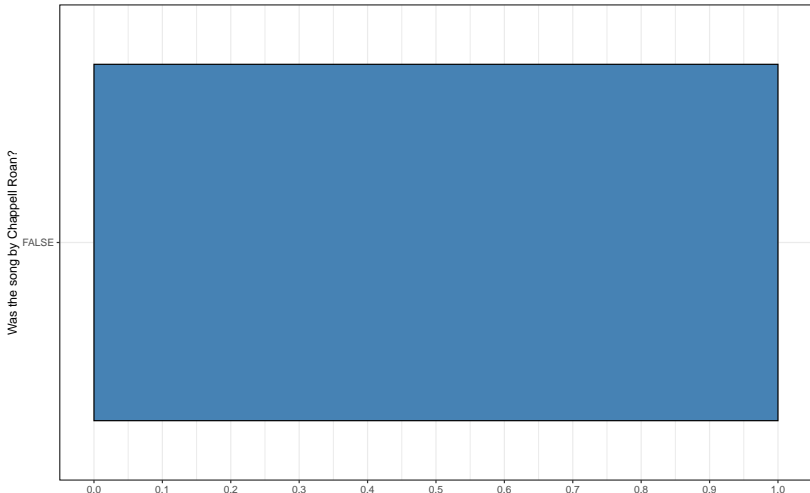
Proportion of Chappell Roan songs



How well does the sample proportion represent the population proportion?

Should we listen to 1 song?

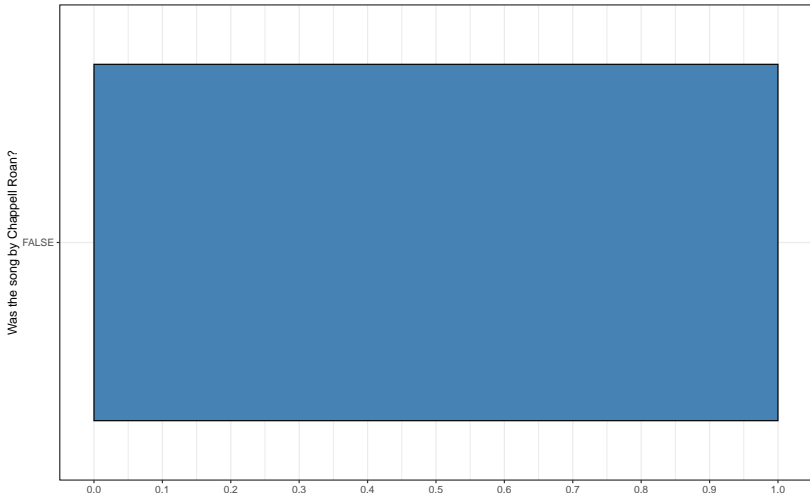
Proportion of Songs by Chappell Roan on the
"Taylor Swift Radio" Spotify Playlist
Number of songs listened to: $n = 1$



How well does the sample proportion represent the population proportion?

Should we listen to 5 songs?

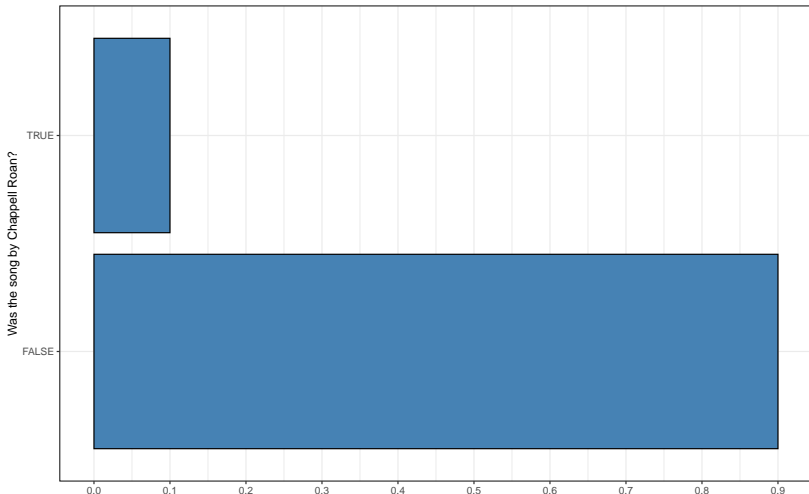
Proportion of Songs by Chappell Roan on the
"Taylor Swift Radio" Spotify Playlist
Number of songs listened to: $n = 5$



How well does the sample proportion represent the population proportion?

Should we listen to 10 songs?

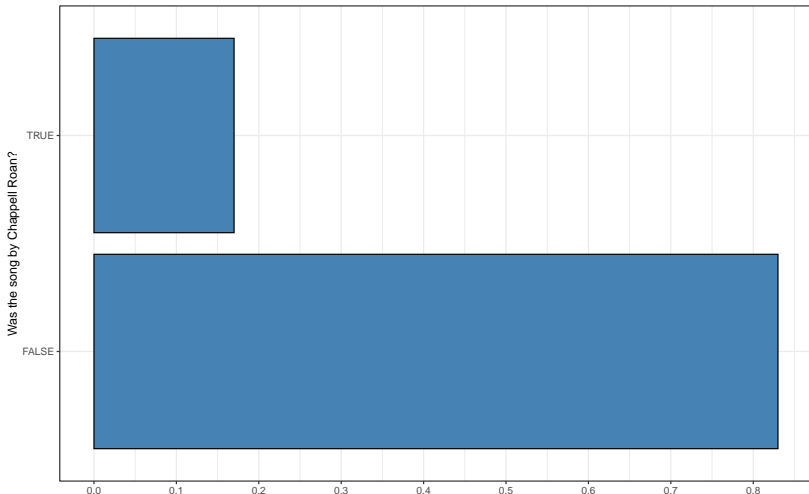
Proportion of Songs by Chappell Roan on the
"Taylor Swift Radio" Spotify Playlist
Number of songs listened to: $n = 10$



How well does the sample proportion represent the population proportion?

Should we listen to 100 songs?

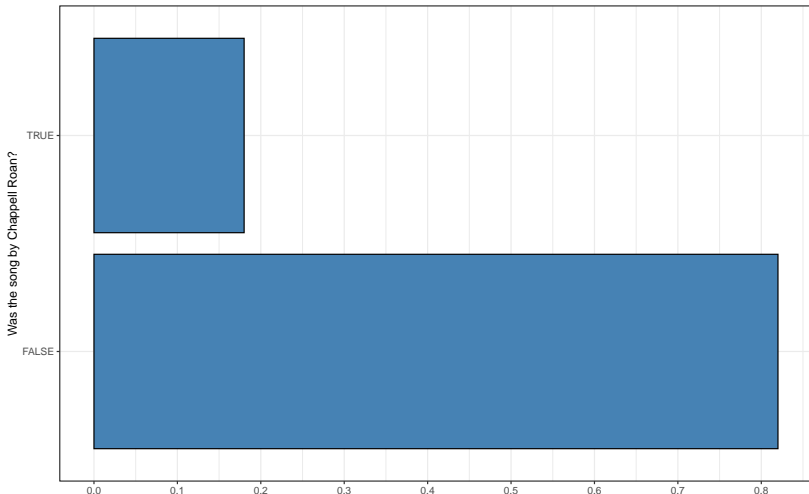
Proportion of Songs by Chappell Roan on the
"Taylor Swift Radio" Spotify Playlist
Number of songs listened to: $n = 100$



How well does the sample proportion represent the population proportion?

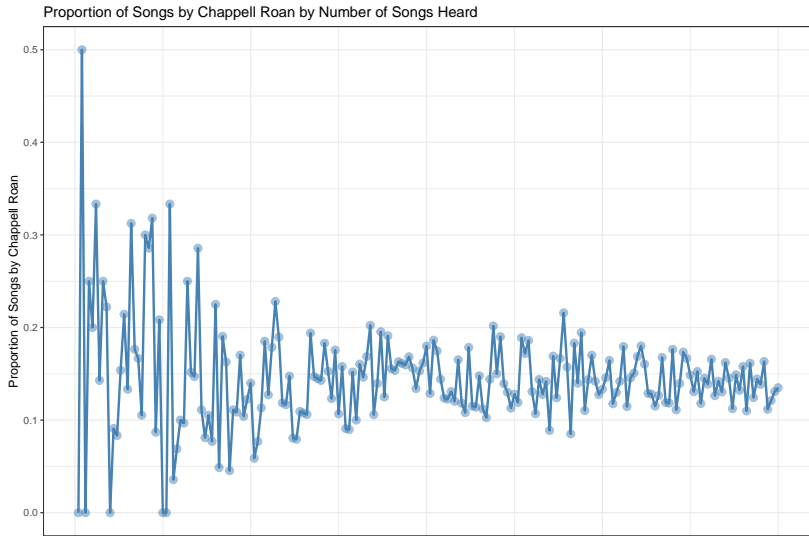
Should we listen to 200 songs?

Proportion of Songs by Chappell Roan on the
"Taylor Swift Radio" Spotify Playlist
Number of songs listened to: $n = 200$



Law of Large Numbers

As more observations are collected, the sample proportion \hat{p}_n of a particular outcome approaches the population proportion p of that outcome.



Disjoint Outcomes

Outcomes are **disjoint** or **mutually exclusive** if they cannot both happen at the same time

- ▶ The next song played cannot be by both Taylor Swift and Chappell Roan

Non-disjoint outcomes can occur at the same time.

- ▶ The next song played could be by Charlie xcx OR Billie Eilish, because they collaborated on a song