Class 09

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-09-18

Homework 2

- 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.

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.
- **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 Wednesday 9/25/2024 from 2:30pm - 4:00pm. If you cannot make it, reach out to me to try and schedule an appointment.

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?

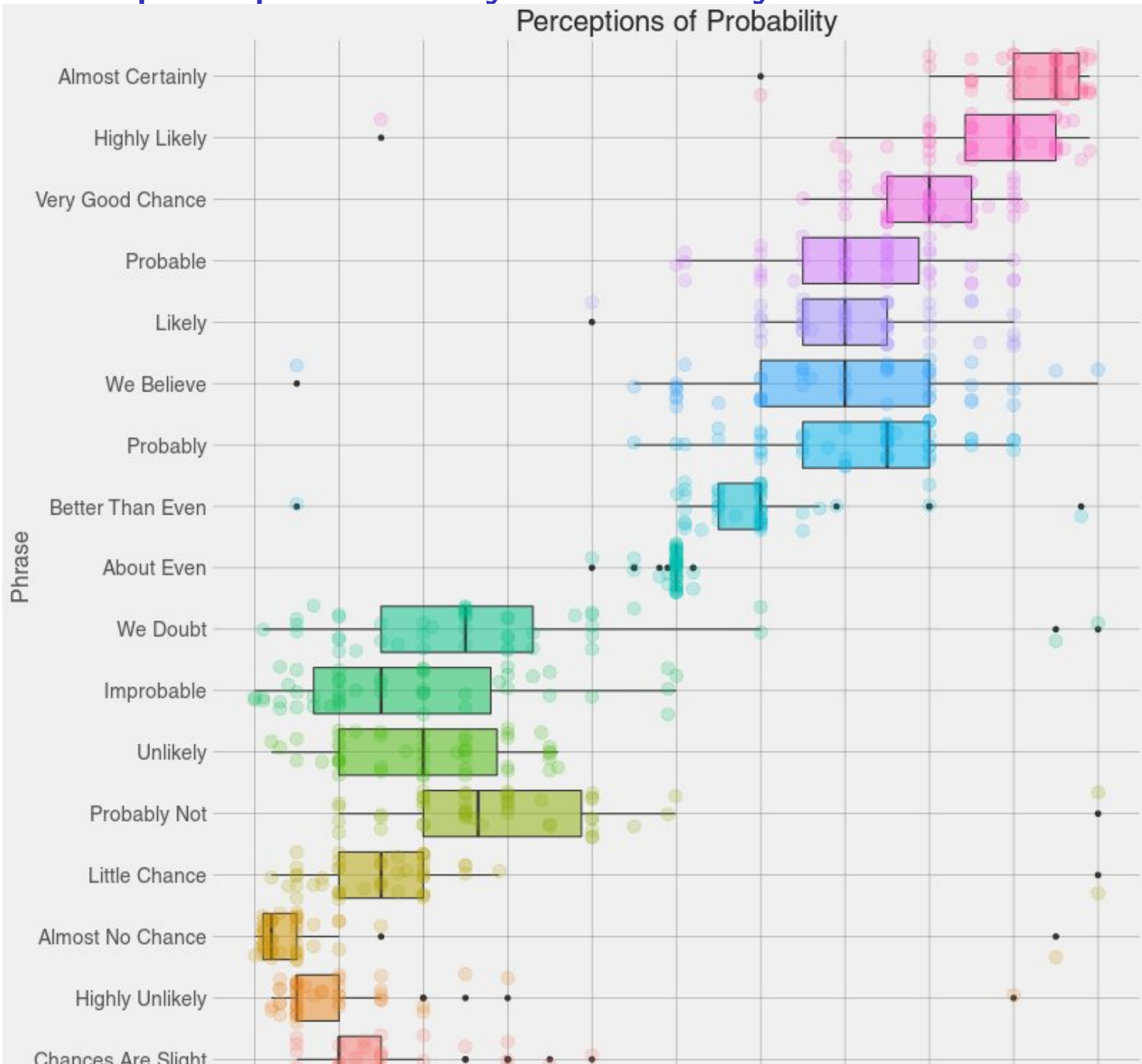
"Highly likely"

"Probably"

"About even"

"Almost no chance"

People interpret probability differently



So what is probability?

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?

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 ("true" shuffle)
  - 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

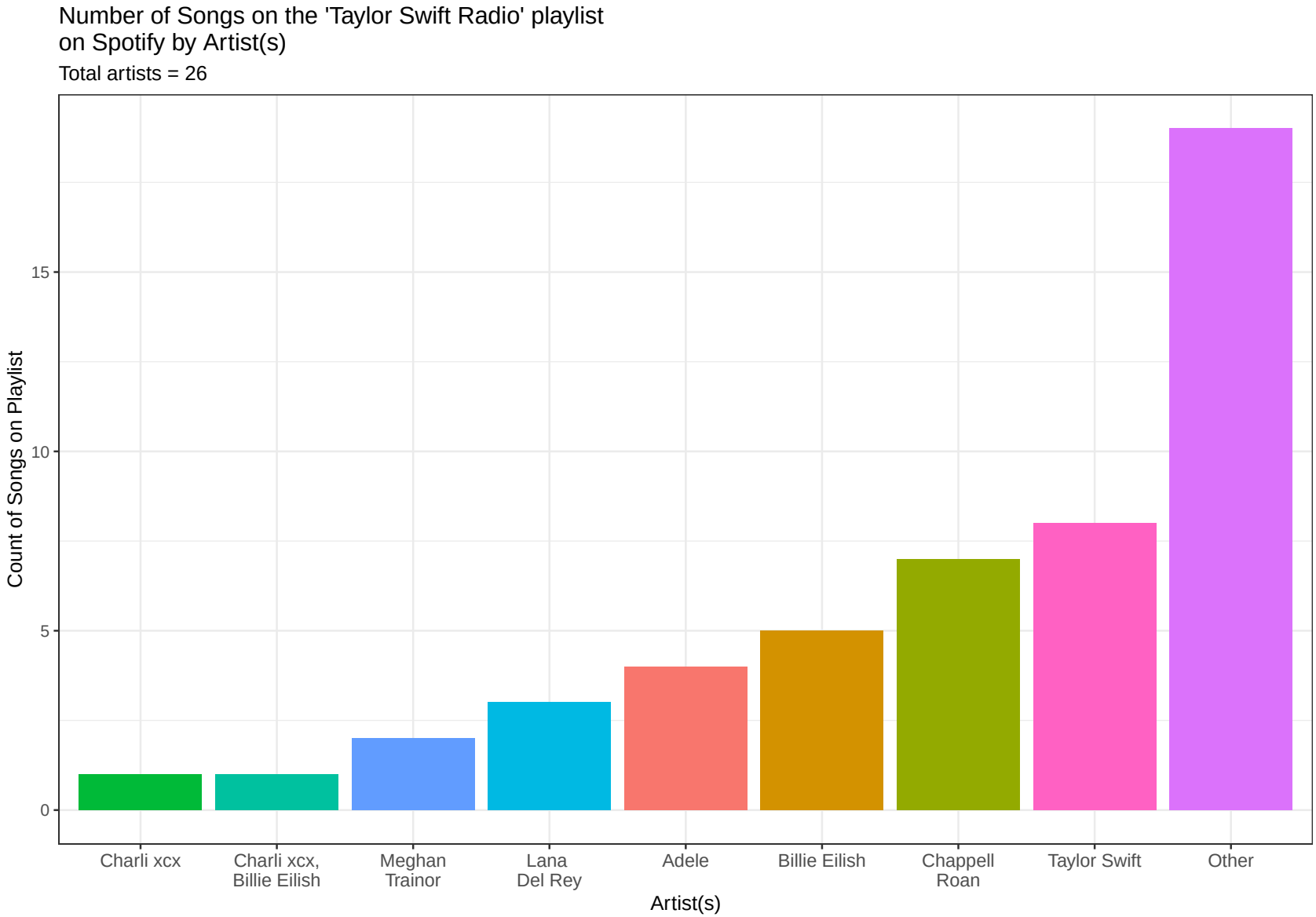
What went wrong?

- 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 frequently than others simply because they have more songs in the library/on the playlist
  - Each song is equally likely to play next (uniform), but not each artist (not uniform)
  - Artists/albums with more songs also more likely to play in a row
- 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 3 times in a row
  - It's also possible for the same song to play twice in a row

Example: Spotify Playlists



What if shuffle was truly random?

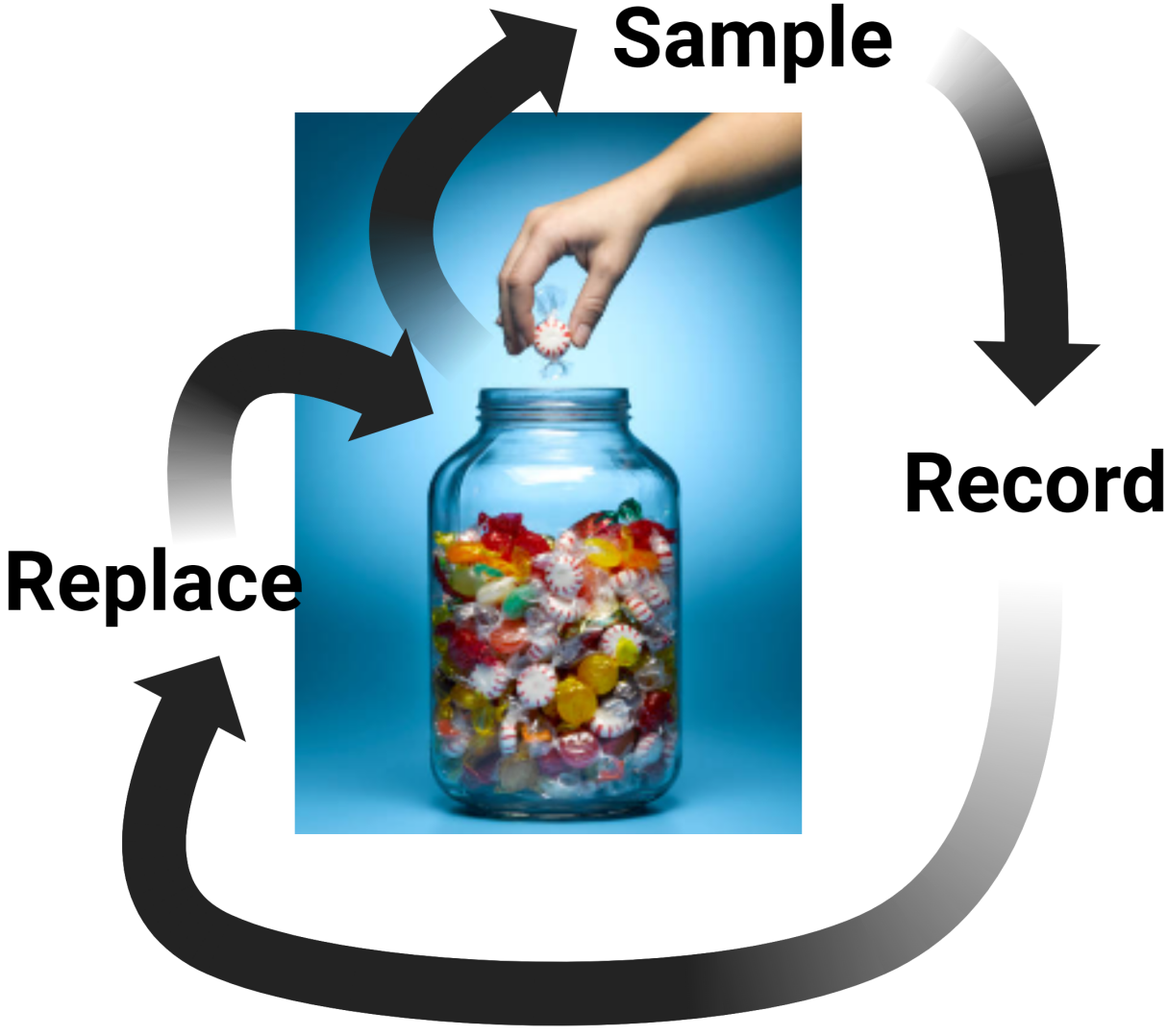
Each time the song changes, every song on the playlist is eligible to be played next

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

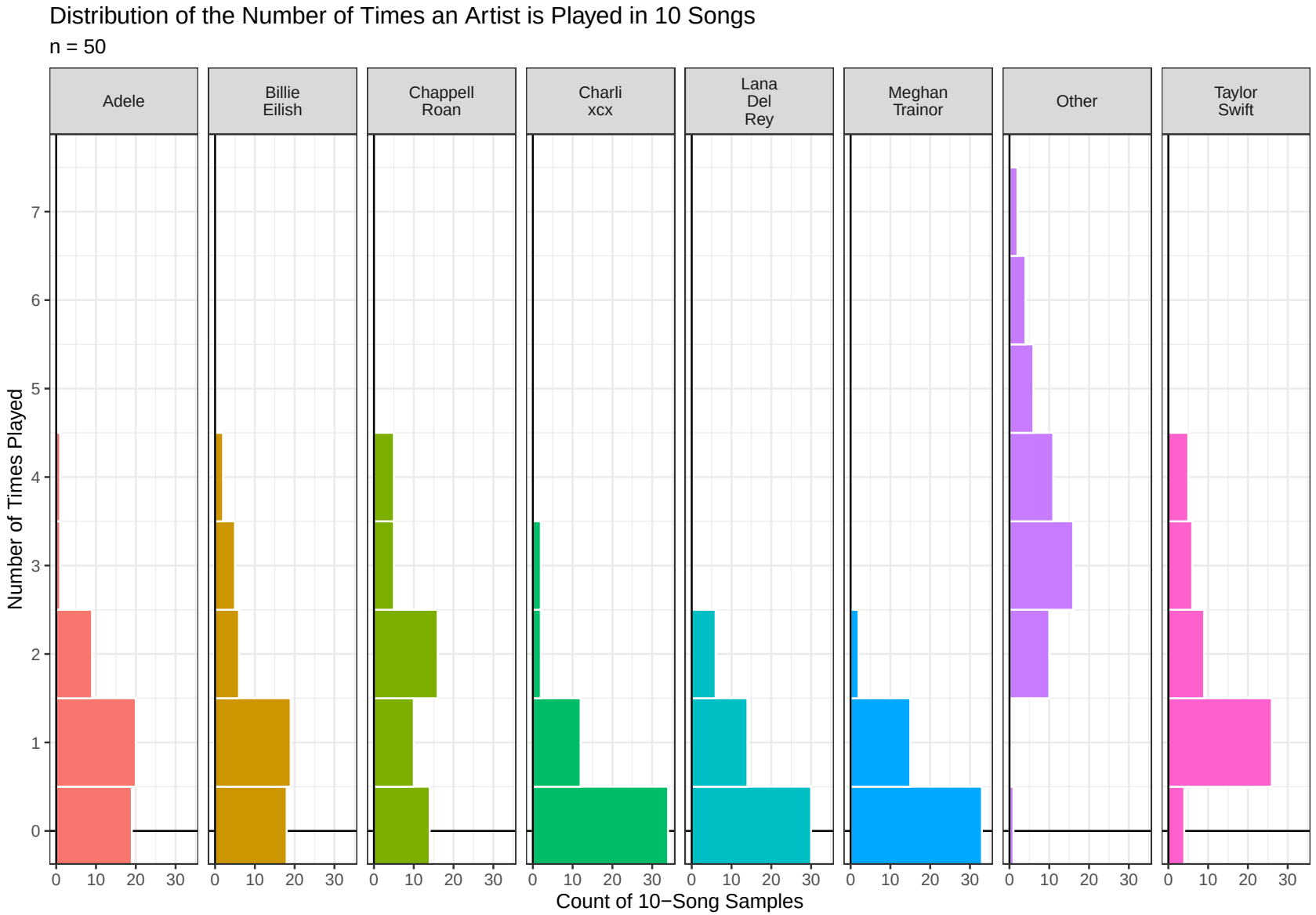
We call this ***sampling with replacement***.

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

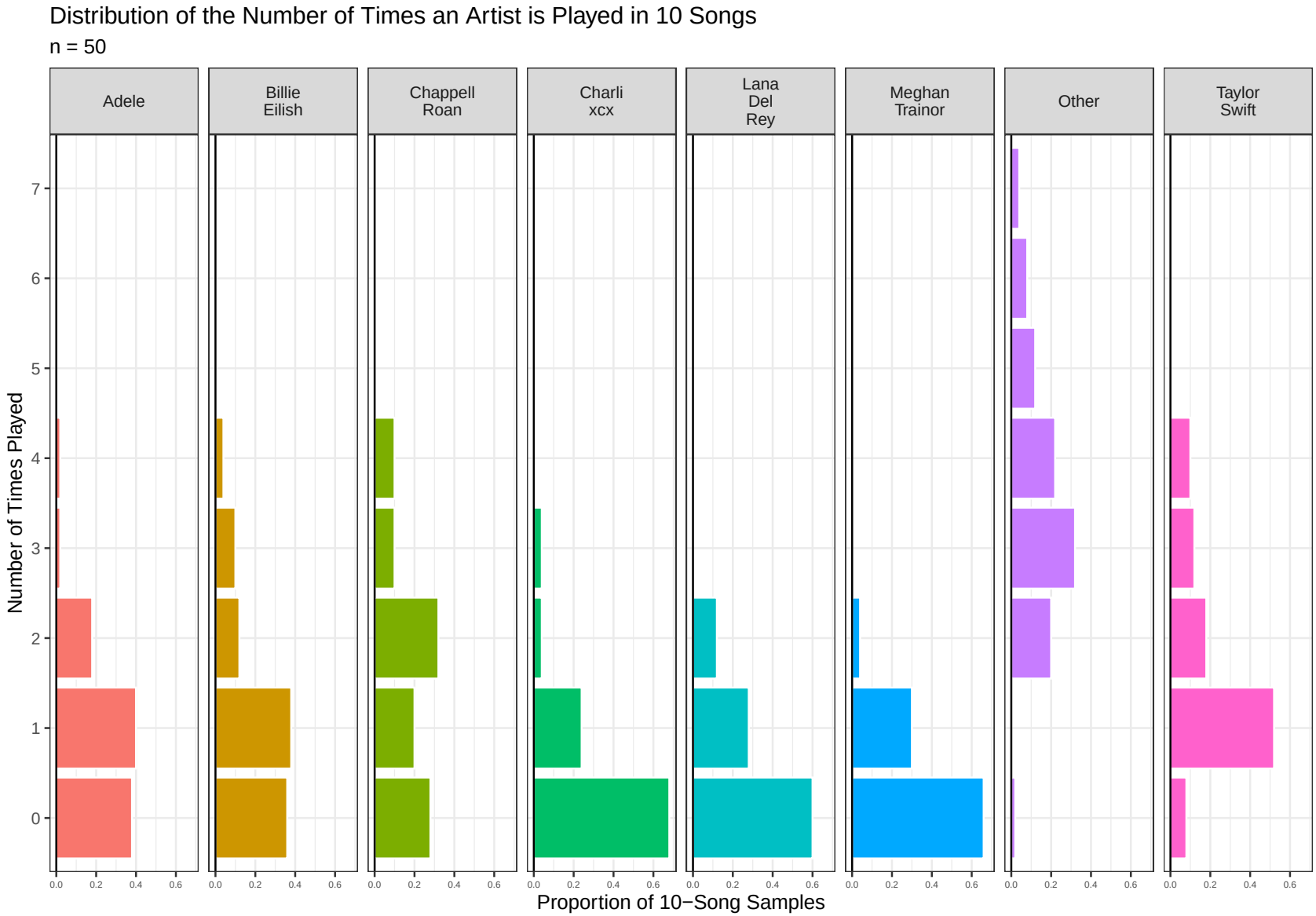
Sampling With Replacement



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



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



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

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

Defining the sample space

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

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

For this example, the **sample space** could be all the songs on the playlist (n = 50) or all the artists who perform them (n = 26).

Example sample spaces

Flipping a Coin

SAMPLE SPACE

{Head, Tail}

Uniform

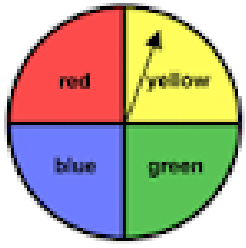
Rolling a Six Sided Dice

SAMPLE SPACE

{1, 2, 3, 4, 5, 6}

Uniform

Spinning a 4 color spinner



SAMPLE SPACE

{Red, Yellow, Green, Blue}

Uniform

Rolling a Weighted Dice

SAMPLE SPACE

{4, 5, 6}

Not Uniform

Picking a flavor of ice cream

SAMPLE SPACE

{Chocolate, Vanilla, Strawberry}

Uniform

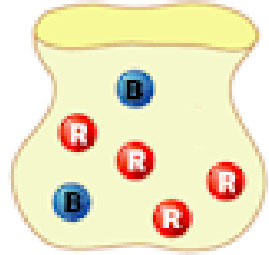
Determining the gender of baby

SAMPLE SPACE

{Boy, Girl}

Uniform

Picking from a bag of marbles

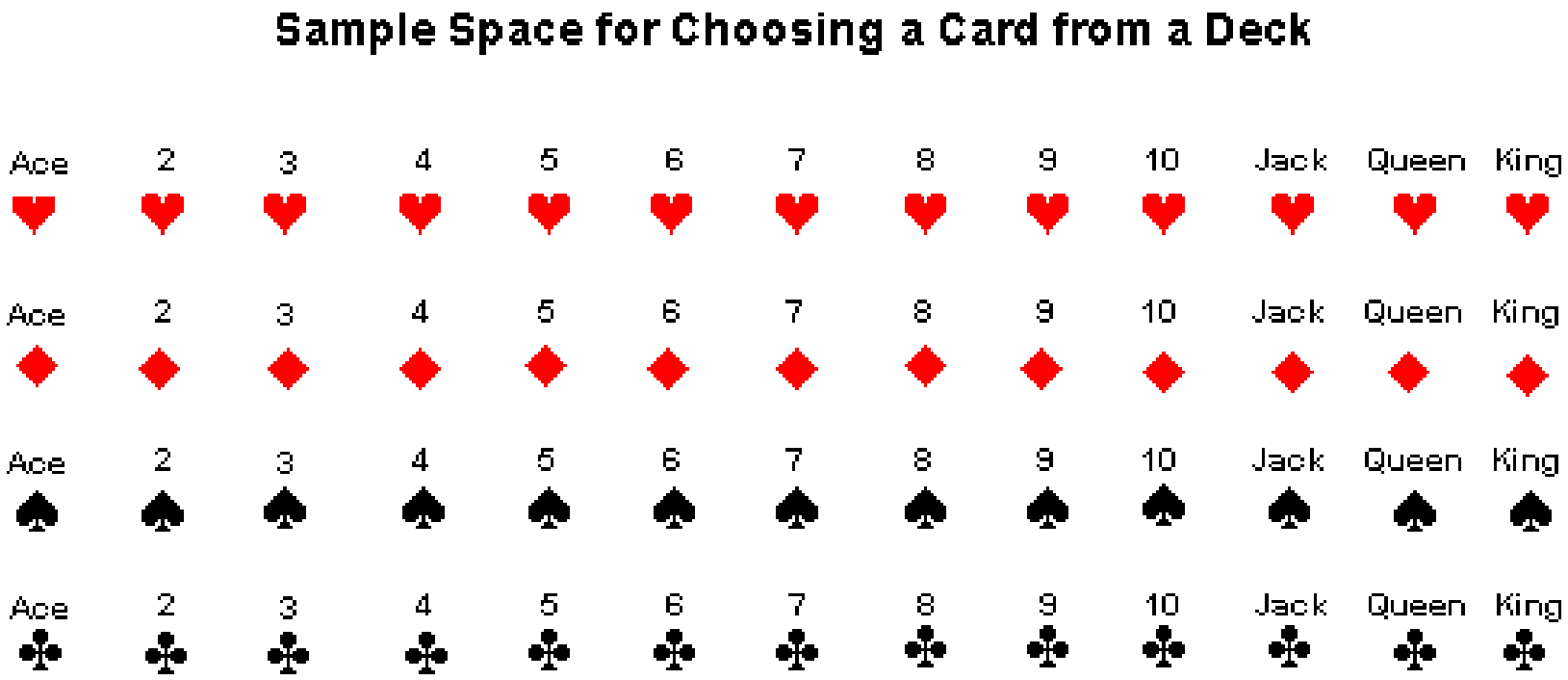


SAMPLE SPACE

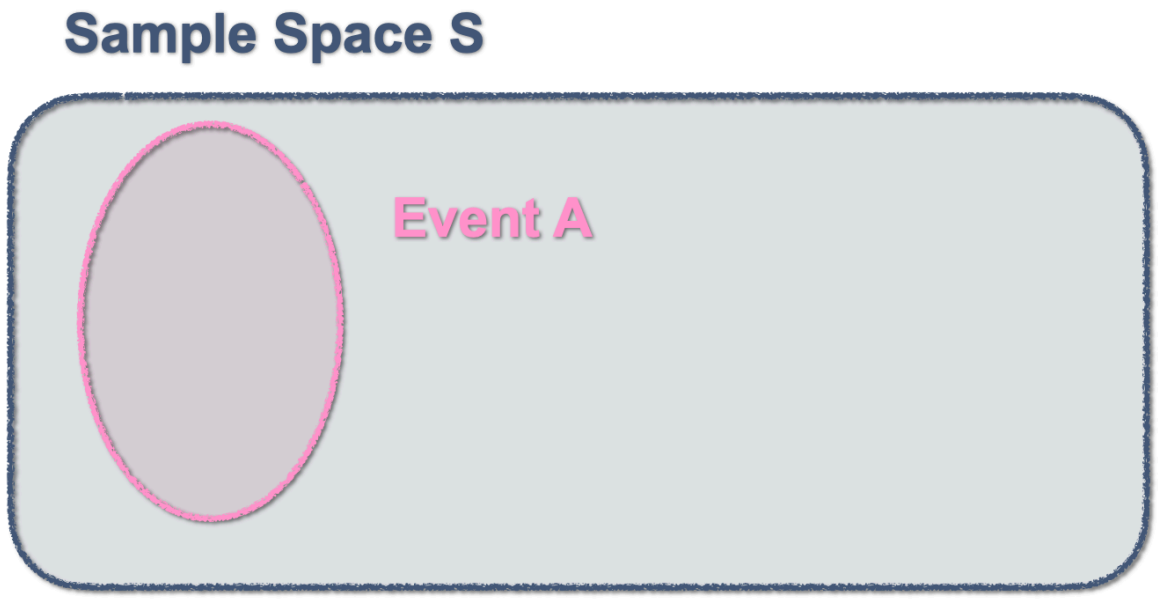
{Blue, Red}

Not Uniform

Another sample space

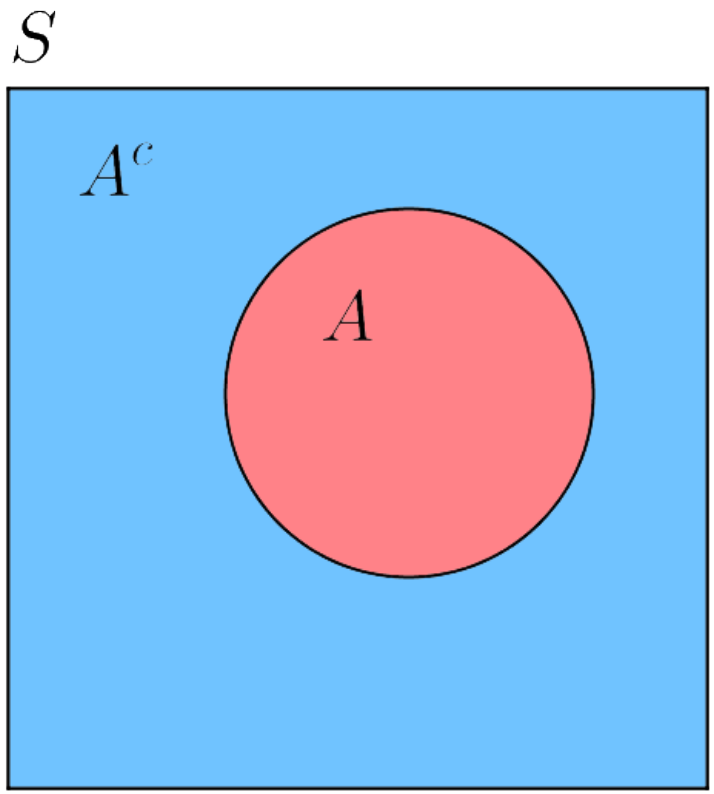


Representing the sample space - 1 event



Representing the sample space - complements

In the sample space S, the complement of event A occurring is event A *not* occurring. This is written as A^C or A'.



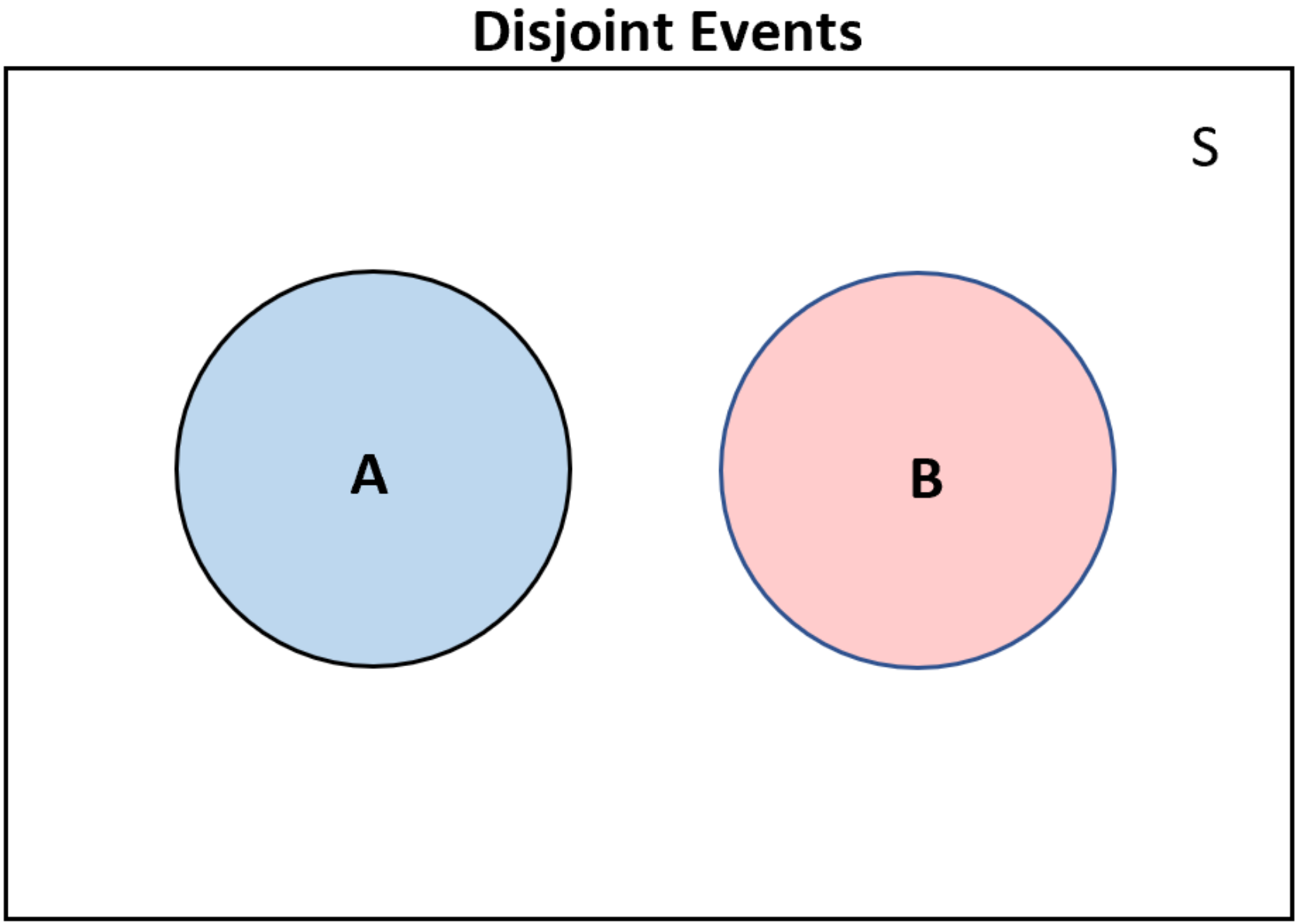
Sample Space S, event A, and complement A^c

2 Events - Disjoint Outcomes

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

- Taylor Swift and Adele did not collaborate on any songs on this playlist
- The next song played can either be by Taylor Swift OR by Chappell Roan but not by Taylor Swift AND Chappell Roan
- The events "The next song is by Taylor Swift" and "The next song is by Chappell Roan" are ***disjoint*/*mutually exclusive***

Defining 2 disjoint events in the sample space

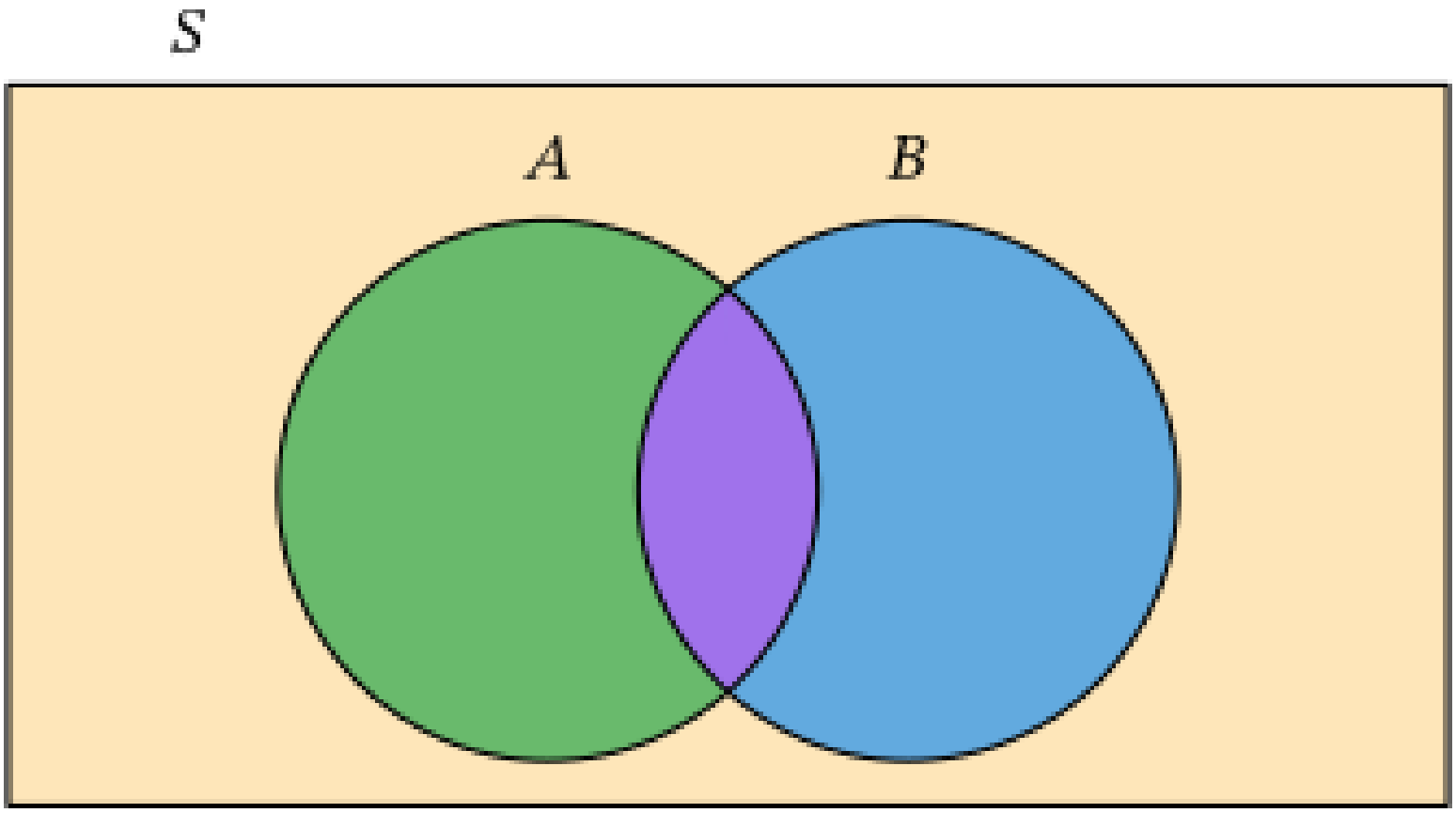


2 Events - Non-Disjoint Outcomes

Non-disjoint outcomes can occur at the same time.

- Charli xcx and Billie Eilish collaborated on a song on this playlist
- The next song played could be by Charli xcx or by Billie Eilish OR by BOTH Charli xcx and Billie Eilish.
- The events "The next song is by Charli xcx" and "The next song is by Billie Eilish" are ***non-disjoint***

Defining 2 non-disjoint events in the sample space



Example: Non-Disjoint Outcomes

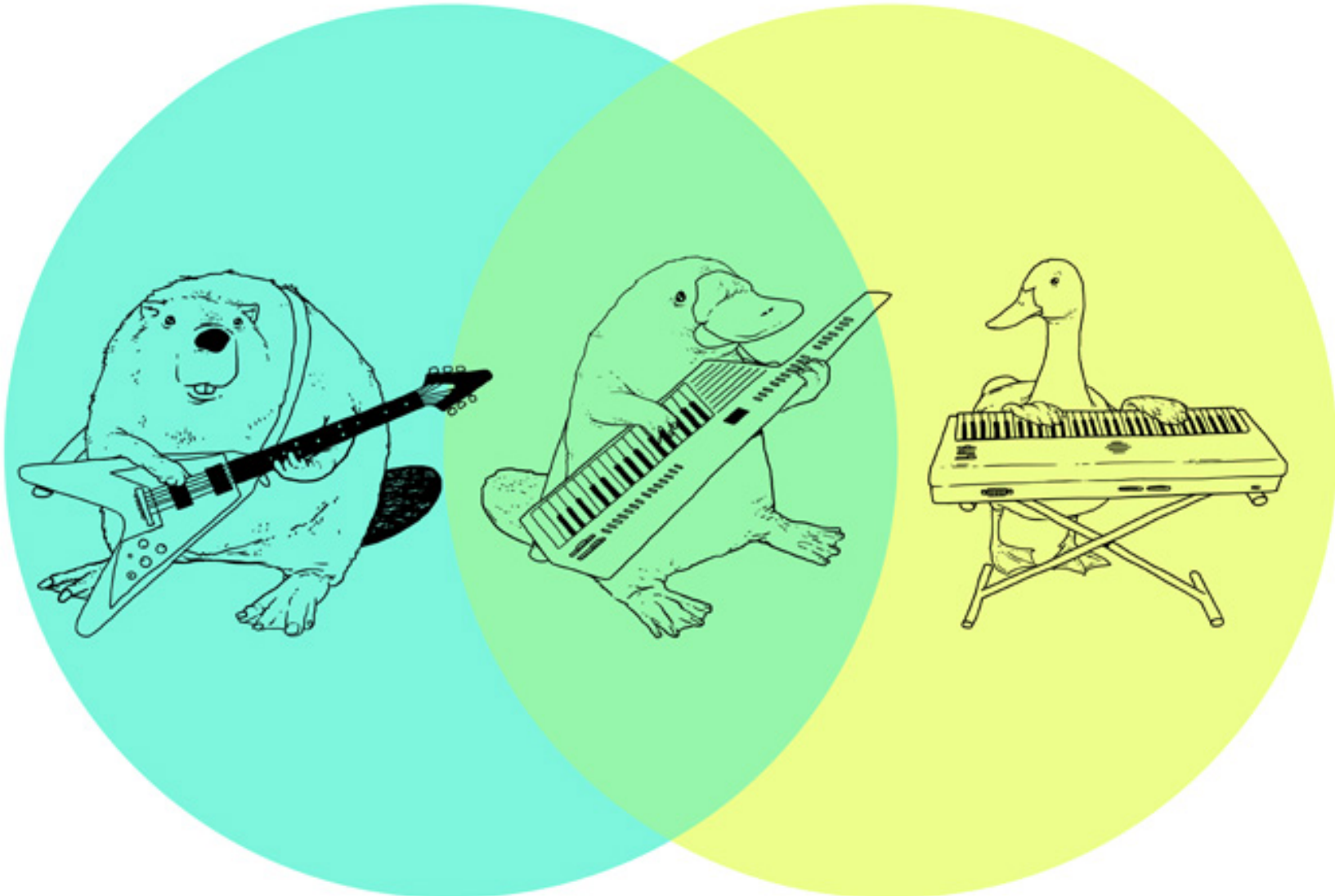


Figure 3: The beaverduck from Tenso Graphics

Calculating probabilities

- Probabilities are proportions, or the number of observations with a particular value divided by...
  - the total number of observations in a sample (n) for the sample proportion ($\hat{p}_n$)
  - the total number of outcomes in the sample space (s) for the population proportion (p)
- Proportions range from 0 (no observations/outcomes) to 1 (all observations/outcomes)
- Also may be a percentage, ranging from 0% to 100% (multiply proportion by 100)

Probability notation and calculation

$$\text{Probability}(\text{EventA}) = \text{P}(\text{A})$$

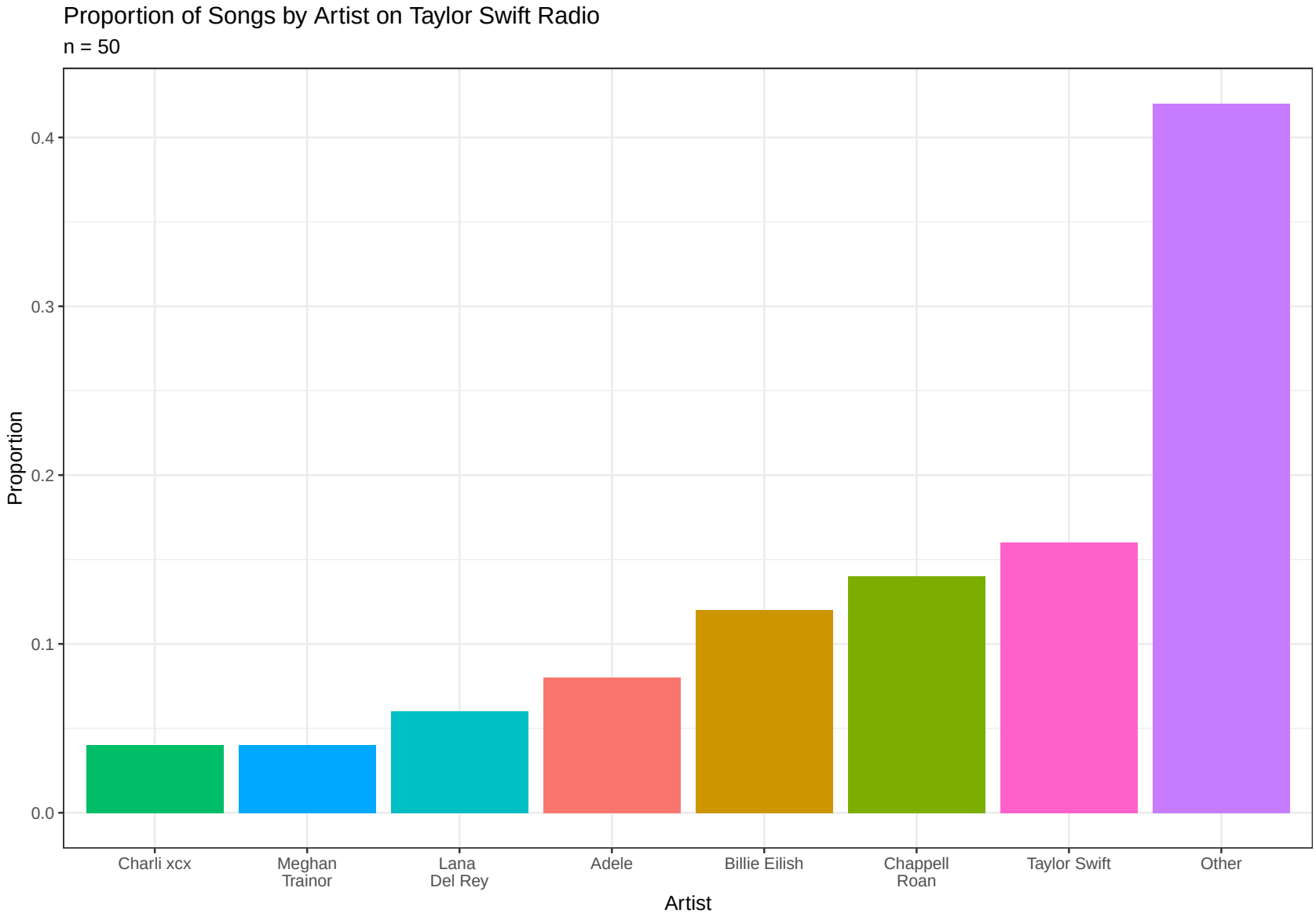
$$\begin{aligned}\text{SampleProbability}(\text{A}) &= \frac{\text{count}(\text{observation} = \text{A})}{\text{count}(\text{observationsinsample})} \\ &= \hat{p}_n\end{aligned}$$

$$\begin{aligned}\text{PopulationP}(\text{A}) &= \frac{\text{count}(\text{event} = \text{A})}{\text{count}(\text{eventsinsamplespace})} \\ &= p\end{aligned}$$

Calculating Probabilities with Complements

$$\begin{aligned} \mathrm{P}(\mathrm{S}) &= 1 \\ &= \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{A}^{\mathrm{C}}) \\ &= \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{A}^{'}) \end{aligned}$$

Proportion of Songs by Artist (Population Probability)



Calculating the Population Probability (p) for a Single Event

$$p = \text{PopulationProbability}(\text{NextSongbyChappellRoan})$$

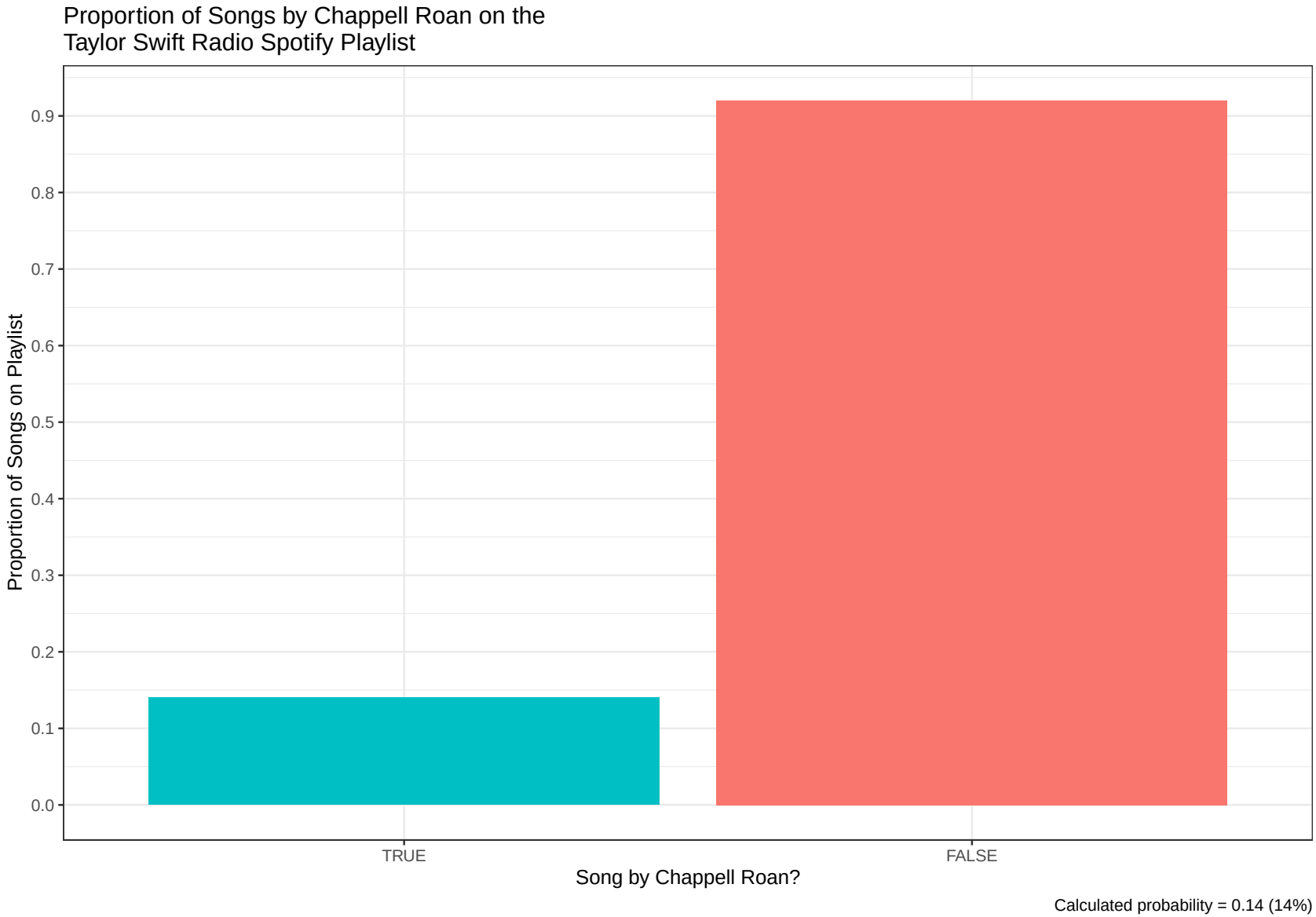
- The sample space for the population probability that the next song is by Chappell Roan when there is "true" shuffle or ***sampling with replacement*** is all songs on the playlist ($n = 50$).
- Chappell Roan has 7 songs on the playlist, so the event "The next song is by Chappell Roan" occurs 7 times within the sample space.

Calculating the Population Probability (p) for a Single Event

The population probability p of the next song being by Chappell Roan is...

$$
\begin{aligned}
p &= \mathrm{P}(\mathrm{NextSongbyChappellRoan}) \\
&= \frac{\mathrm{count}(\mathrm{SongsByChappellRoan})}{\mathrm{count}(\mathrm{TotalPossibleSongs})} \\
&= \frac{7}{50} \\
&= 0.14 \\
&= 14\%
\end{aligned}
$$

Proportion of Chappell Roan songs



Calculated probability = 0.14 (14%)

Calculating the Sample Probability ($\hat{p}_n$) for a Single Event

- The sample space for the sample probability of the next song being by Chappell Roan when there is "true" shuffle or ***sampling with replacement*** is the number of songs listened to so far ($n = 1+$).
- Each time a Chappell Roan song is played, an event is counted / recorded.

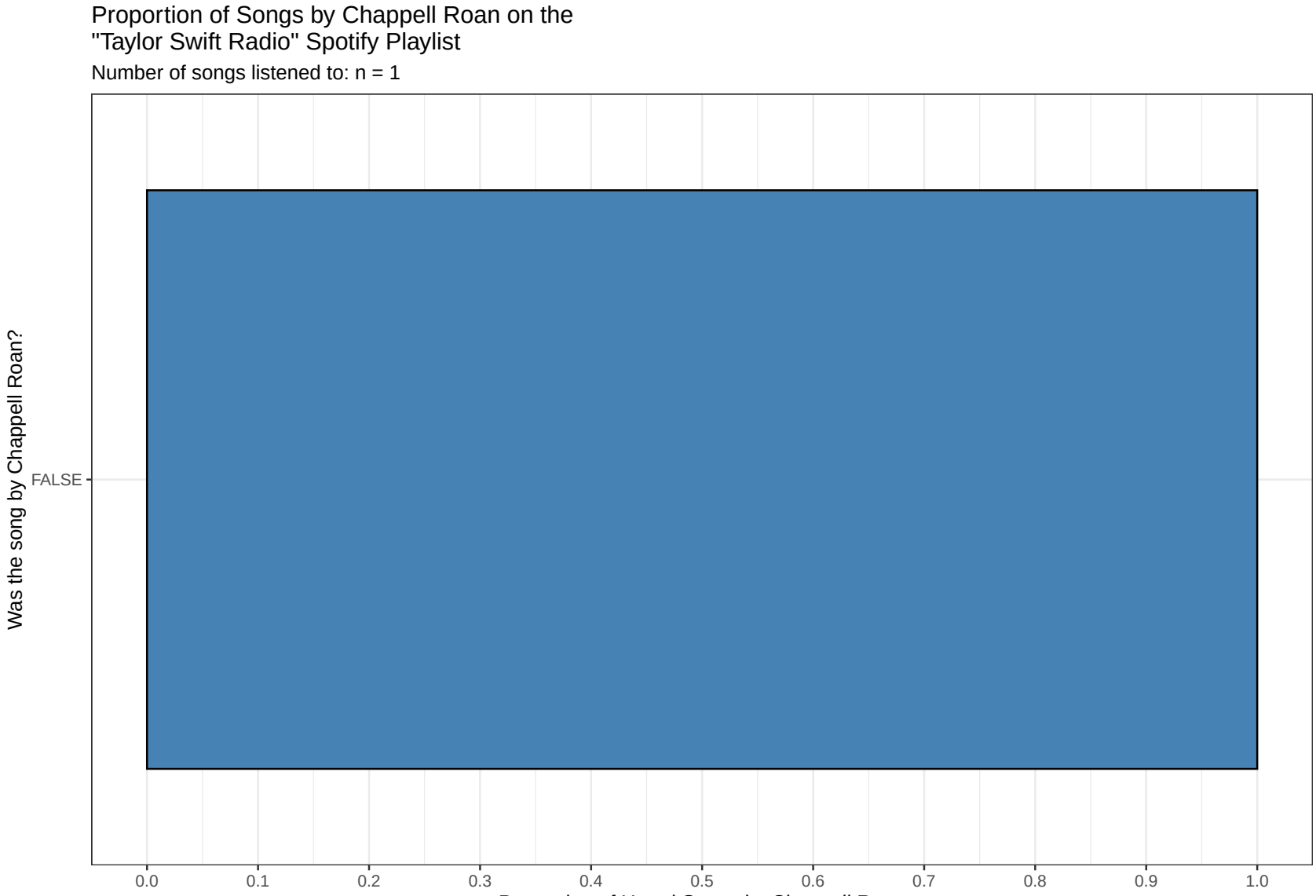
Calculating the Sample Probability ($\hat{p}_n$) for a Single Event

The population probability of the next song being by Chappell Roan is...

$$\begin{aligned}\hat{p}_n &= \mathrm{P}(\mathrm{NextSongbyChappellRoan}) \\ &= \frac{\mathrm{count}(\mathrm{SongsHeardByChappellRoan})}{\mathrm{count}(\mathrm{TotalSongsHeard})} \\ &= \frac{x}{n}\end{aligned}$$

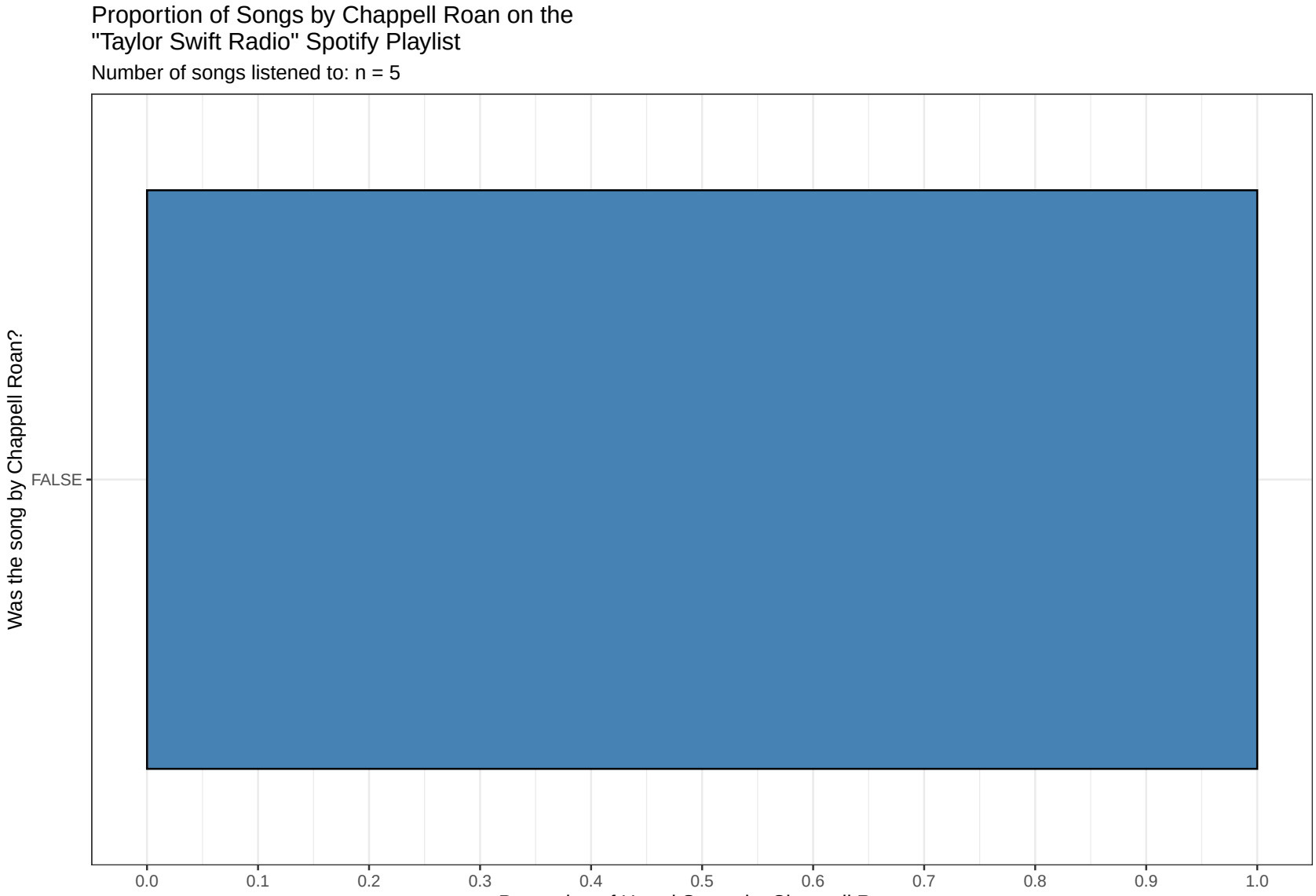
How well does a sample proportion represent the population proportion?

Should we listen to 1 song?



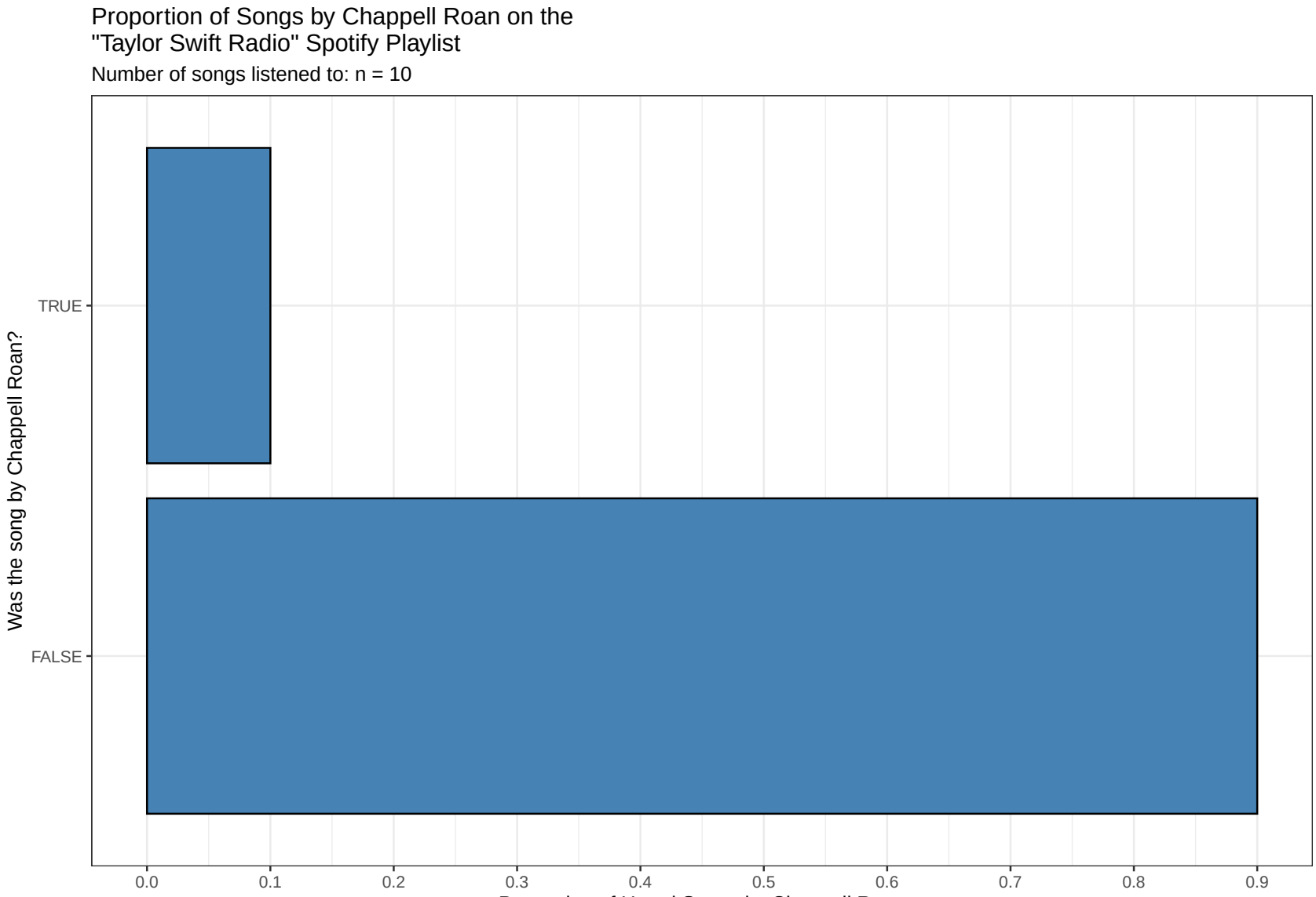
How well does the sample proportion represent the population proportion?

Should we listen to 5 songs?



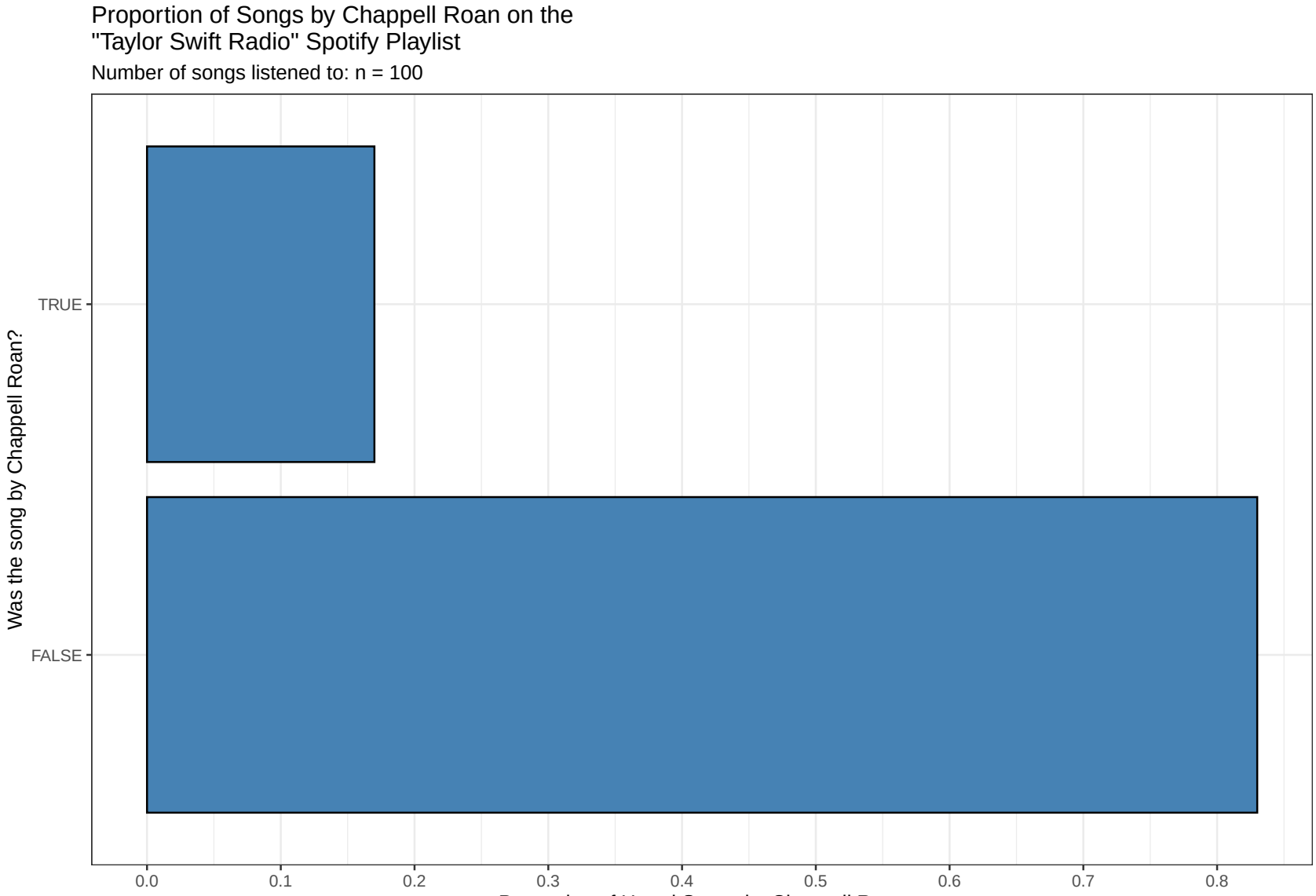
How well does the sample proportion represent the population proportion?

Should we listen to 10 songs?



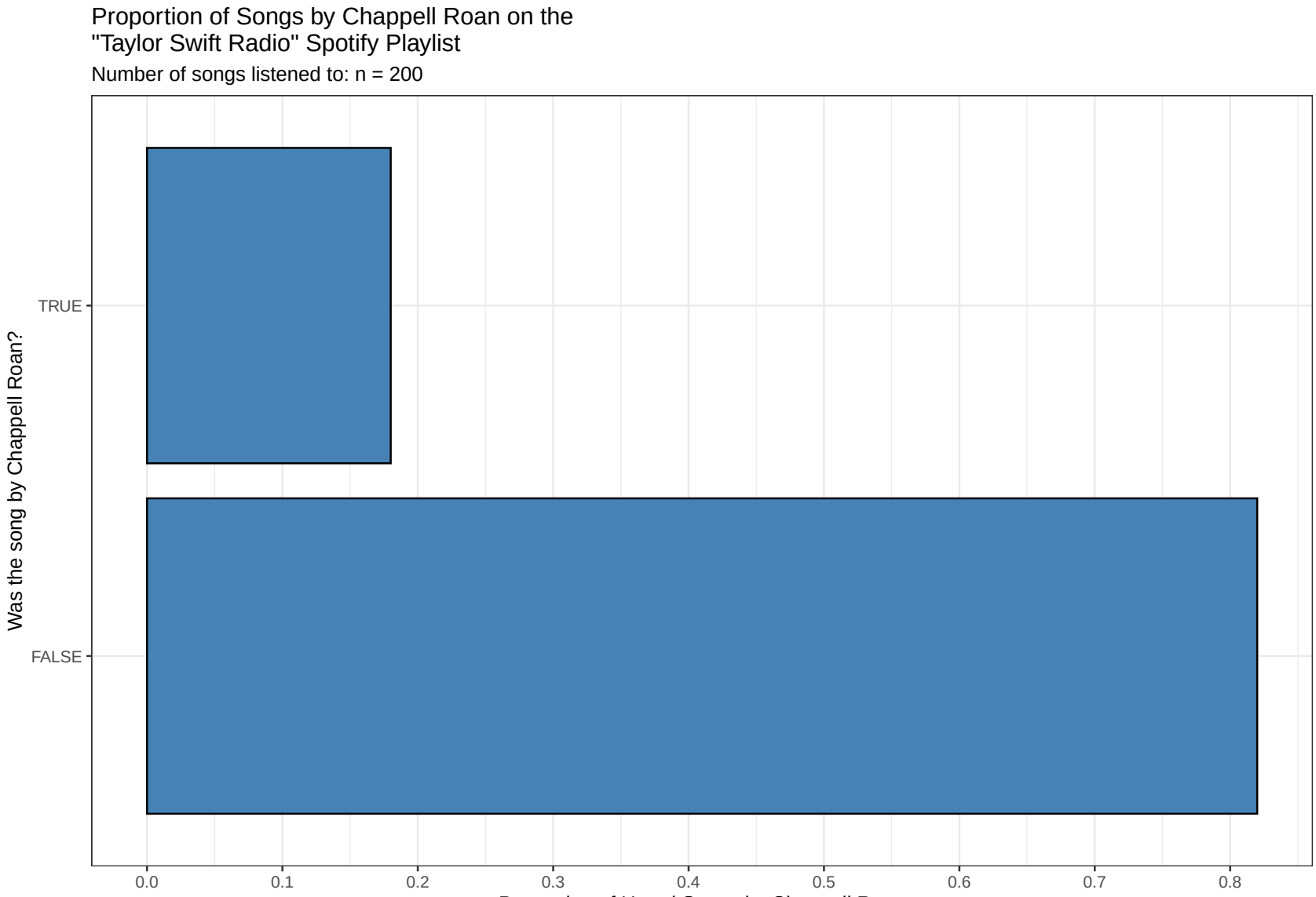
How well does the sample proportion represent the population proportion?

Should we listen to 100 songs?



How well does the sample proportion represent the population proportion?

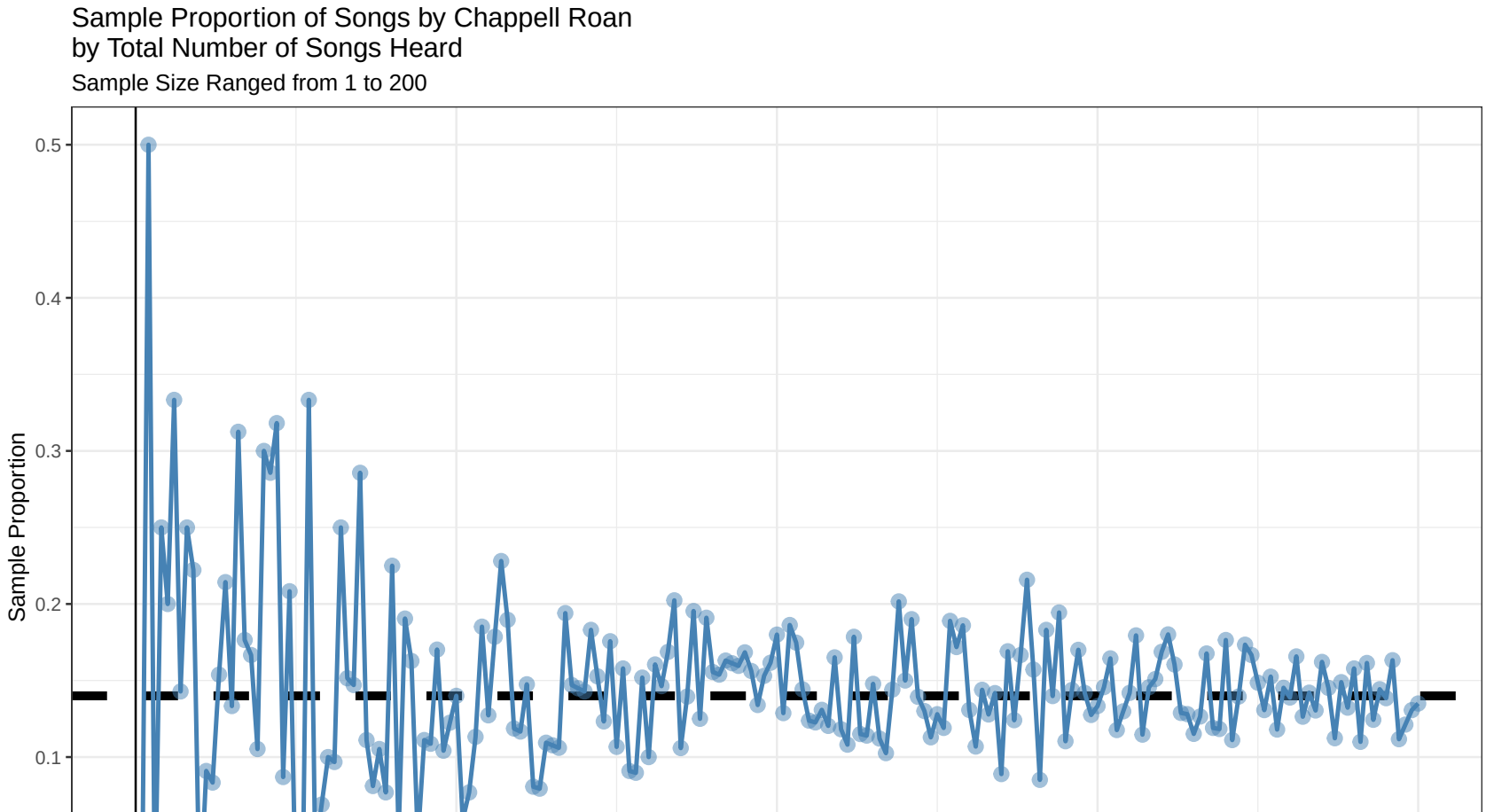
Should we listen to 200 songs?



Law of Large Numbers

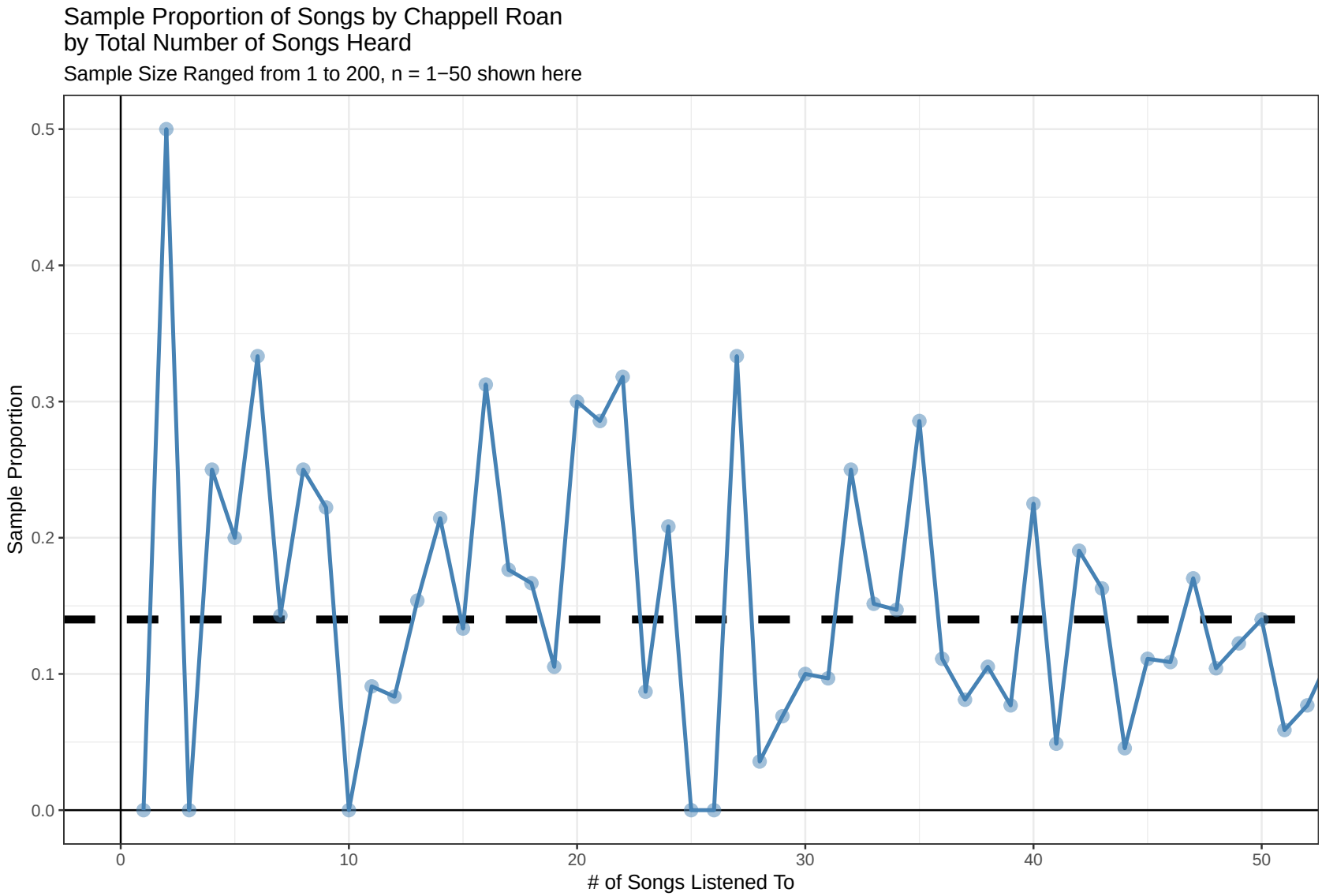
How well the sample proportion $\hat{p}_n$ represents the population proportion p depends on the size of the denominator.

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



Small Sample Size / Small Denominator = Unreliable

The sample proportion is an unreliable estimator of the population proportion when the sample size is small.



Large Sample Size / Small Denominator = Reliable Sample Proportion

The sample proportion is a reliable estimator of the population proportion when the sample size is large.