Lab 3: Probability

Julia Ferris

2023-09-21

```r
library(tidyverse)
library(openintro)
library(ggplot2)
library(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 4th Edition. You can read this by typing
## vignette('os4') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

Exercise 1

What does a streak length of 1 mean, i.e. how many hits and misses are in a streak of 1? What about a streak length of 0?

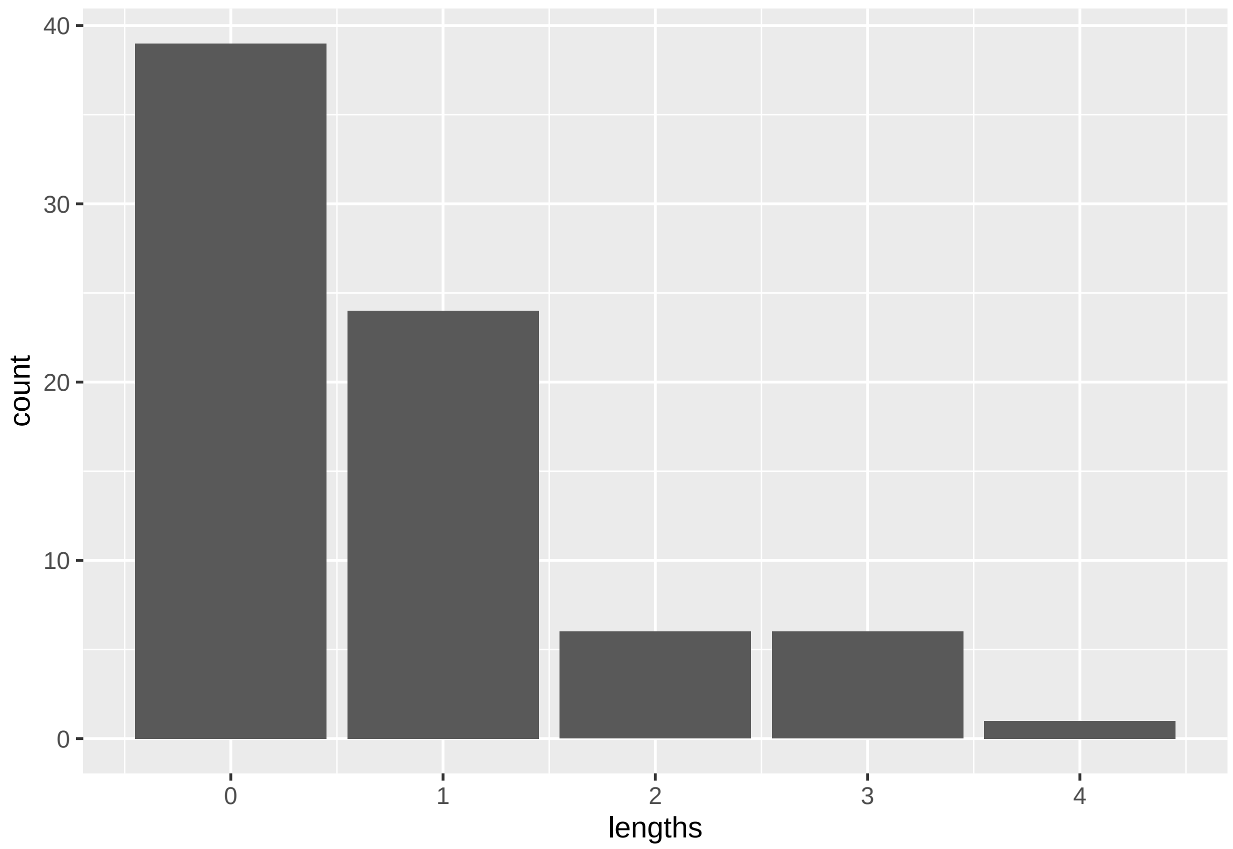
A streak length of 1 means that in the streak, one shot was a hit and one shot was a miss. A streak of length 0 means one shot was a miss.

Exercise 2

Describe the distribution of Kobe's streak lengths from the 2009 NBA finals. What was his typical streak length? How long was his longest streak of baskets? Make sure to include the accompanying plot in your answer.

This is a right-skewed distribution. His typical streak length is 0. Of streaks greater than 0, his typical streak is of length 1. His longest streak has a length of 4.

```r
kobe_streak <- calc_streak(kobe_basket$shot)
kobe_streak_df <- data.frame(lengths = c(kobe_streak))
ggplot(kobe_streak_df, aes(x = lengths)) +
  geom_bar()
```



Exercise 3

In your simulation of flipping the unfair coin 100 times, how many flips came up heads? Include the code for sampling the unfair coin in your response. Since the markdown file will run the code, and generate a new sample each time you Knit it, you should also "set a seed" before you sample. Read more about setting a seed below.

24 flips came up heads.

```
coin_outcomes <- c("heads", "tails")
set.seed(123456)
sim_unfair_coin <- sample(coin_outcomes, size = 100, replace = TRUE,
                          prob = c(0.2, 0.8))
table(sim_unfair_coin)
```

```
## sim_unfair_coin
## heads tails
##    24    76
```

Exercise 4

What change needs to be made to the sample function so that it reflects a shooting percentage of 45%? Make this adjustment, then run a simulation to sample 133 shots. Assign the output of this simulation to a new object called sim_basket.

To change the shooting percentage to 45%, the variable prob must be added with values of 0.45 for hit and 0.55 for miss.

```
shot_outcomes <- c("H", "M")
set.seed(123456)
sim_basket <- sample(shot_outcomes, size = 133, replace = TRUE, prob = c(0.45, 0.55))
table(sim_basket)
```

```
## sim_basket
##  H  M
## 54 79
```

Exercise 5

Using calc_streak, compute the streak lengths of sim_basket, and save the results in a data frame called sim_streak.

Insert any text here.

```
sim_streak <- data.frame(lengths = calc_streak(sim_basket))
glimpse(sim_streak)
```

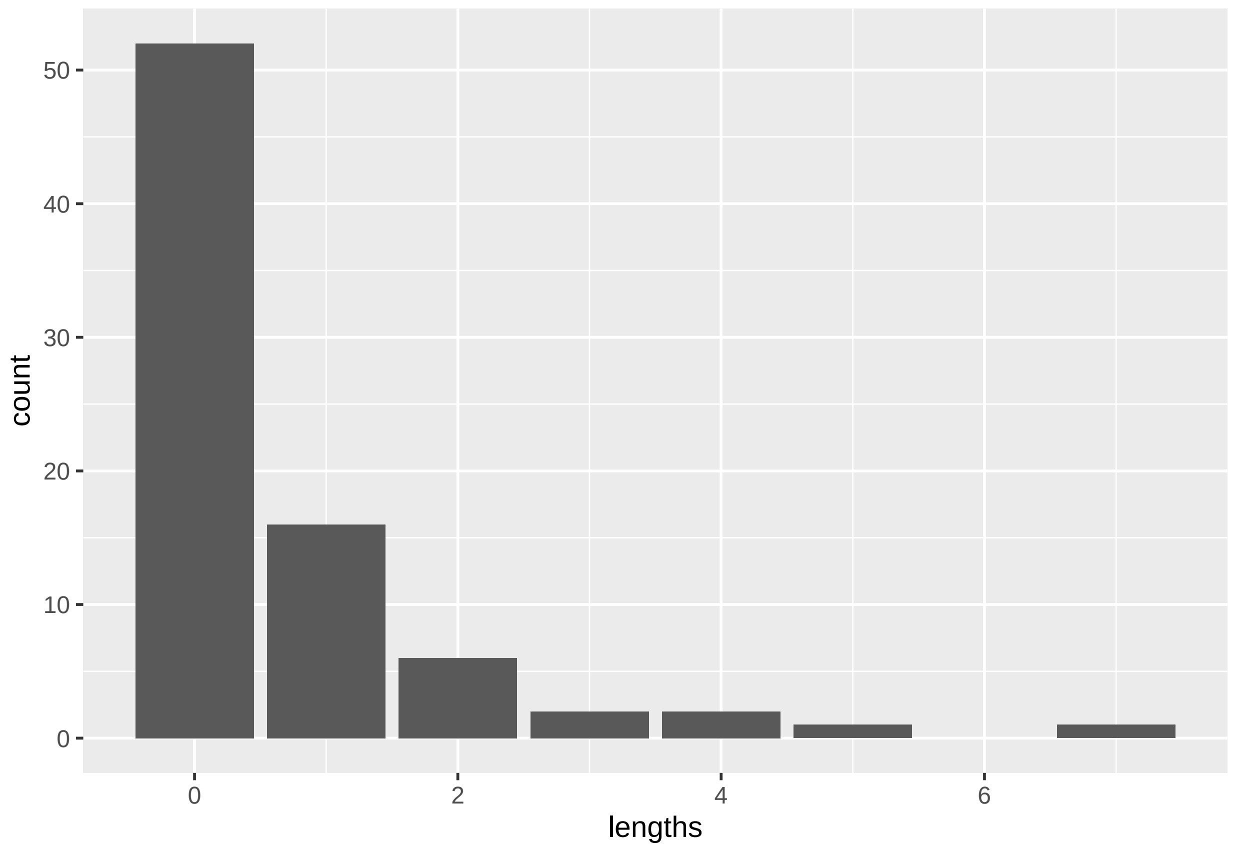
```
## Rows: 80
## Columns: 1
## $ lengths <dbl> 2, 0, 0, 0, 0, 0, 1, 7, 0, 1, 0, 0, 0, 0, 4, 0, 1, 0, 1, 4, 2,~
```

Exercise 6

Describe the distribution of streak lengths. What is the typical streak length for this simulated independent shooter with a 45% shooting percentage? How long is the player's longest streak of baskets in 133 shots? Make sure to include a plot in your answer.

This is a right-skewed distribution of streak lengths. The typical streak length is 0. For lengths of 1 or more, the typical streak length is 1. The longest streak is 7.

```
ggplot(data = as.data.frame(sim_streak), aes(x = lengths)) +
  geom_bar()
```

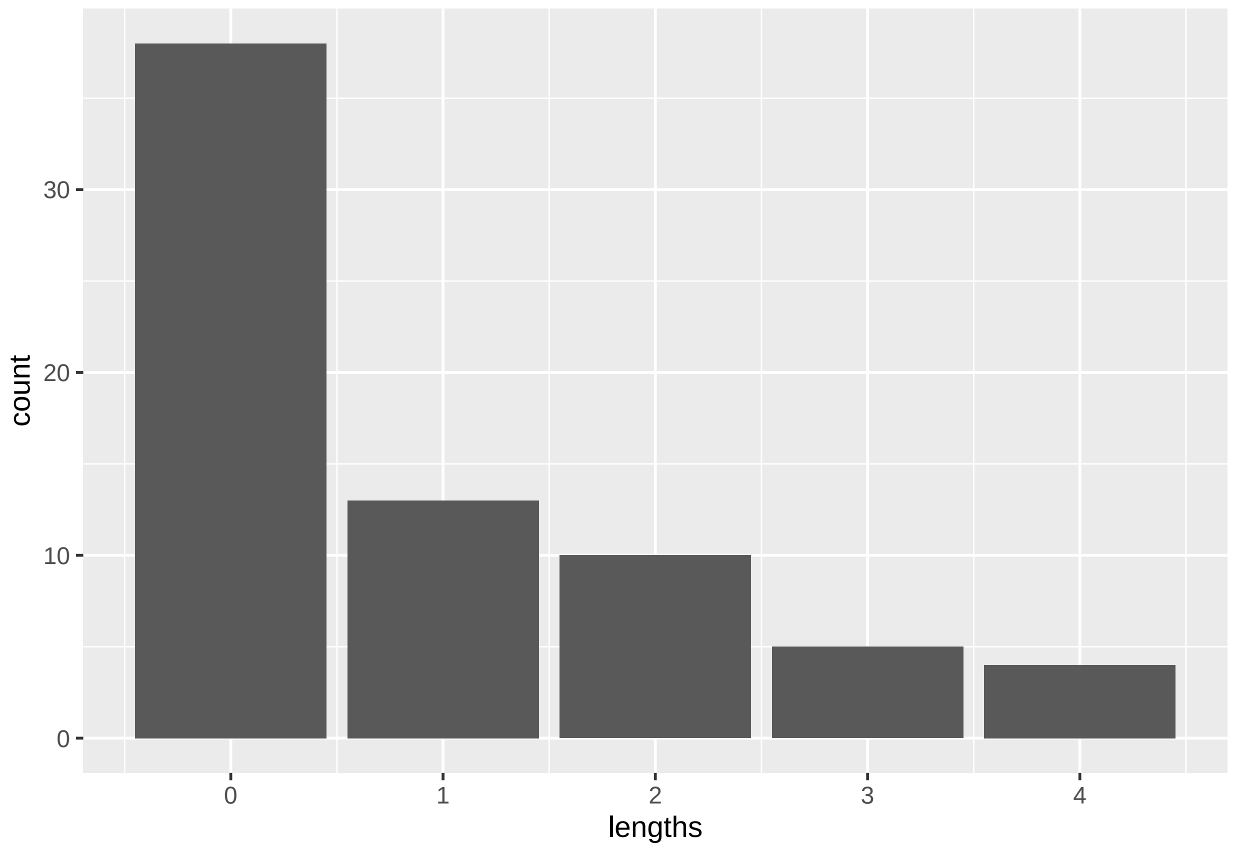


Exercise 7

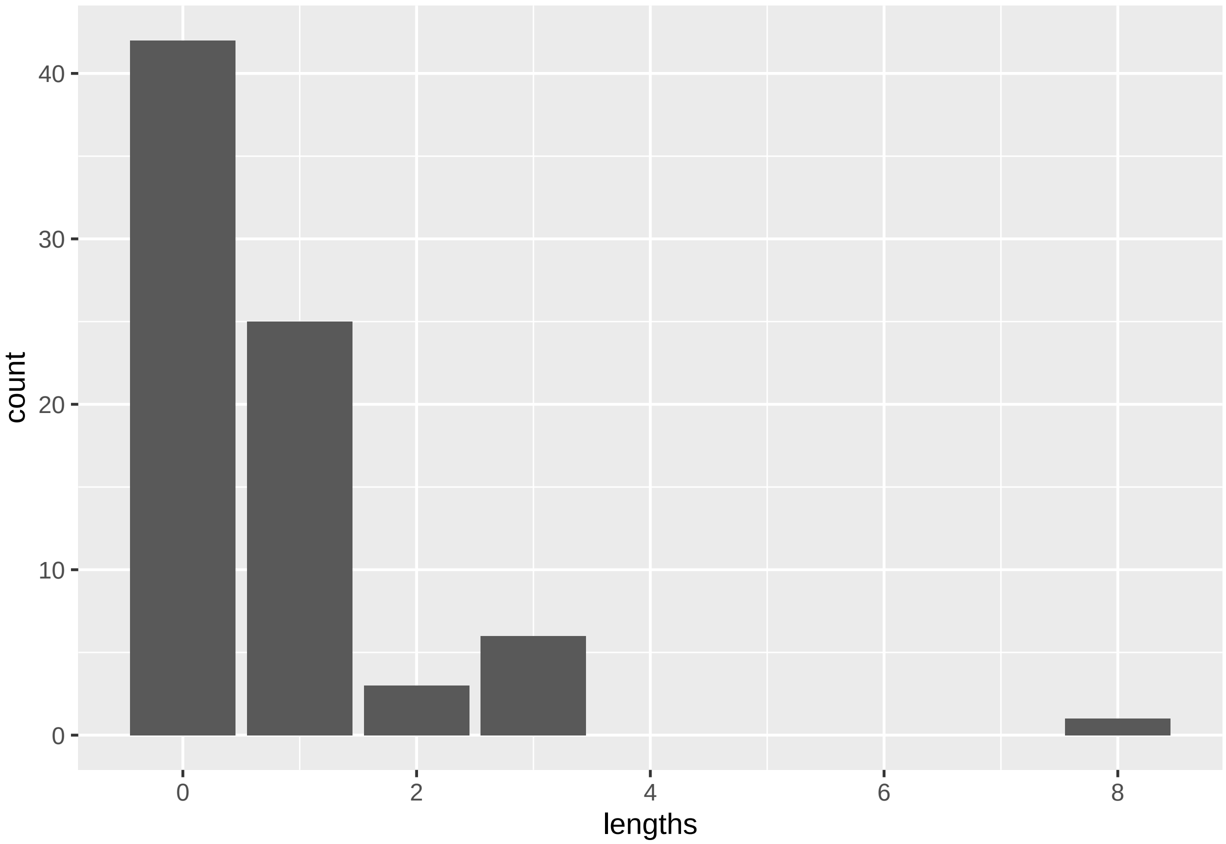
If you were to run the simulation of the independent shooter a second time, how would you expect its streak distribution to compare to the distribution from the question above? Exactly the same? Somewhat similar? Totally different? Explain your reasoning.

If I were to set the seed the same as before, then the distribution would be the same. However, I will set the seed to a new value. I expect the streak distribution to be extremely similar to the question above. The reasoning is that smaller streaks are much more likely to occur, so I expect the distribution to be right-skewed as long as the probability of a hit is 45%. The examples below confirm my assumption.

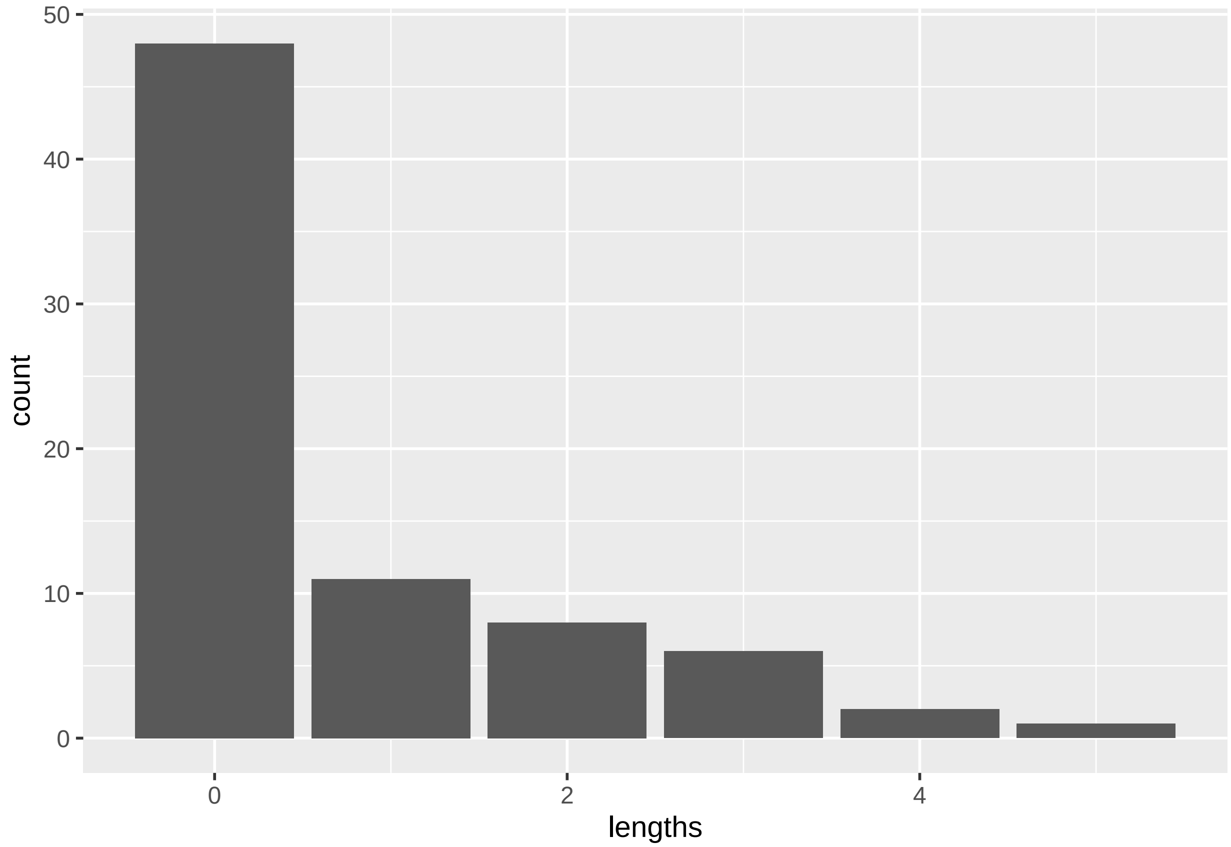
```r
set.seed(4321)
sim_basket2 <- sample(shot_outcomes, size = 133, replace = TRUE, prob = c(0.45, 0.55))
sim_streak2 <- data.frame(lengths = calc_streak(sim_basket2))
ggplot(data = as.data.frame(sim_streak2), aes(x = lengths)) +
  geom_bar()
```



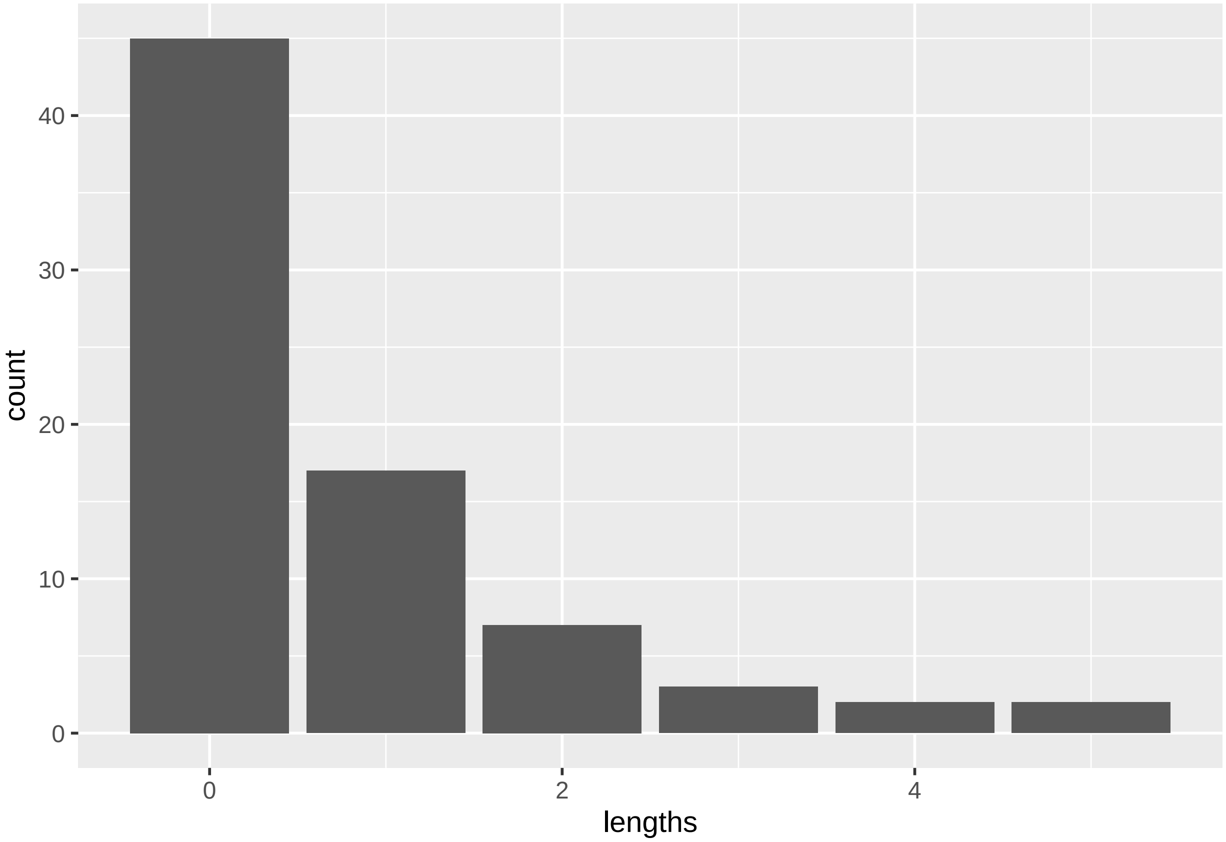
```
set.seed(85)
sim_basket2 <- sample(shot_outcomes, size = 133, replace = TRUE, prob = c(0.45, 0.55))
sim_streak2 <- data.frame(lengths = calc_streak(sim_basket2))
ggplot(data = as.data.frame(sim_streak2), aes(x = lengths)) +
  geom_bar()
```



```r
set.seed(1010)
sim_basket2 <- sample(shot_outcomes, size = 133, replace = TRUE, prob = c(0.45, 0.55))
sim_streak2 <- data.frame(lengths = calc_streak(sim_basket2))
ggplot(data = as.data.frame(sim_streak2), aes(x = lengths)) +
  geom_bar()
```



```
set.seed(2222)
sim_basket2 <- sample(shot_outcomes, size = 133, replace = TRUE, prob = c(0.45, 0.55))
sim_streak2 <- data.frame(lengths = calc_streak(sim_basket2))
ggplot(data = as.data.frame(sim_streak2), aes(x = lengths)) +
  geom_bar()
```

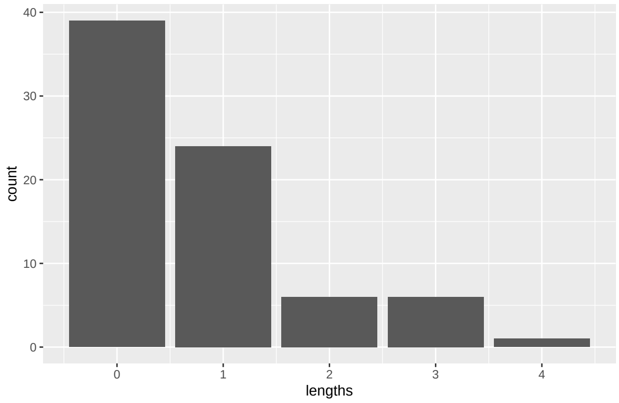


Exercise 8

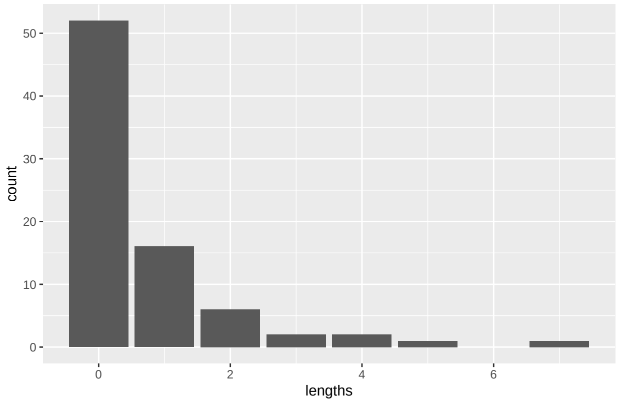
How does Kobe Bryant's distribution of streak lengths compare to the distribution of streak lengths for the simulated shooter? Using this comparison, do you have evidence that the hot hand model fits Kobe's shooting patterns? Explain.

The distributions are both right-skewed, and they are very similar. The only difference is that in Kobe Bryant's distribution, streaks of lengths 2 and 3 had the same frequency. Due to the minor differences, I do not have evidence that the hot hand model fits Kobe's shooting patterns. If he fit the hot hand model, he should have a higher frequency of lengths 2, 3, and more. This would make sense because making one basket should increase the likelihood of making another basket. For example, if his accuracy is 45% for the first basket and 60% for the second basket, then he should have some lengths of 1 but more lengths of 2. If his accuracy for making a third basket is 70%, then he should have a higher frequency for a length of 3. However, this is not the case. Also, an independent shooter showed decreasing frequencies with longer streak lengths, just like Kobe Bryant. Since both had a similar pattern, it appears that he does not fit the hot hand model.

```r
ggplot(data = as.data.frame(kobe_streak_df), aes(x = lengths)) +
  geom_bar() +
  labs(title = "Kobe Bryant's Streaks")
ggplot(data = as.data.frame(sim_streak), aes(x = lengths)) +
  geom_bar() +
  labs(title = "Simulated Shooter's Streaks")
```



Kobe Bryant's Streaks



Simulated Shooter's Streaks