Probability

Zachary Palmore

2020-09-13

```
library(tidyverse)
library(openintro)
library(ggpubr)
```

Pre-exercise

Checking for access to the kobe_basket data.

```
glimpse(kobe_basket)
```

```
## Rows: 133
## Columns: 6
## $ vs          <fct> ORL, ORL, ORL, ORL, ORL, ORL, ORL, ORL, ORL, ORL, ORL, ...
## $ game        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ quarter     <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3...
## $ time        <fct> 9:47, 9:07, 8:11, 7:41, 7:03, 6:01, 4:07, 0:52, 0:00, 6...
## $ description <fct> Kobe Bryant makes 4-foot two point shot, Kobe Bryant mi...
## $ shot        <chr> "H", "M", "M", "H", "H", "M", "M", "M", "M", "H", "H", ...
```

This data set contains 133 rows with 6 variables.

Verifying the first 9 rows of data. It should display the sequence of shots as:

- H M | M | H H M | M | M | M

indicating hits (H) and misses (M).

```
kobe_basket[1:9, 1:6]
```

```
## # A tibble: 9 x 6
##   vs     game quarter time  description                                    shot
##   <fct> <int> <fct>   <fct> <fct>                                          <chr>
## 1 ORL       1 1       9:47  Kobe Bryant makes 4-foot two point shot        H
## 2 ORL       1 1       9:07  Kobe Bryant misses jumper                      M
## 3 ORL       1 1       8:11  Kobe Bryant misses 7-foot jumper               M
## 4 ORL       1 1       7:41  Kobe Bryant makes 16-foot jumper (Derek Fishe~ H
## 5 ORL       1 1       7:03  Kobe Bryant makes driving layup                H
## 6 ORL       1 1       6:01  Kobe Bryant misses jumper                      M
## 7 ORL       1 1       4:07  Kobe Bryant misses 12-foot jumper              M
## 8 ORL       1 1       0:52  Kobe Bryant misses 19-foot jumper              M
## 9 ORL       1 1       0:00  Kobe Bryant misses layup                       M
```

It does and furthermore, it is identical to the expected data set where every row displays a shot made by Kobe.

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 one shot was taken, and it was made, then another shot was taken, and it missed. In other words, there were two shots taken. During which, the first shot made it into the basket. The other shot missed, thus ending the streak.

A streak length of 0 simply means a shot was taken and it missed. There were no further shots made in this streak because to have a streak, one must continue to make it into the basket.

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.

Using the function calc_streak from the openintro package, the streaks are calculated then stored in a new data frame.

```r
streaks <- calc_streak(kobe_basket$shot)
streaks[1:5, 1]
```

```
## [1] 1 0 2 0 0
```

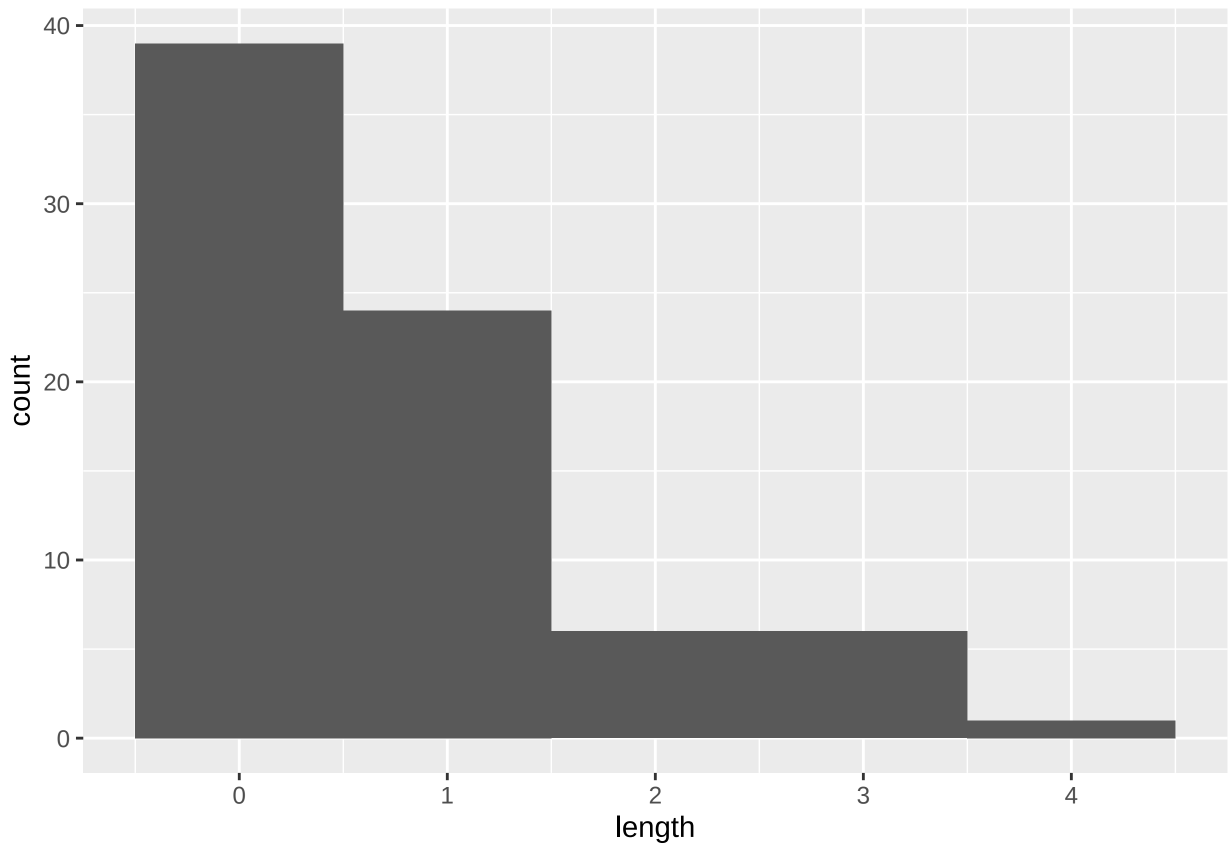
This data frame contains one column of the name "length" that indicates the length of each of Kobe's streaks from the 2009 NBA finals.

```r
summary(streaks$length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.7632  1.0000  4.0000
```

The range is 4, with smallest streak length at zero and longest at 4. From these summary statistics we can also see the mean is less than 1 (specifically 0.76) with a median of 0. The data is heavily concentrated towards smaller streak lengths with a thin tail stretching toward the right.

```r
ggplot(data = streaks, aes(x = length)) +
  geom_histogram(binwidth = 1)
```



Here, it is much easier to see a right skewed distribution with the most frequent streak length at 0.

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.

Simulating the coin:

```
coin_outcomes <- c("heads", "tails")
sample(coin_outcomes, size = 1, replace = TRUE)
```

```
## [1] "heads"
```

```
sim_fair_coin <- sample(coin_outcomes, size = 100, replace = TRUE)
sim_fair_coin
```

```
##   [1] "tails" "heads" "heads" "tails" "tails" "heads" "tails" "heads" "heads"
##  [10] "tails" "heads" "heads" "tails" "tails" "heads" "heads" "heads" "heads"
##  [19] "tails" "heads" "heads" "heads" "heads" "tails" "heads" "tails" "heads"
##  [28] "heads" "heads" "heads" "heads" "heads" "heads" "tails" "tails" "heads"
##  [37] "heads" "heads" "heads" "heads" "heads" "tails" "tails" "heads" "tails"
##  [46] "tails" "tails" "tails" "heads" "tails" "tails" "tails" "tails" "heads"
##  [55] "heads" "heads" "tails" "heads" "heads" "tails" "heads" "heads" "tails"
##  [64] "heads" "tails" "tails" "tails" "heads" "tails" "heads" "heads" "tails"
##  [73] "tails" "heads" "tails" "heads" "tails" "tails" "heads" "tails" "tails"
##  [82] "tails" "heads" "heads" "tails" "heads" "tails" "heads" "tails" "tails"
##  [91] "tails" "heads" "tails" "tails" "tails" "tails" "tails" "tails" "tails"
## [100] "tails"
```

```r
table(sim_fair_coin)
```

```
## sim_fair_coin
## heads tails
##    50    50
```

Simulating the unfair coin with a set seed.

```r
set.seed(10312020)
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
##    28    72
```

In this simulation of an unfair coin, 28 flips were heads and 72 were tails.

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.

Changes might include adjusting the probability of the shots to have it weighted towards missing at 0.55, and hits at 0.45 given Kobe's shooting percentage of 45%. Another good change is the adjustment of the sample size to 133 to run a more realistic simulation.

```r
set.seed(11262020)
possibilities <- c("H", "M")
sim_basket <- sample(possibilities,
                     size = 133,
                     replace = TRUE,
                     prob = c(0.45, 0.55))
table(sim_basket)
```

```
## sim_basket
##  H  M
## 49 84
```

With this simulation there are 84 misses and 49 hits.

Exercise 5

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

```
sim_streak <- calc_streak(sim_basket)
table(sim_streak)
```

```
## sim_streak
##  0  1  2  3  4
## 54 18  9  3  1
```

This still shows a lot of misses with more at 0 than any other streak length.

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.

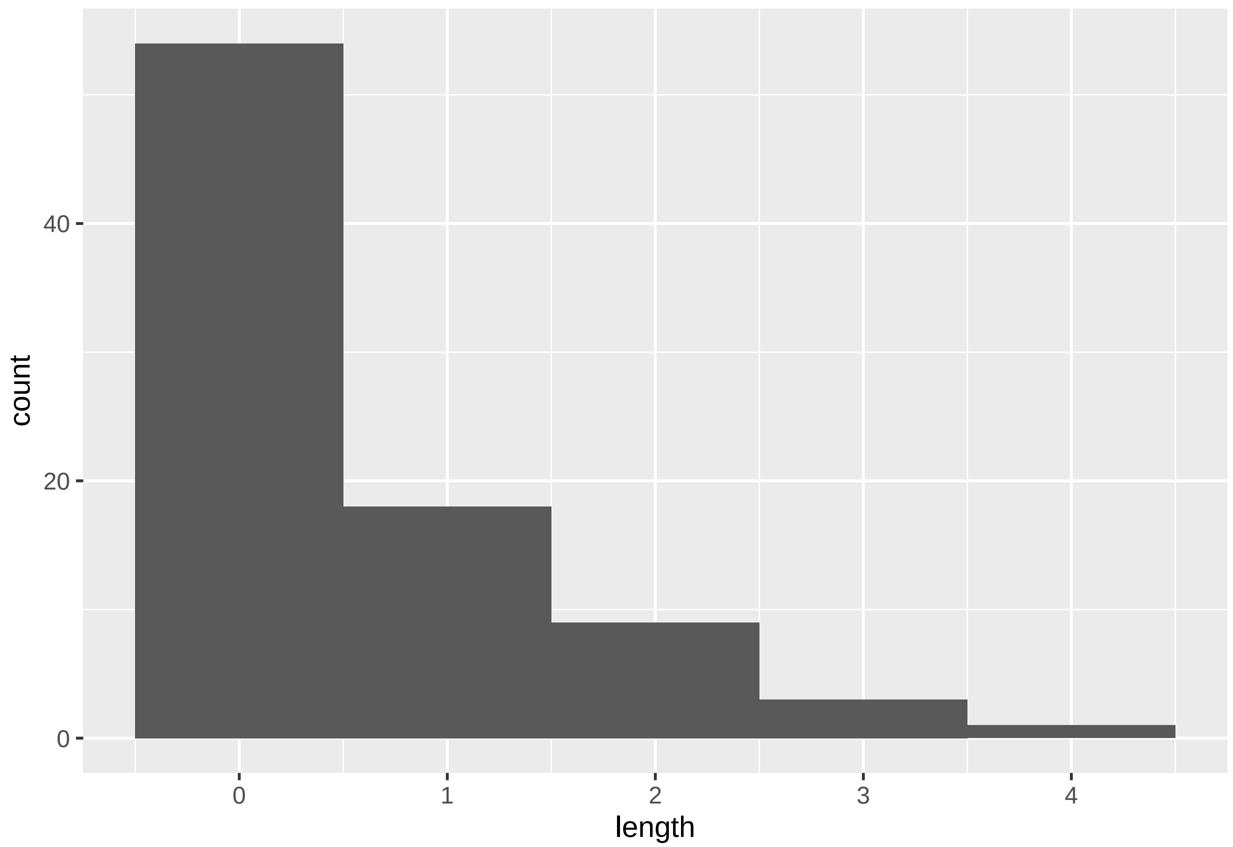
We already know the distribution is going to be concentrated on the misses given the frequencies of the sim_streak table above.

```
summary(sim_streak)
```

```
##      length
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.5765
##  3rd Qu.:1.0000
##  Max.   :4.0000
```

Here we can see the range is the same with the smallest streak at 0 and longest still at 4. The median is also zero again but the mean has changed. In this independent shooter scenario with a 45% shooting percentage, the mean is smaller than the real version with Kobe.

```
ggplot(data = sim_streak, aes(x = length)) +
  geom_histogram(binwidth = 1)
```



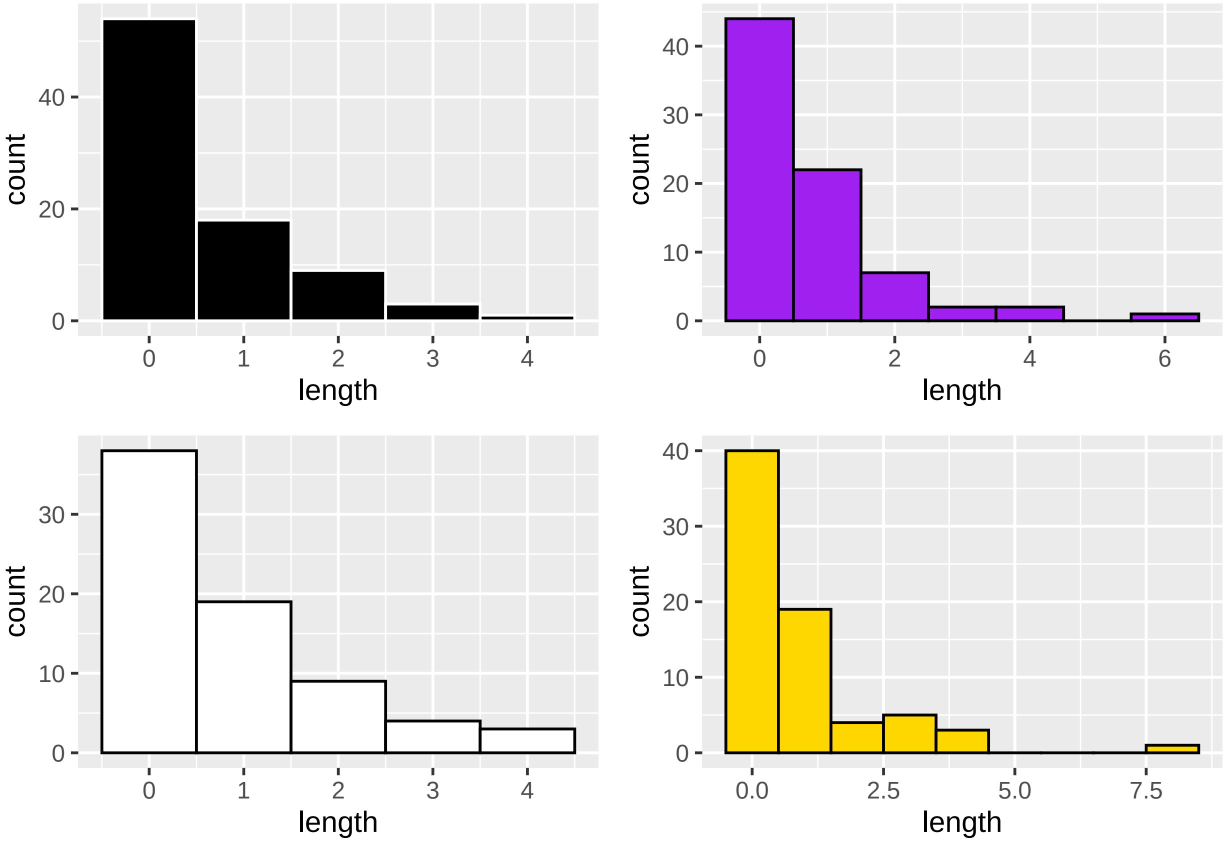
This histogram shows a right skewed distribution with a mode of 0 again. However, in this distribution the streak length of 0 occurs much more frequently than in the original shots taken with Kobe.

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.

It should be somewhat similar because the probabilities are the same and the random selection process of values of H or M follows the same rules. However, I will test this with 3 more graphs.

```r
# original sim
set.seed(11262020)
possibilities <- c("H", "M")
sim_basket <- sample(possibilities,
                     size = 133,
                     replace = TRUE,
                     prob = c(0.45, 0.55))
sim_streak <- calc_streak(sim_basket)
plot1 <- ggplot(data = sim_streak, aes(x = length)) +
  geom_histogram(binwidth = 1, fill = "black", color = "white")
# next plot
set.seed(12252020)
possibilities <- c("H", "M")
sim_basket2 <- sample(possibilities,
                      size = 133,
                      replace = TRUE,
                      prob = c(0.45, 0.55))
sim_streak2 <- calc_streak(sim_basket2)
plot2 <- ggplot(data = sim_streak2, aes(x = length)) +
  geom_histogram(binwidth = 1, fill = "purple", color = "black")
# Next plot
set.seed(01012021)
possibilities <- c("H", "M")
sim_basket3 <- sample(possibilities,
                      size = 133,
                      replace = TRUE,
                      prob = c(0.45, 0.55))
sim_streak3 <- calc_streak(sim_basket3)
plot3 <- ggplot(data = sim_streak3, aes(x = length)) +
  geom_histogram(binwidth = 1, fill = "white", color = "black")
# next plot
set.seed(02142021)
possibilities <- c("H", "M")
sim_basket4 <- sample(possibilities,
                      size = 133,
                      replace = TRUE,
                      prob = c(0.45, 0.55))
sim_streak4 <- calc_streak(sim_basket4)
plot4 <- ggplot(data = sim_streak4, aes(x = length)) +
  geom_histogram(binwidth = 1, fill = "gold", color = "black")
# Bring them all together
ggarrange(plot1, plot2, plot3, plot4)
```



While there are two independent shooters with slightly longer streak lengths, all distributions remain right skewed with modes of 0.

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.

It appears as though we do not have strong evidence for the hot hand model of Kobe's shooting patterns.

To review, the sims streak data were gathered into the same data frame along with Kobe's.

```
# combining sim stats and creating a new variable
# name with kobe streak data
sims_streaks <- rbind(sim_streak, sim_streak2, sim_streak3, sim_streak4)
kobe_streak <- streaks
```

Their statistics look like this;

```
summary(sims_streaks)
```

```
##      length
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.7403
##  3rd Qu.:1.0000
##  Max.   :8.0000
```

```
summary(kobe_streak)
```

```
##      length
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.7632
##  3rd Qu.:1.0000
##  Max.   :4.0000
```

```
# checking if the mean was computed the desired way
 sum_sim_means <- (mean(sim_streak$length) +
    mean(sim_streak2$length) +
  mean(sim_streak3$length) +
    mean(sim_streak4$length))
 sim_means <- sum_sim_means/4
 sim_means
```

```
## [1] 0.7477867
```

Although Kobe has a marginally higher mean, the independent shooter simulation has a much higher max streak length at 8 hits to Kobe's max of 4. There were more misses than hits in any simulation run, however, Kobe's streaks also follow this same pattern.

...