Lab3 - Probability

Jawaid Hakim

2022-09-24

Contents


```
knitr::opts_chunk$set(warning = FALSE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

Your investigation will focus on the performance of one player: Kobe Bryant of the Los Angeles Lakers. His performance against the Orlando Magic in the 2009 NBA Finals earned him the title Most Valuable Player and many spectators commented on how he appeared to show a hot hand. The data file we'll use is called kobe_basket.

```
glimpse(kobe_basket)
```

```
## Rows: 133
## Columns: 6
## $ vs          <fct> ORL, 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, 1~
## $ quarter     <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3~
## $ time        <fct> 9:47, 9:07, 8:11, 7:41, 7:03, 6:01, 4:07, 0:52, 0:00, 6:35~
## $ description <fct> Kobe Bryant makes 4-foot two point shot, Kobe Bryant misse~
## $ shot        <chr> "H", "M", "M", "H", "H", "M", "M", "M", "M", "H", "H", "H"~
```

1 Exercise 1

Just looking at the string of hits and misses, it can be difficult to gauge whether or not it seems like Kobe was shooting with a hot hand. One way we can approach this is by considering the belief that hot hand shooters tend to go on shooting streaks. For this lab, we define the length of a shooting streak to be the number of consecutive baskets made until a miss occurs.

For example, in Game 1 Kobe had the following sequence of hits and misses from his nine shot attempts in the first quarter:

H M | M | H H M | M | M | M

You can verify this by viewing the first 9 rows of the data in the data viewer.

Within the nine shot attempts, there are six streaks, which are separated by a "|" above. Their lengths are one, zero, two, zero, zero, zero (in order of occurrence).

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?

The length of a streak is the number of hits in that streak and can be represented by the regular expression H*M - i.e. 0 or more hits followed by 1 miss. A streak of length 1 contains exactly 1 hit followed by 1 miss. A streak of length 0 contains 0 hits and 1 miss.

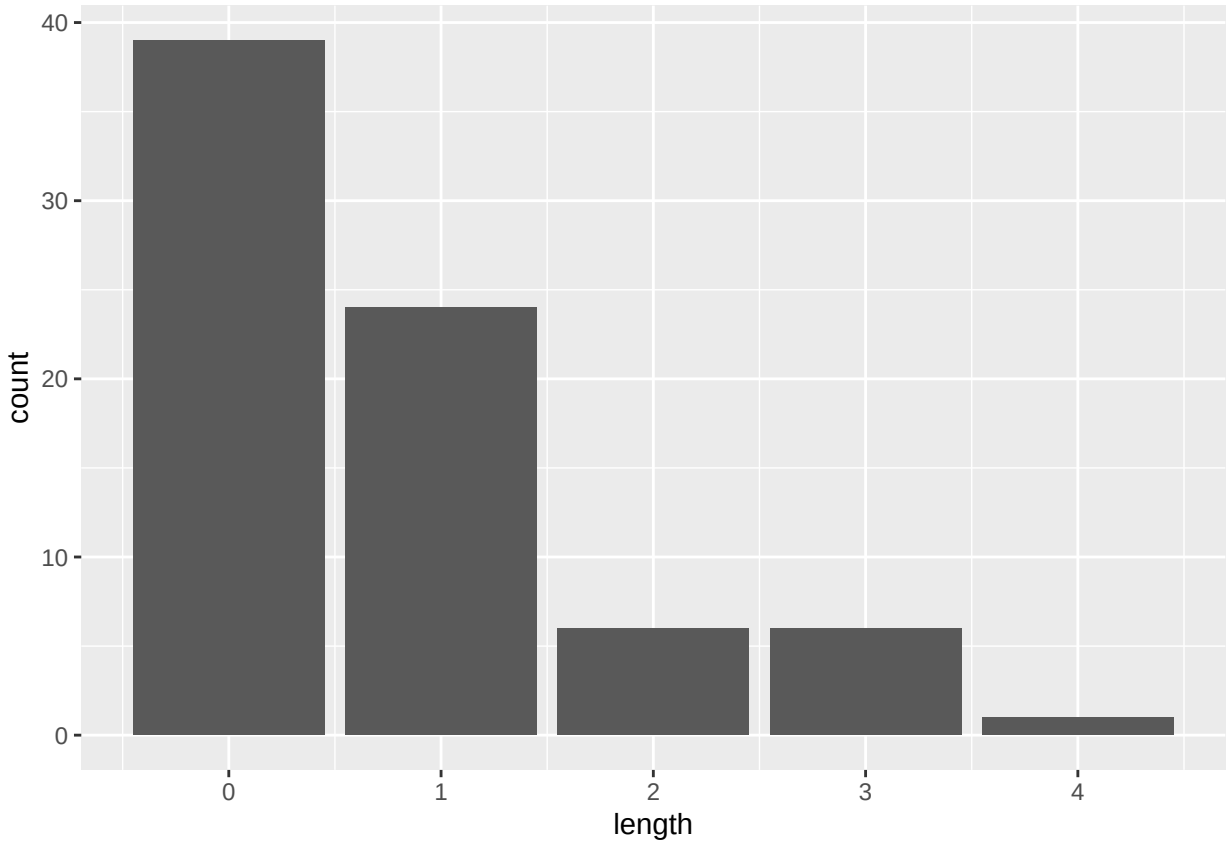
2 Exercise 2

Counting streak lengths manually for all 133 shots would get tedious, so we'll use the custom function calc_streak to calculate them, and store the results in a data frame called kobe_streak as the length variable.

```
kobe_streak <- calc_streak(kobe_basket$shot)
```

We can then take a look at the distribution of these streak lengths.

```
ggplot(data = kobe_streak, aes(x = length)) +
  geom_bar()
```



> 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.

There is a clear peak (streaks of length 0) on left and a tail (streaks of length 4) on the right. This is not a normal distribution, looks like a *unimodal* distribution.

The typical streak (mode) was of length 0 - i.e. 0 hits. The longest (max) streak was of length 4.

```
my.mode <- function (v) {
    uv <- unique(v)
    uv[which.max(tabulate(match(v, uv)))]
}

my.mode(kobe_streak$length)
```

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

```
max(kobe_streak$length)
```

```
## [1] 4
```

3 Exercise 3

Since there are only two elements in coin_outcomes, the probability that we "flip" a coin and it lands heads is 0.5. Say we're trying to simulate an unfair coin that we know only lands heads 20% of the time. We can adjust for this by adding an argument called prob, which provides a vector of two probability weights.

```
coin_outcomes <- c('heads', 'tails')
set.seed(1981)
sim_unfair_coin <- sample(coin_outcomes,
                          size = 100,
                          replace = TRUE,
                          prob = c(0.2, 0.8))
```

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.

In simulation of flipping the unfair coin 100 times, 14 flips came up heads.

```
set.seed(1981)

coin_outcomes <- c('heads', 'tails')

sim_unfair_coin <- sample(coin_outcomes,
                          size = 100,
                          replace = TRUE,
                          prob = c(0.2, 0.8)) # P(H) = 20%, P(T) = 80%

length(sim_unfair_coin[sim_unfair_coin == 'heads'])
```

```
## [1] 14
```

4 Exercise 4

Simulating a basketball player who has independent shots uses the same mechanism that you used to simulate a coin flip. To simulate a single shot from an independent shooter with a shooting percentage of 50% you can type:

```
shot_outcomes <- c("H", "M")           # H = Hit, M = Miss
sim_basket <- sample(shot_outcomes,
                     size = 1,
                     replace = TRUE)
```

To make a valid comparison between Kobe and your simulated independent shooter, you need to align both their shooting percentage and the number of attempted shots.

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.

```
sim_basket <- sample(shot_outcomes,
                     size = 133,
                     replace = TRUE,
                     prob = c(0.45, 0.55)) # P(H) = 45%, P(M) = 55%
sim_basket
```

```
##   [1] "M" "M" "M" "H" "M" "H" "M" "M" "M" "H" "H" "H" "M" "M" "M" "M" "M" "H"
##  [19] "M" "M" "M" "M" "M" "M" "M" "M" "H" "M" "H" "M" "H" "H" "M" "M" "M" "H"
##  [37] "M" "H" "M" "M" "H" "H" "M" "M" "M" "M" "M" "H" "M" "H" "H" "H" "M" "H"
##  [55] "M" "H" "M" "M" "H" "H" "M" "H" "M" "M" "M" "M" "M" "M" "M" "H" "M" "H"
##  [73] "H" "M" "H" "H" "H" "M" "H" "M" "M" "M" "M" "M" "M" "H" "M" "H" "H" "M" "M"
##  [91] "M" "H" "M" "M" "H" "M" "M" "H" "H" "H" "M" "M" "M" "M" "H" "H" "M" "M" "H"
## [109] "H" "M" "M" "M" "H" "M" "M" "M" "M" "H" "M" "H" "M" "H" "H" "M" "M" "H" "M"
## [127] "H" "H" "M" "H" "M" "M" "M"
```

5 Exercise 5

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

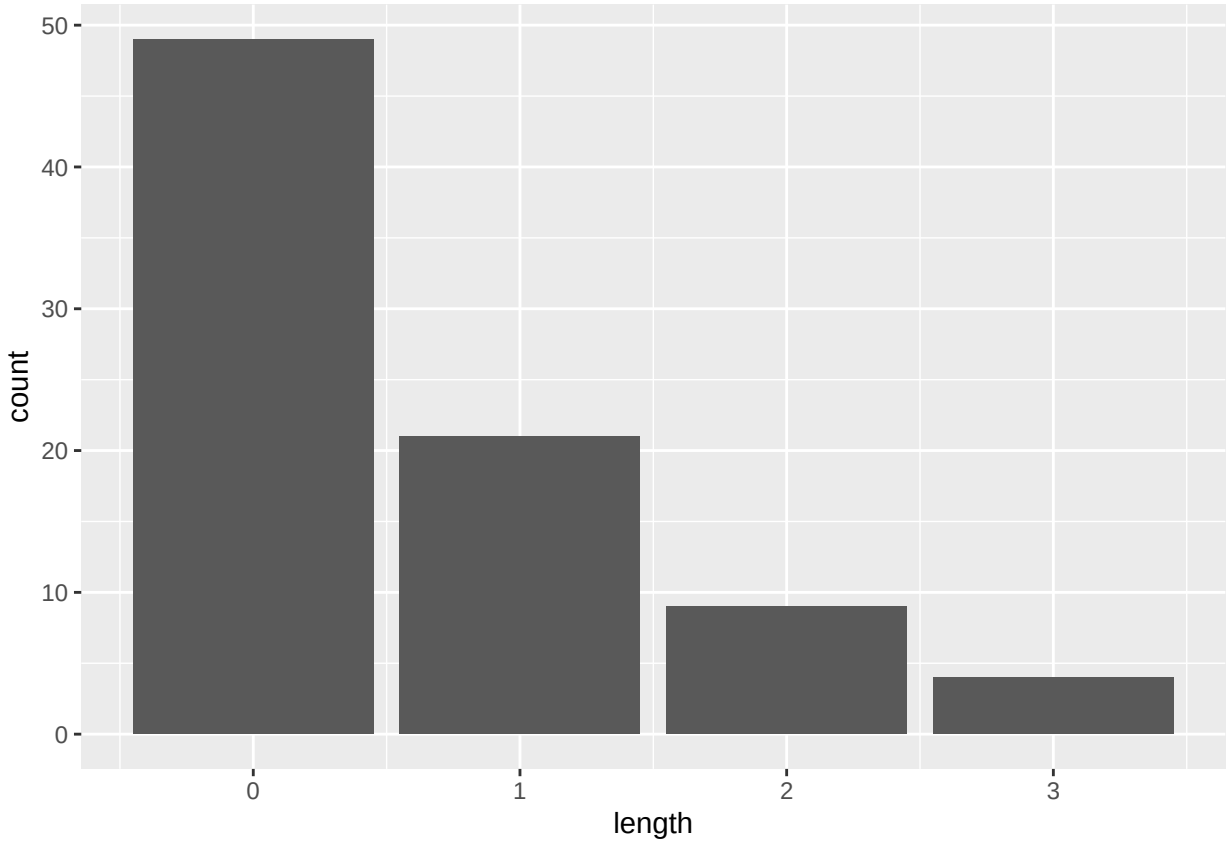
```
sim_streak <- calc_streak(sim_basket)
```

6 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.

Based on mode, the typical streak length of the simulated shooter is 0 (no hits) and the longest length is 3.

```
ggplot(data = sim_streak, aes(x = length)) +
  geom_bar()
```




```
my.mode(sim_streak$length)
```

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

```
max(sim_streak$length)
```

```
## [1] 3
```

7 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.

Both distributions should be identical since we explicitly set the seed of the random number generator. If the seed was removed (or modified) then the streak distribution would be similar but not identical. This variation is an inherent property of the random number generator used for the simulation.

8 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.

Both distributions look *unimodel*, are right skewed, and have mode equal to 0. Stream lengths of the two distributions, although not identical, are similar. For example, unlike the simulated shooter, who has 0 streaks of length 4, Kobe has 1 streak of length 4.

One would expect significant number of longer streaks if the hot hand model was valid. We can conclude that Kobe's streaks are random and not due to a hot hand.

```
kobe_streak_counts <- table(kobe_streak$length)
kobe_streak_counts
```

```
##
##  0  1  2  3  4
## 39 24  6  6  1
```

```
sim_streak_counts <-table(sim_streak$length)
sim_streak_counts
```

```
##
##  0  1  2  3
## 49 21  9  4
```