

# Data 606 Lab 3

Joao De Oliveira

2025-09-27

## Load packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2     3.5.2      v tibble    3.2.1
## v lubridate  1.9.4      v tidyr     1.3.1
## v purrr      1.0.4
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

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

## The Data

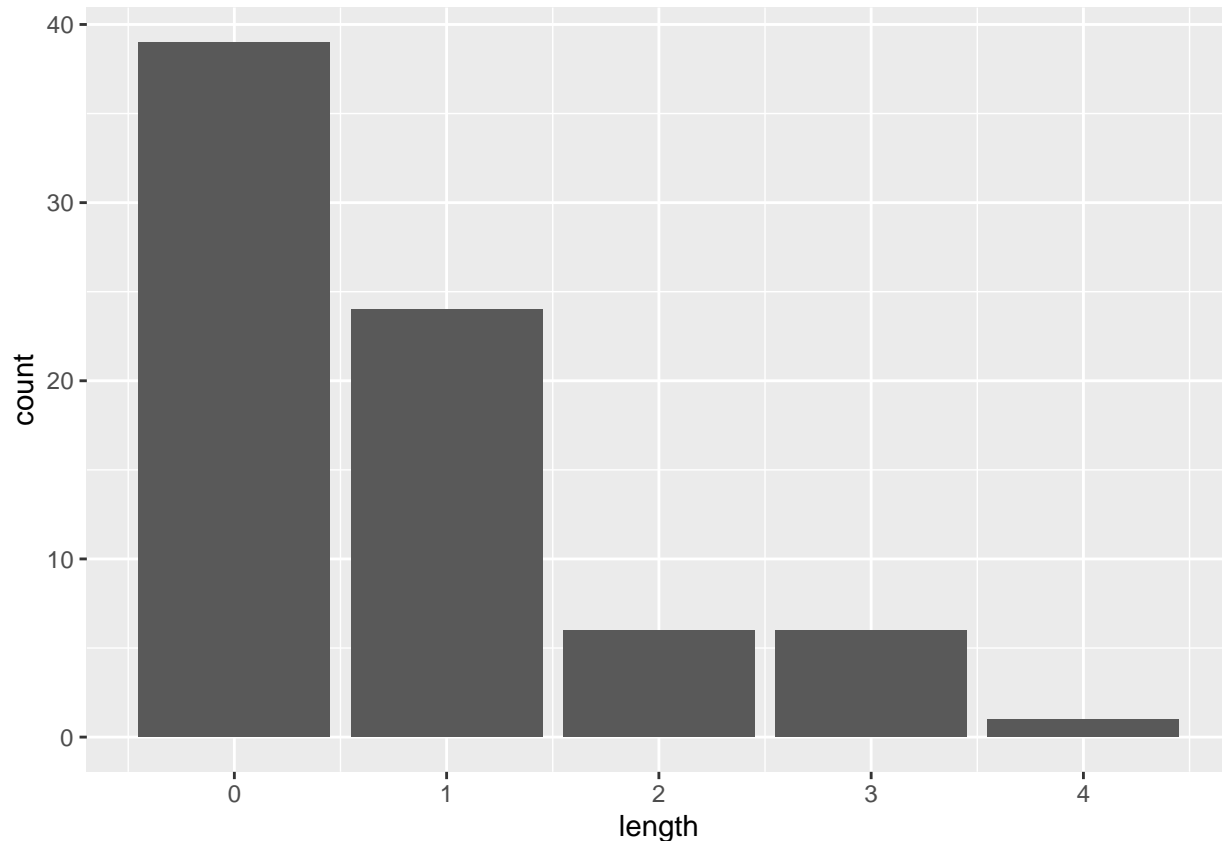
```
glimpse(kobe_basket)
```

```
## 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~
## $ 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, 6:35~
## $ 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"~
```

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?

```
kobe_streak <- calc_streak(kobe_basket$shot)
ggplot(data = kobe_streak, aes(x = length)) +
  geom_bar()
```

Answer 1: A streak with length of 1 means that there are two shots represented in that streak (shot with index 0 and shot with index 1). So, in a streak length of 0 there is one single shot



with index 0”

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.

Answer 2: Kobe Bryant’s streak lengths distribution from the 2009 NBA finals is right-skewed since the length with more counts is clearly 0 (1 shot), meaning most successful shots were isolated shots. However, the number of times with 2 consecutive shots is also pretty relevant. The longest streak had a length of 4 (5 shots).

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

Question 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.

```
## [1] "tails"
```

```
set.seed(54962)
```

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

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

```
table(sim_fair_coin)
```

```
## sim_fair_coin
## heads tails
##    59    41
```

```
sim_fair_coin <- sample(coin_outcomes, size = 100, replace = TRUE, prob = c(0.2, 0.8))
```

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

Answer 3: In my simulation there are 59 heads.

```
## sim_basket
## H M
## 64 69
```

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`.

```

# exercise 5: Using calc_streak, compute the streak lengths of sim_basket, and save the results in a data frame
# Answer 5:
sim_streak <- calc_streak(sim_basket)

# exercise 6: Describe the distribution of streak lengths. What is the typical streak length for this simulation?

sim_streak_table <- sim_streak |>
  dplyr::count(length, name = "count") |>
  dplyr::mutate(prop = count / sum(count))

mode_length <- sim_streak_table$length[which.max(sim_streak_table$count)]
max_length <- max(sim_streak$length)

mode_length

```

Answer 4: The changes are: changing the size of the sample to 133 (size = 133) and the probabilities of 45 and 55 (prob = c(0.45, 0.55))

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

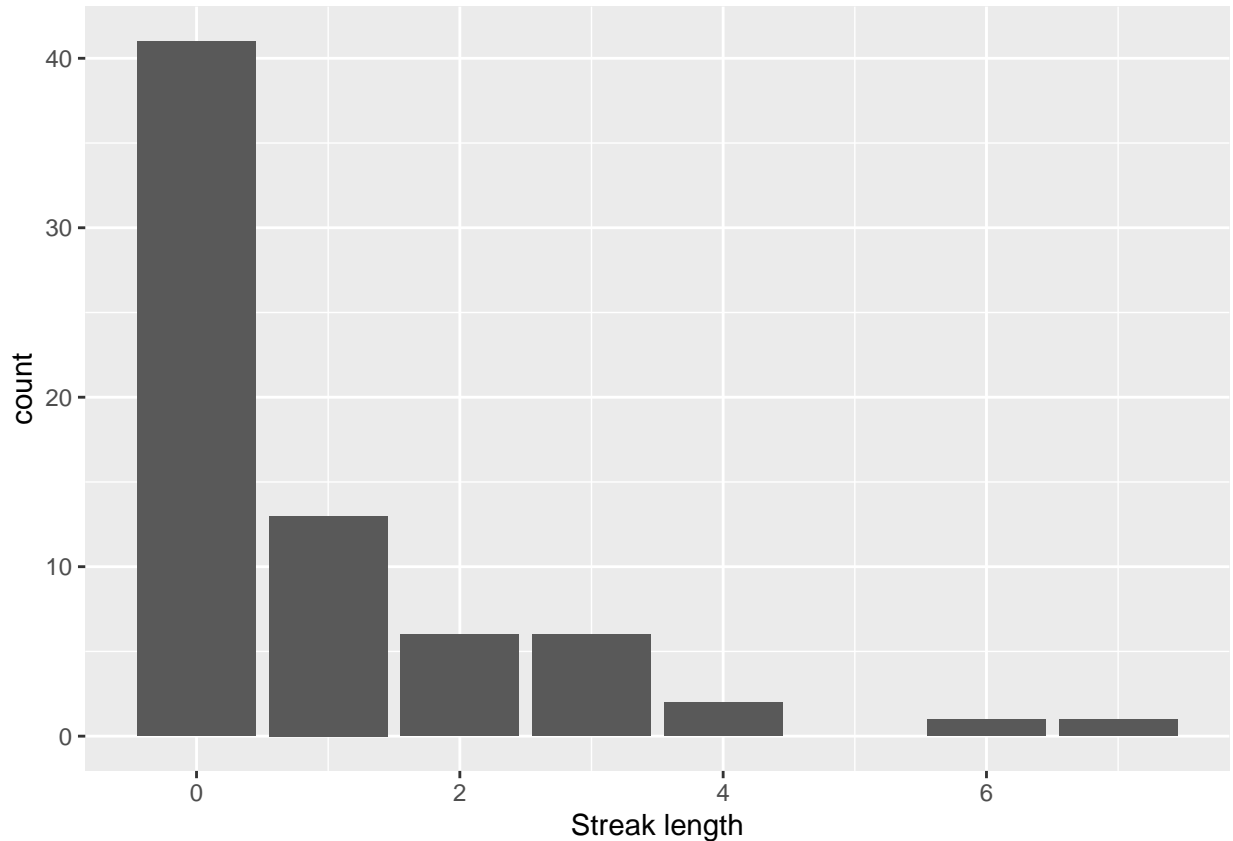
```
max_length
```

```
## [1] 7
```

```

ggplot(sim_streak, aes(x = length)) +
  geom_bar() +
  labs(x = "Streak length", y = "count")

```



**Answer 6:** This distribution is also right-skewed, meaning most of the data is on the left side (1 shot streaks, frequency decreases for more than 1 shot streaks). The mode (typical streak length) is 0 (1 shot) and max length (longest streak) was 7 (8 shots).

**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.

**Answer 7:** I expect the distribution to have a similar pattern but not exactly the same since randomness changes the exact output values but not the general conclusions about the distribution.

**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.

**Answer 8:** We can see that both Kobe and the independent shooter's distributions are pretty similar, right-skewed with a mode of 0 (1 shot). In reality, Kobe's longest streak was considerably smaller than the independent shooter longest streak ( $4 < 7$ ) showing that a real game and the pressure that it involves probably impacts the ability to have long streaks. So, the data shows that there is no reason that claim that Kobe had a hot hand during the 2009 finals.