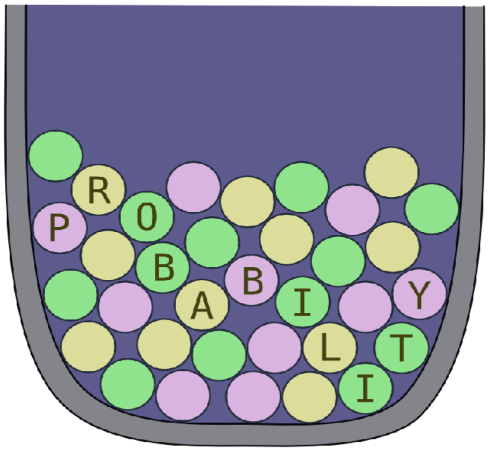
IT137IU: Data Analysis
Lab#8/Assignment#8: Probability



Introduction

This lab intends to provide practice working with important terms and concepts related to probability. It also serves as an example of how you can use the computational power of `R` to uncover new insights that might be difficult to reveal with pencil and paper calculations.

Basketball players who make several baskets in succession are described as having a *hot hand*. Fans and players have long believed in the hot hand phenomenon, which refutes the assumption that each shot is independent of the next. However, a 1985 paper in attachment pdf file by Gilovich, Vallone, and Tversky collected evidence that contradicted this belief and showed that successive shots are independent events.

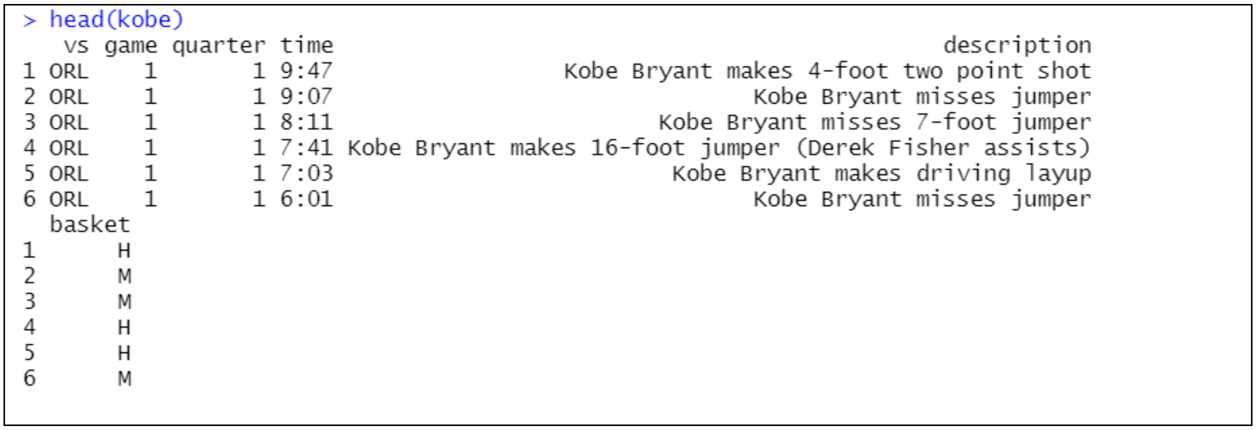
In this lab, we will analyze the performance of the late Kobe Byrant against the Orlando Magic in the 2009 NBA finals on route to his fifth and final championship (his performance led to him being the finals MVP). More specifically, we will look at his made and missed shot attempts to investigate whether Kobe tended to go on shooting streaks or if his shots were independent of each other. The goals for this lab are:

1. Think about the effects of independent and dependent events,
2. Learn how to simulate shooting streaks in R,
3. To compare a simulation to actual data to determine if the hot hand phenomenon appears to be real.

1. The Kobe Dataset

I uploaded a file on Blackboard namely **Kobe.csv**. Download the file and save it into the same folder where your Lab 8 files reside. In this dataset, Kobe took 133 shot attempts during the entire five-game series. The outcome of each shot is recorded in the variable “basket” as either “H” for a hit or made shot, or as “M” for a missed shot. Additional information, such as game, quarter, time, and description are also included in the dataset.

```
## This reads the data

kobe <- read.csv("Kobe.csv")

## This will load a custom function that we'll use in this lab

source("functions.R")

## Print the first few rows of the Kobe dataset

head(kobe)
```



```
> head(kobe)
   vs game quarter time                                          description
1 ORL    1       1 9:47                   Kobe Bryant makes 4-foot two point shot
2 ORL    1       1 9:07                                Kobe Bryant misses jumper
3 ORL    1       1 8:11                         Kobe Bryant misses 7-foot jumper
4 ORL    1       1 7:41 Kobe Bryant makes 16-foot jumper (Derek Fisher assists)
5 ORL    1       1 7:03                          Kobe Bryant makes driving layup
6 ORL    1       1 6:01                                Kobe Bryant misses jumper
  basket
1      H
2      M
3      M
4      H
5      H
6      M
```

<u>Exercise 1:</u> [20pts] What proportion of Kobe's shot attempts did he hit?

2. Shooting Streak

Looking at any individual string of hits and misses, it is difficult to tell whether Kobe has a hot hand or not. To make things more manageable, we'll aggregate the data by looking at *consecutive made shots until a miss occurs* using the `calc_streak` function.

To demonstrate how this works, the code below prints Kobe's first 9 shots in Game 1

```
kobe$basket[1:9]
```

```
## [1] "H" "M" "M" "H" "H" "M" "M" "M" "M"
```

We can see the first "streak" contains 2 shots (HM) with 1 hit, the second was 1 shot (M) with 0 hits, the third was 3 shots (HHM) with 2 hits, the fourth, fifth, and sixth were all 1 shot (M) with 0 hits. Compare this with the output of the `calc_streak` function that is seen below:

```
calc_streak(kobe$basket[1:9])
```

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

Exercise 2: **[20pts]** Apply the `calc_streak` function to the entire data (not just the first 9 shots as shown above), storing the result in an object called “kobe_streaks”. Then use the `table` function to calculate a frequency table of Kobe’s different shooting streaks. How long was his longest streak of baskets?

3. Compared to What

We’ve shown that Kobe had some long shooting streaks, but are they long enough to support the belief that he had hot hands? What can we compare them to?

To answer these questions, let’s return to the idea of ***independence***. Two processes are independent if the outcome of one process doesn’t affect the outcome of the second. If each shot that a player takes is an independent process, having made or missed your first shot will not affect the probability that you will make or miss your second shot.

A shooter with a hot hand will have shots that are not independent of one another. Specifically, if the shooter makes his first shot, the hot hand model says he will have a higher probability of making his second shot.

Let’s suppose for a moment that the hot hand model is valid for Kobe. During his career, the percentage of time Kobe makes a basket (i.e. his shooting percentage) is about 44%, or in probability notation,

$$P(\text{shot } 1 = H) = 0.44$$

If he makes the first shot and has a hot hand (not independent shots), then the probability that he makes his second shot would go up to, let’s say, 54%,

$$P(\text{shot } 2 = H | \text{shot } 1 = H) = 0.54$$

As a result of these increased probabilites, you’d expect Kobe to have longer streaks. Compare this to the skeptical perspective where Kobe does not have a hot hand, where each shot is independent of the next. If he hits his first shot, the probability that he makes the second is still 0.44.

$$P(\text{shot } 2 = H | \text{shot } 1 = H) = 0.44$$

In other words, making the first shot did nothing to affect the probability that he’d make his second shot. If Kobe’s shots are independent, then he’d have the same probability of hitting every shot regardless of his past shots: 44%.

Now that we’ve phrased the situation in terms of independent shots, let’s return to the question: how do we tell if Kobe’s shooting streaks are long enough to indicate that he has hot hands? We can compare his streak lengths to someone without hot hands: an independent shooter.

4. Simulation in R

While we don’t have any data from a shooter, we know to have independent shots, that sort of data is very easy to simulate in R. In a simulation, you set the ground rules of a random process and then the computer uses random numbers to generate an outcome that adheres to those rules. As a simple example, you can simulate flipping a fair coin with the following.

Simulating a Fair Coin

The code below shows how to simulate flipping a fair coin two times:

```
outcomes <- c("heads", "tails")
sample(outcomes, size = 2, replace = TRUE)
```

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

In this example, the vector `outcomes` can be viewed as a container holding slips of paper with the label’s “heads” and “tails”. The `sample` function can be thought of as drawing from this container a certain number of times (size) with or without replacement (we asked for replacement here). Try running the sample function several times to get a feel for the random process at work.

If you wanted to simulate flipping a fair coin 1000 times, you could either run the function 1000 times or, more simply, adjust the size argument, which governs how many samples to draw. Save the resulting vector of heads and tails in a new object called `coin_outcomes`. Note that for random processes which are repeated many times, the table function can be useful in summarizing the results:

```
outcomes <- c("heads", "tails")
coin_outcomes <- sample(outcomes, size = 1000, replace = TRUE)
table(coin_outcomes)
```

```
## coin_outcomes
## heads tails
##   505   495
```

Simulating a Weighted Coin

Sometimes we'll want to simulate scenarios where the outcomes aren't 50-50. This can be done by specifying the probability of each individual outcome within the `sample` function:

```
coin_outcomes <- sample(outcomes, size = 1000, replace = TRUE, c= (0.8,0.2))
```

Notice the difference in outcomes when specifying an 80% probability of heads and a 20% probability of tails. Additionally, you should recognize that `sample` will expect you to provide a vector input to the `prob` argument whose length matches your vector `outcomes` (so we couldn't just say `prob = .8` here, instead we need to specify probabilities for every outcome we listed). Note that if the `prob` argument is not used, `sample` will assume that all of the outcomes provided are equally likely.

Exercise 3: [**20pts**] In your simulation of flipping the unfair coin, how many flips came up heads?

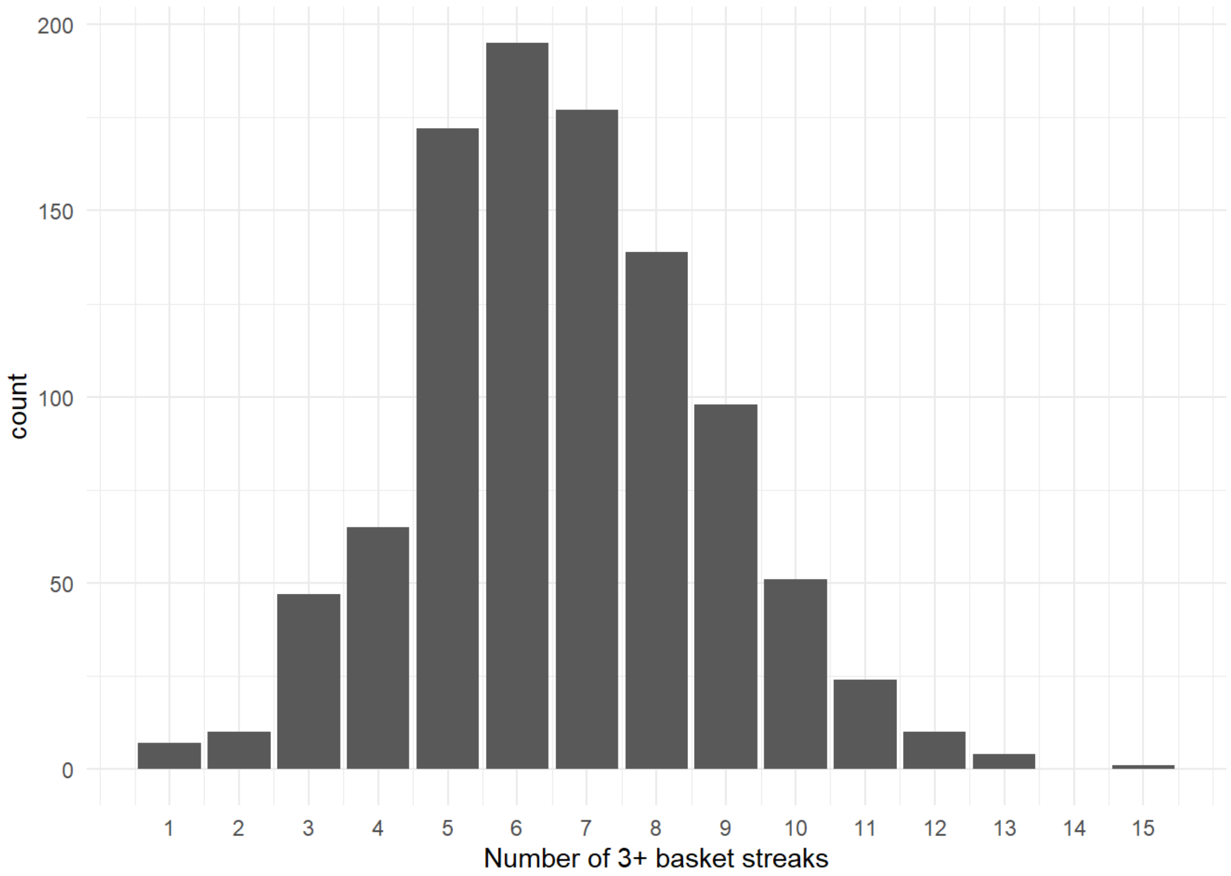
Simulating an Independent Shooter

In this application, we're interested in how Kobe compares to an independent shooter. If Kobe's performance in the finals significantly deviates from the outcomes, you'd expect under the *independence* model we'd have evidence supporting the *hot hand* model.

Exercise 4: **[20pts]** Using the coin-flip examples described above, simulate 133 shot attempts according to the independence model (i.e.: a 44% chance of hit). Then, use the `calc_streak` and `table` functions to summarize the results of this simulation. How does this table compare to the table of the real data? Include your code in the proper block and write your written answer in the space below the block. (Hint: be sure to use "H" and "M" to denote your outcomes, otherwise `calc_streak` won't work properly).

5. Replication

Simulating an independent shooter *once* and comparing it to Kobe's performance isn't enough to sufficiently evaluate the plausibility of the null model. Instead, we should simulate many times and see how unlikely a performance like Kobe's might be.



Exercise 5: **[20pts]** Based upon the histogram above, which displays the number of 3+ basket shooting streaks from 1000 simulated independent shooters each attempting 133 shots, do you believe Kobe's 2009 finals performance supports the existence of the *hot hand* phenomenon? (Hint: use this histogram to estimate the *probability* of seeing Kobe's 2009 finals numbers if the independence model were true).

What to submit:

Your submission should include the following:

1. Lab report answers all exercises above and source code.
2. Please create a folder called "yourname_studentID_lab8" that includes all the required files and generate a zip file called "yourname_studentID_lab8.zip".
3. Please submit your work (.zip) to Blackboard.