Homework 1

STAT 399: Statistical Computing and Data Visualization with R

Do not include your name, keep anonymous for peer review

Due: Tuesday 4/14/2020 by 10:00 pm

Instructions

- Make a copy of this template to write out your solution, and rename it before knitting it the first time as file as `LastnameFirstname_HW01.Rmd`, where you should replace `Lastname` and `Firstname` by your own last name and first name, respectively.
- Inside this .Rmd file do not include any personal identifier (such as your name, Odin ID, etc.).
- Knit your Rmd file as html and upload both the Rmd and html files with your solution to D2L in `Activities > Assignments > Homework1` before Tuesday April 14th at 4:30 pm.

Objectives for this week's homework

1. **R Programming Objectives**

- Practice subsetting
- Using loops
- Creating your own functions
- Working with data frames in R

2. **Statistics Objectives**

- Think about *statistical independence*
- Obtain frequencies from categorical data
- Obtain conditional probabilities with real data
- Learn how to simulate data under particular assumptions and derive meaningful inference from it

We do not expect to resolve this controversy, but in this homework we'll take a stab at the problem using two approaches that could be considered to tackle questions like this.

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 by Gilovich, Vallone, and Tversky collected evidence that contradicted this belief and showed that successive shots are independent events (http://psych.cornell.edu/sites/default/files/Gilo.Vallone.Tversky.pdf). This paper started a great controversy that continues to this day, as you can see by Googling *hot hand basketball*.

Our investigation will focus on Kobe Bryant's performance with the Los Angeles Lakers in the 2009 NBA finals when playing against the Orlando Magic, which earned him the title *Most Valuable Player*. Many spectators commented on him having a *hot hand*. Let's load some data from those games and look at the data structure with `str`.

```
load("kobefiles.RData")
str(kobe)
```

```
## 'data.frame':    133 obs. of  8 variables:
##  $ vs         : Factor w/ 1 level "ORL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ game       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ quarter    : Factor w/ 5 levels "1","1OT","2",..: 1 1 1 1 1 1 1 1 1 3 ...
##  $ time       : Factor w/ 116 levels "00:00.0","00:00.5",..: 114 109 102 100 96 85 64 21 11 91 ...
```

```
##  $ description: Factor w/ 80 levels "Bryant 3pt Shot: Made (16 PTS) Assist: Bynum (1 AST)  ",..: 40
##  $ basket     : chr  "H" "M" "M" "H" ...
##  $ streak     : num  1 1 2 3 3 3 4 5 6 7 ...
##  $ shot.num   : num  1 2 1 1 2 3 1 1 1 1 ...
```

In this data frame, each row contains a shot taken by Kobe Bryant. If he hit the shot (made a basket), a hit, `H`, is recorded in the column named `basket`, otherwise a miss, `M`, is recorded.

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

To verify this use the following command:

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

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).

Question 1.1

Think about what a streak of length 0, length 1, length 2, etc. mean (i.e. how many hits and misses are in a streak of length $m = 0, 1, 2, \ldots$). Using as input the variable or variables from the `kobe` data that you consider necessary, build the function `get_streak`, which returns the length of each streak.

```
##Build your function here
```

Now, use your function `get_streak` to calculate the lengths of all shooting streaks and then use the R functions `summary`, `quantile` and `barplot` to describe the distribution of Kobe's streak lengths from the 2009 NBA finals (use `?` to find out about their usage). What was his typical streak length? How long was his longest streak of baskets?

Question 1.2

So 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? Consider the idea of *statistical independence*. A shooter with a hot hand will have shots that are *not* independent of one another. Specifically, if the shooter makes their first shot, the hot hand model says they will have a *higher* probability of making their second shot.

During Kobe's career, the percentage of time he makes a basket (i.e. his shooting percentage) is about 45%, or equivalently

$$P(\text{shot } 1 = \text{H}) = 0.45.$$

If hot hands is really a thing, then when Kobe 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, 60%,

$$P(\text{shot } 2 = \text{H} \,|\, \text{shot } 1 = \text{H}) = 0.60.$$

Because of these increased probabilites, you'd expect Kobe to have longer streaks. Now, if *hot hands* are just a myth, and each shot is independent of the next. When Kobe hits his first shot, the probability that he makes the second is still 0.45.

$$P(\text{shot } 2 = \text{H} \,|\, \text{shot } 1 = \text{H}) = 0.45.$$

Now, having expressed the problem in this way we may assess if Kobe's shooting streaks are long enough to indicate that he has hot hands. Here are two possible ways: 1) calculating the conditional probabilities and 2) comparing Kobe's streak lengths to someone without hot hands (a simulated independent shooter).

Part 1.2.a – Conditional probabilities

1. With the data, a logical statement, and the function `mean`, first calculate the total percentage of shots that resulted in a basket in the 2009 NBA finals, as

$$\frac{\#\text{ hits}}{\text{total }\#\text{ shots}}$$

2. We need to filter out the streaks that had at least the first shot resulting in a Hit – by doing this we are conditioning the data to make the conditional statement. Since those streaks in which Kobe made the first shot have two shots or more, use the variable "shot.num" in the dataset to calculate

$$P(\text{shot } 2 = \text{H} \,|\, \text{shot } 1 = \text{H}) = \frac{\#(\text{shot } 2 = \text{H} \cup \text{shot } 1 = \text{H})}{\#(\text{shot } 2 = \text{H} \cup \text{shot } 2 = \text{M})} = \frac{\#\text{shot } 2 = \text{H}}{\#\text{ shot } 2}$$

by identifying those observations corresponding to the second shots (i.e., those with "shot.num==2").

3. Is there evidence to think that Kobe has *hot hands*? How reliable is this conclusion? Provide an objective argument to justify your answer.

Part 1.2.b – Simulating independent shooters

The second alternative is to compare Kobe's streak lengths to the streak lengths of shooters without hot hands, or in other words to independent shooters. We don't have any data from shooters we know to have independent shots, but this type 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. To simulate a single shot from an independent shooter with a shooting percentage of 50% we can use the code below (switch the chunk option `eval=FALSE` to `eval=TRUE` so that this chunk is evaluated).

```
outcomes <- c("H", "M")
sample(outcomes, size = 1, replace = TRUE, prob = c(0.5,0.5))
```

Keep in mind that to make a valid comparison between Kobe and our simulated independent shooter, we need to align both their shooting percentage and the number of attempted shots.

If you want to learn more about `sample` or any other function, recall that you can always check out its help file.

```
?sample
```

Work on the following problems:

1. Simulate 133 shots from an independent shooter *comparable* to Kobe and using `calc_streak`, compute the streak lengths of this independent shooter.
2. Use the R functions `table` and `quantile` to compare the streak length distribution of Kobe and that for this independent shooter
3. Build the R function `sim_generation`, which takes the number of independent shooters (N) to simulate, the number of shots taken by each shooter (m), and the shooting percentage ($perc$); and returns a data frame containing the outcomes for all shot taken by a single shooter in each row.

4. Use the function `sim_generation` to simulate $N = 500$ shooters taking $m = 133$ shots and $perc = 0.45$.
5. Get creative, use the results from the previous exercise to evaluate if Kobe has in fact hot hands.

This homework was created by adapting materials from