DATA 606 - Lab2

Joshua Sturm

09/07/2017

Getting Started

Our 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. Let's load some data from those games and look at the first several rows.

```
load("more/kobe.RData")
head(kobe)
```

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

In this data frame, every row records 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]
```

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

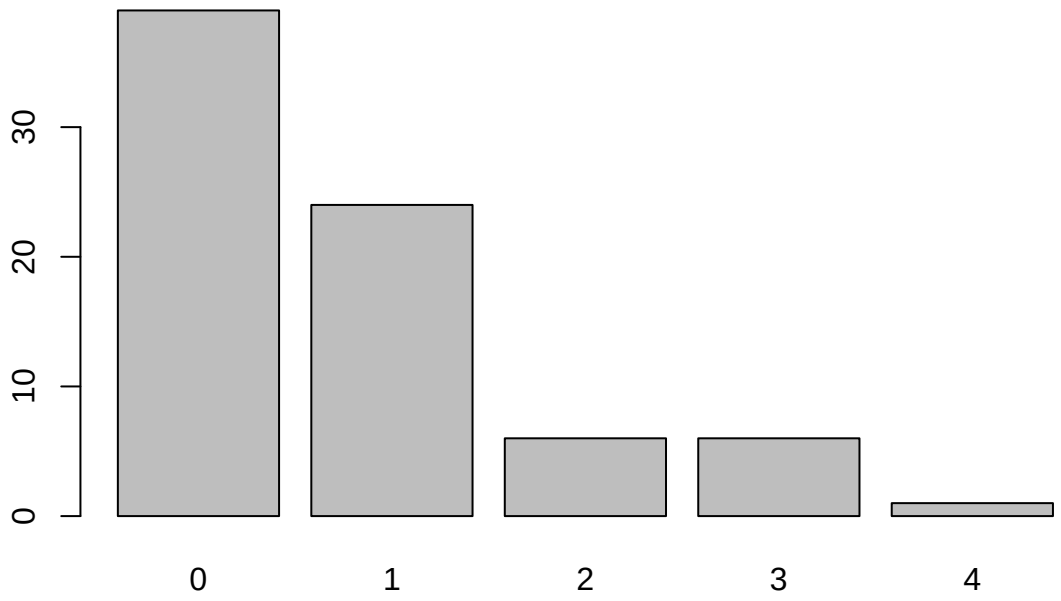
Within the nine shot attempts, there are six streaks, which are separated by a "|" above. Their lengths are one, zero, two, zero, zero, zero (in order of occurrence).

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 of 1 would simply be only one consecutive hit. A streak of 0 would be a miss.
```

The custom function `calc_streak`, which was loaded in with the data, may be used to calculate the lengths of all shooting streaks and then look at the distribution.

```
kobe_streak <- calc_streak(kobe$basket)
barplot(table(kobe_streak))
```



Note that instead of making a histogram, we chose to make a bar plot from a table of the streak data. A bar plot is preferable here since our variable is discrete – counts – instead of continuous.

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?

```
It's right skewed, in that the majority were misses, with very few actual hit streaks. The typical strea
```

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 effect 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 45%, or in probability notation,

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

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, 60%,

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

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 hit his first shot, the probability that he makes the second is still 0.45.

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

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: 45%.

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.

Simulations 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.

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

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

The vector `outcomes` can be thought of as a hat with two slips of paper in it: one slip says `heads` and the other says `tails`. The function `sample` draws one slip from the hat and tells us if it was a head or a tail.

Run the second command listed above several times. Just like when flipping a coin, sometimes you'll get a heads, sometimes you'll get a tails, but in the long run, you'd expect to get roughly equal numbers of each.

If you wanted to simulate flipping a fair coin 100 times, you could either run the function 100 times or, more simply, adjust the `size` argument, which governs how many samples to draw (the `replace = TRUE` argument indicates we put the slip of paper back in the hat before drawing again). Save the resulting vector of heads and tails in a new object called `sim_fair_coin`.

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

To view the results of this simulation, type the name of the object and then use `table` to count up the number of heads and tails.

```
sim_fair_coin
```

```
##   [1] "tails" "tails" "tails" "heads" "tails" "heads" "heads" "tails"
##   [9] "tails" "heads" "heads" "heads" "tails" "heads" "heads" "tails"
##  [17] "heads" "heads" "tails" "heads" "tails" "tails" "heads" "tails"
##  [25] "tails" "heads" "heads" "heads" "tails" "heads" "tails" "heads"
##  [33] "heads" "tails" "tails" "heads" "heads" "heads" "tails" "heads"
```

```
##  [41] "tails" "tails" "heads" "heads" "tails" "tails" "tails" "heads"
##  [49] "heads" "tails" "tails" "heads" "heads" "heads" "heads" "tails"
##  [57] "tails" "tails" "heads" "tails" "heads" "tails" "tails" "tails"
##  [65] "heads" "heads" "heads" "heads" "tails" "heads" "tails" "heads"
##  [73] "tails" "heads" "tails" "tails" "heads" "tails" "tails" "tails"
##  [81] "tails" "tails" "tails" "tails" "heads" "heads" "tails" "tails"
##  [89] "tails" "tails" "tails" "tails" "tails" "tails" "tails" "tails"
##  [97] "heads" "heads" "heads" "tails"
```

```
table(sim_fair_coin)
```

```
## sim_fair_coin
## heads tails
##    45    55
```

Since there are only two elements in `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.

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

`prob=c(0.2, 0.8)` indicates that for the two elements in the `outcomes` vector, we want to select the first one, `heads`, with probability 0.2 and the second one, `tails` with probability 0.8. Another way of thinking about this is to think of the outcome space as a bag of 10 chips, where 2 chips are labeled "head" and 8 chips "tail". Therefore at each draw, the probability of drawing a chip that says "head"" is 20%, and "tail" is 80%.

3. In your simulation of flipping the unfair coin 100 times, how many flips came up heads?

```
sim_unfair_coin <- sample(outcomes, size = 100, replace = TRUE, prob = c(0.2, 0.8))
table(sim_unfair_coin)
```

```
## sim_unfair_coin
## heads tails
##    19    81
```

Simulating the Independent Shooter

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

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

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.

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`.

```
set.seed(472)
sim_basket <- sample(outcomes, size = 133, replace = TRUE, prob=c(0.45, 0.55))
table(sim_basket)
```

```
## sim_basket
##  H  M
## 73 60
```

We can customize the probability for each variable with the prob() function.

```
I used the function 'set.seed' to make the data reproducible.
```

With the results of the simulation saved as `sim_basket`, we have the data necessary to compare Kobe to our independent shooter. We can look at Kobe's data alongside our simulated data.

```
table(kobe$basket)
```

```
##
##  H  M
## 58 75
```

Both data sets represent the results of 133 shot attempts, each with the same shooting percentage of 45%. We know that our simulated data is from a shooter that has independent shots. That is, we know the simulated shooter does not have a hot hand.

On your own

Comparing Kobe Bryant to the Independent Shooter

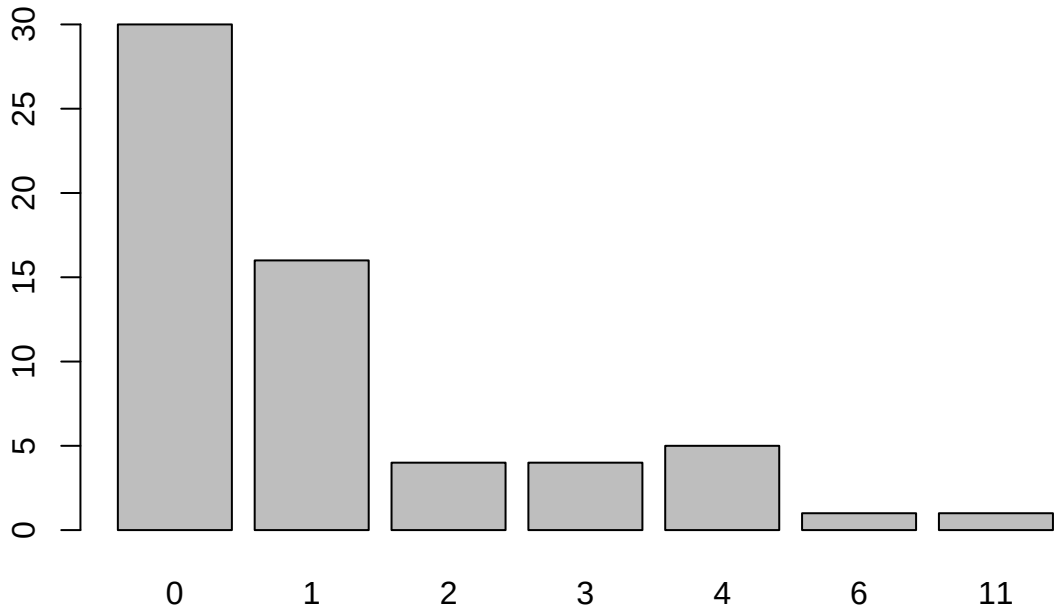
Using `calc_streak`, compute the streak lengths of `sim_basket`.

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

```
I calculated the streak for our simulated player, and assigned it to a new variable 'sim_streak'
```

1. 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?

```
barplot(table(sim_streak))
```



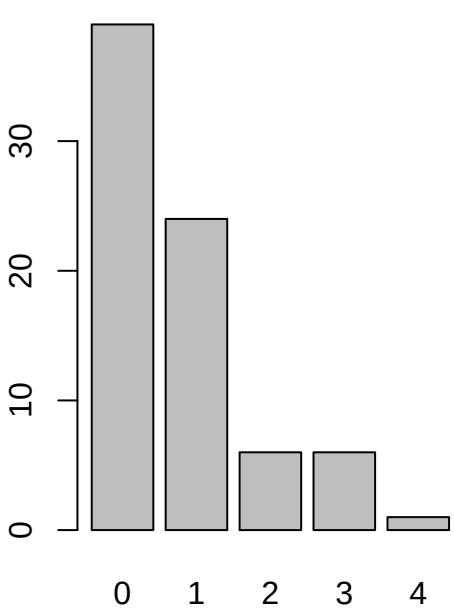

```
The distribution is unimodal and right skewed. The majority of shots were misses (i.e. streak of 0). Th
```

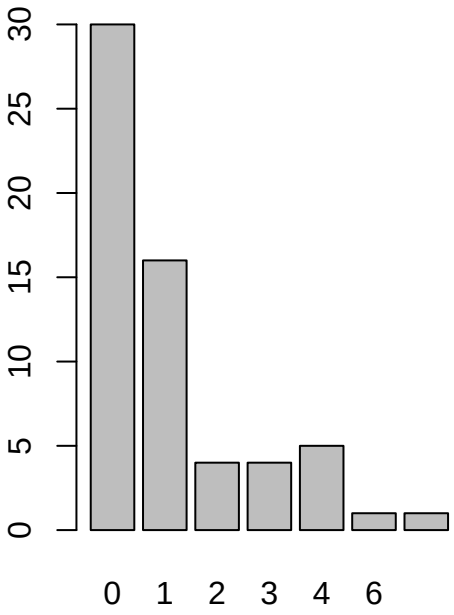
2. If you were to run the simulation of the independent shooter a second time, how would you expect its streak distribution to compare to the distribution from the question above? Exactly the same? Somewhat similar? Totally different? Explain your reasoning.

```
Assuming we didn't use 'set.seed' in both runs, we would get slightly different results, but they'd be
```

3. 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.

```
par(mfrow=c(1,2))
barplot(table(kobe_streak))
barplot(table(sim_streak))
```





While the independent shooter had some longer streaks, overall, the distributions of his and kobe's sho