da460\_assignment2\_grahn

jason grahn

2018-10-05

download.file("http://www.openintro.org/stat/data/kobe.RData", destfile = "kobe.RData")  
load("kobe.RData")  
head(kobe,5)

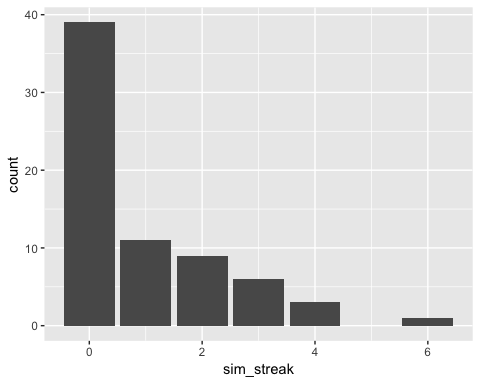
## vs game quarter time  
## 1 ORL 1 1 9:47  
## 2 ORL 1 1 9:07  
## 3 ORL 1 1 8:11  
## 4 ORL 1 1 7:41  
## 5 ORL 1 1 7:03  
## description basket  
## 1 Kobe Bryant makes 4-foot two point shot H  
## 2 Kobe Bryant misses jumper M  
## 3 Kobe Bryant misses 7-foot jumper M  
## 4 Kobe Bryant makes 16-foot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

## 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?

#first we have to generate the data for the simulated player  
# the outcomes vector  
outcomes <- c("H", "M")  
  
#then run the shooting simulation with the 45% hit rate  
sim\_basket <- sample(outcomes, size = 133, replace = TRUE, prob = c(0.45, (1-.45)))  
  
#role that into a table  
tibble(sim\_basket)

## # A tibble: 133 x 1  
## sim\_basket  
## <chr>   
## 1 M   
## 2 M   
## 3 M   
## 4 M   
## 5 H   
## 6 H   
## 7 H   
## 8 M   
## 9 M   
## 10 H   
## # ... with 123 more rows

##run the calc\_streak function to generate the kobe\_steak dataset  
sim\_streak <- calc\_streak(sim\_basket)  
  
#make a barchart and save it for reference  
sim\_bar <- tibble(sim\_streak) %>%   
 ggplot() +  
 geom\_bar(aes(x = sim\_streak))  
  
#display that barchart  
sim\_bar



#and lets save those counts to their own dataframe  
sim\_aggregate <- tibble(sim\_streak) %>%   
 group\_by(sim\_streak) %>%   
 summarise(sim\_count = n())

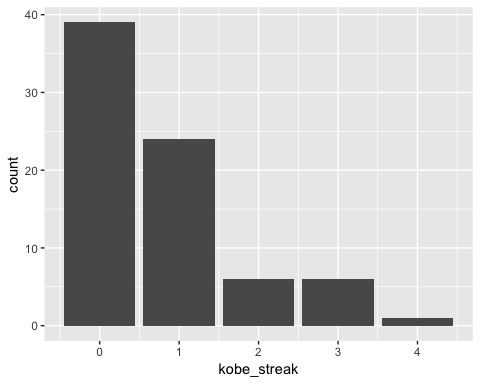
The distrubution is right skewed with a maximum streak of 6. The typical streak length, if we don’t include streaks of zero, is *one*.

## 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.

No code is required to answer this question. Running the simulation a second time will still maintain the right skewness, but the individual values of the streaks will change. It would likely appear similar, but with different values for each of the streak lengths. This is because the simulation doesn’t include proportions of streak lengths, just random selections of ‘hit’ or ‘miss’.

## 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.

#run the calc\_streak function to generate the kobe\_steak dataset  
kobe\_streak <- calc\_streak(kobe$basket)  
  
#and let's plot that, saving the bar chart  
kobe\_bar <- tibble(kobe\_streak) %>%   
 ggplot(aes(x = kobe\_streak)) +  
 geom\_bar()  
  
#show that plot  
kobe\_bar



#but more importantly, we need to see the counts of each next to each other.  
kobe\_aggregate <- tibble(kobe\_streak) %>%   
 group\_by(kobe\_streak) %>%   
 summarise(kobe\_count = n()) %>%   
 rename(streak\_length = kobe\_streak)

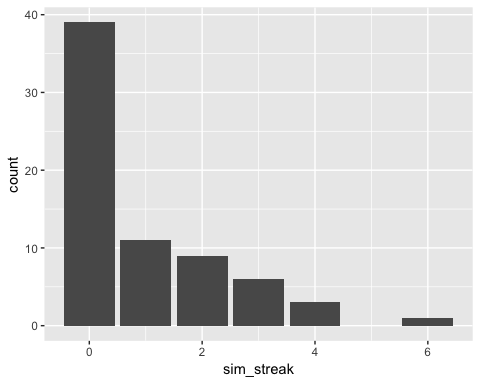
#save the two tables as one, so we can compare  
streak\_difference <- sim\_aggregate %>%   
 rename(streak\_length = sim\_streak) %>%   
 full\_join(kobe\_aggregate) %>%   
 replace\_na(list(kobe\_count = 0)) %>%   
 replace\_na(list(sim\_count = 0)) %>%   
 mutate(streak\_delta = kobe\_count - sim\_count)

## Joining, by = "streak\_length"

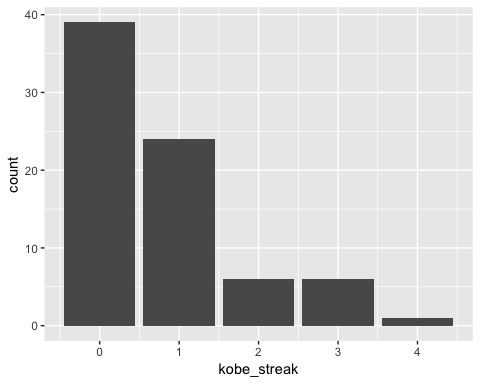
streak\_difference

## # A tibble: 6 x 4  
## streak\_length sim\_count kobe\_count streak\_delta  
## <dbl> <dbl> <dbl> <dbl>  
## 1 0 39 39 0  
## 2 1 11 24 13  
## 3 2 9 6 -3  
## 4 3 6 6 0  
## 5 4 3 1 -2  
## 6 6 1 0 -1

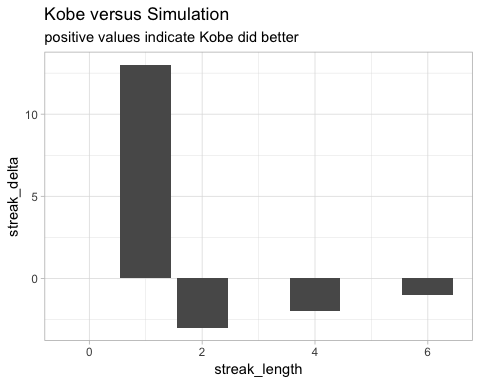
sim\_bar



kobe\_bar



#and plot a viz of the difference  
ggplot(streak\_difference) +  
 geom\_bar(aes(x = streak\_length, y = streak\_delta), stat = "identity") +  
 labs(title = "Kobe versus Simulation",  
 subtitle = "positive values indicate Kobe did better") +   
 theme\_light()



### How does Kobe Bryant’s distribution of streak lengths compare to the distribution of streak lengths for the simulated shooter?

They are both right-skewed streaks for sure, with 1-hit streaks making the longest (with zero-streak lengths excluded). The simulated player has a longer right tail than Kobe because it has streaks that lasted longer, *and* it has less streaks of zero.

### Using this comparison, do you have evidence that the hot hand model fits Kobe’s shooting patterns? Explain.

The simulated model has a higher streak maximum and did quite a bit better at 2-run streaks. If the simulated “hot hand” model were more like Kobe, we would see less streaks last as long. From this simple analysis, it doesn’t seem like we have evidence that the hot hand simulation fit’s Kobe’s shooting patterns, it’s too good!

**Part 2: SAS**

## 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?

* First we generate the simulated data:

data sim\_basket(keep=basket);  
 do i = 1 to 133;  
 flip = rand("Bernoulli",0.45);  
 if flip = 1 then basket="H"; else basket="M";  
 output;  
 end;  
run;

* Then we use calc\_streak to summarise the streaks

%calc\_streak(dset=sim\_basket, streakvar=basket, outset=sim\_streak);

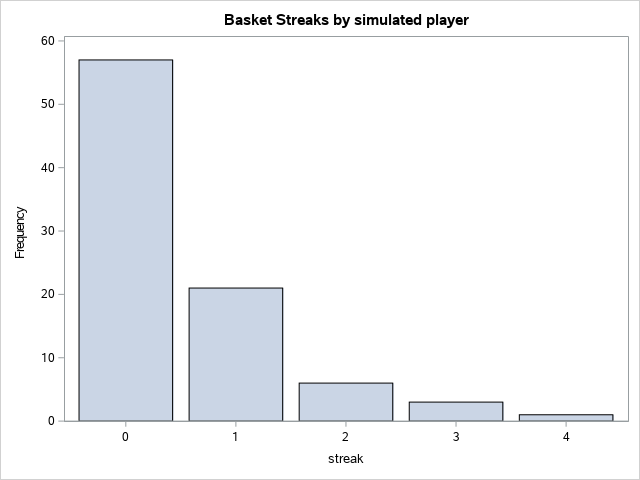
* Then we plot it

proc sgplot data=sim\_streak;

title "Basket Streaks by simulated player";

vbar streak;

run;



The distribution has a right skew. As before we should exclude a streak length of zero because those are just misses, not a streak. The typical streak length is 1 – hardly a streak. The longest streak from *this* run of the simulation us 4, much like Kobe’s.

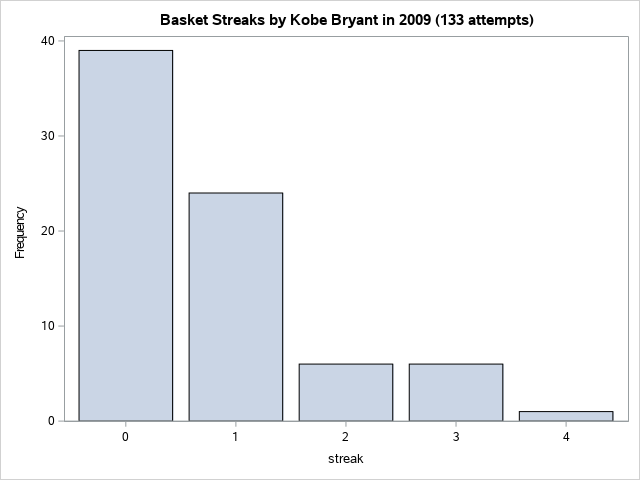
**2. If you were to run the simulation of the independent shooter a second time, how would you expect its streak distribution to compare to the distribution from the question above? Exactly the same? Somewhat similar? Totally different? Explain your reasoning.**

Just as before, no code is required to answer this question. Running the simulation a second time will still maintain the right skew, but the values of the streaks will change. The distribution would probably be similar, but with different values for each of the streak lengths. The simulation doesn’t include proportions of streak lengths, just random selections of ‘hit’ or ‘miss’.

## 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.

Kobe’s distribution is gathered using code from the lab:

proc sgplot data=kobe\_streak;  
 title "Basket Streaks by Kobe Bryant in 2009 (133 attempts)";  
 vbar streak;  
run;



Comparing this to the Sim streak plot above, the distributions are more similar than the simulation run in R! The maximums are both 4, with a streak of 1 being the most common. Both have right skew, but the simulation missed *many* more shots – nearly twenty more. I still don’t believe this simulated model accurately fits Kobe’s shooting pattern, even with the 45% accuracy rate, the misses seem significantly different.