Probability

The Hot Hand

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. This paper started a great controversy that continues to this day, as you can see by Googling *hot hand basketball*.

Data

Your 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. The data file we'll use is called `kobe_basket`.

This data frame contains 133 observations and 6 variables. The `shot` variable in this dataset indicates whether the shot was a hit (`H`) or a miss (`M`).

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:

$$\text{H M} \mid \text{M} \mid \text{H H M} \mid \text{M} \mid \text{M} \mid \text{M}$$

You can verify this by viewing the first 9 rows of the data in the data viewer.

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 length of 1 would be a made shot preceded by a miss (or starting the dataset) and followed by a miss (or ending the dataset)

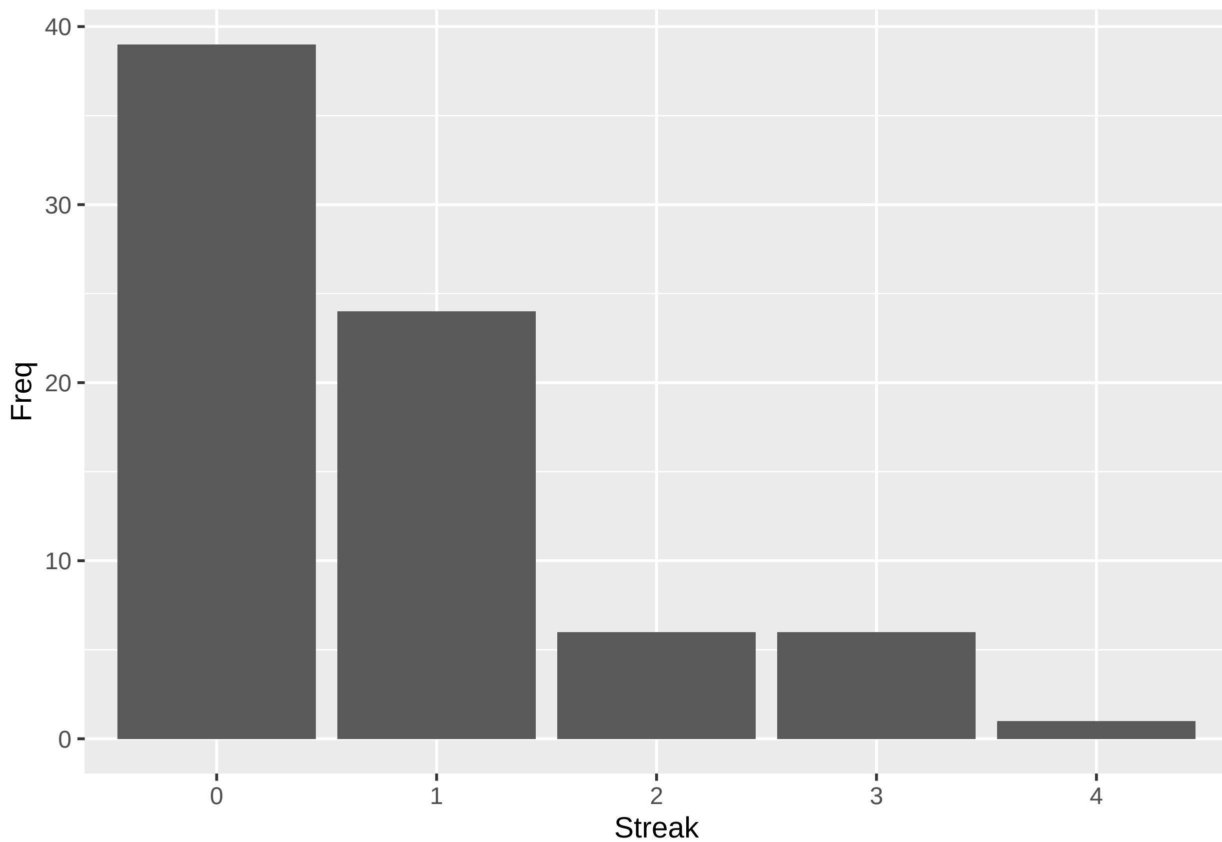
   The point is a streak length of 1 is when Kobe made one shot but not two consectitive shots.

   A streak length of 0 is any subset of 2 consective misses.

```
kobe_streak <- calc_streak(kobe_basket$shot)
```

We can then take a look at the distribution of these streak lengths.

```
kobe_streak_df<- as.data.frame(kobe_streak)

kobe_streak_df2<-as.data.frame(table(kobe_streak_df))
names(kobe_streak_df2)<-c("Streak","Freq")

ggplot(data=kobe_streak_df2, aes(y=Freq, x=Streak)) +
  geom_bar(stat="identity")
```



1. 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.

The typical streak length is just 1. His longest streak is 4 which he only accomplished once.

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 a hot hand? 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 effect 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 a hot hand? We can compare his streak lengths to someone without a hot hand: 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.

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

The vector `coin_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(coin_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
table(sim_fair_coin)
```

Since there are only two elements in `coin_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.

```
# odds of a head is 20%
sim_unfair_coin <- sample(coin_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%.

1. 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.

```
suc_results<-table(sim_unfair_coin)

sprintf("In the 100 trials, without seed,
        where P(H) was .20, the number of heads was, %d", suc_results["heads"])
```

A note on setting a seed: Setting a seed will cause R to select the same sample each time you knit your document. This will make sure your results don't change each time you knit, and it will also ensure reproducibility of your work (by setting the same seed it will be possible to reproduce your results). You can set a seed like this:

```
set.seed(35797)                  # make sure to change the seed
sim_unfair_coin <- sample(coin_outcomes, size = 100, replace = TRUE,
                          prob = c(0.2, 0.8))
suc_results<-table(sim_unfair_coin)

sprintf("In the 100 trials, with seed of 35797,
        where P(H) was .20, the number of heads was, %d", suc_results["heads"])
```

The number above is completely arbitraty. If you need inspiration, you can use your ID, birthday, or just a random string of numbers. The important thing is that you use each seed only once in a document. Remember to do this **before** you sample in the exercise above.

In a sense, we've shrunken the size of the slip of paper that says "heads", making it less likely to be drawn, and we've increased the size of the slip of paper saying "tails", making it more likely to be drawn. When you simulated the fair coin, both slips of paper were the same size. This happens by default if you don't provide a `prob` argument; all elements in the `outcomes` vector have an equal probability of being drawn.

Simulating the Independent Shooter

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

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

To make a valid comparison between Kobe and your simulated independent shooter, you need to align both their shooting percentage and the number of attempted shots.

1. 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`.

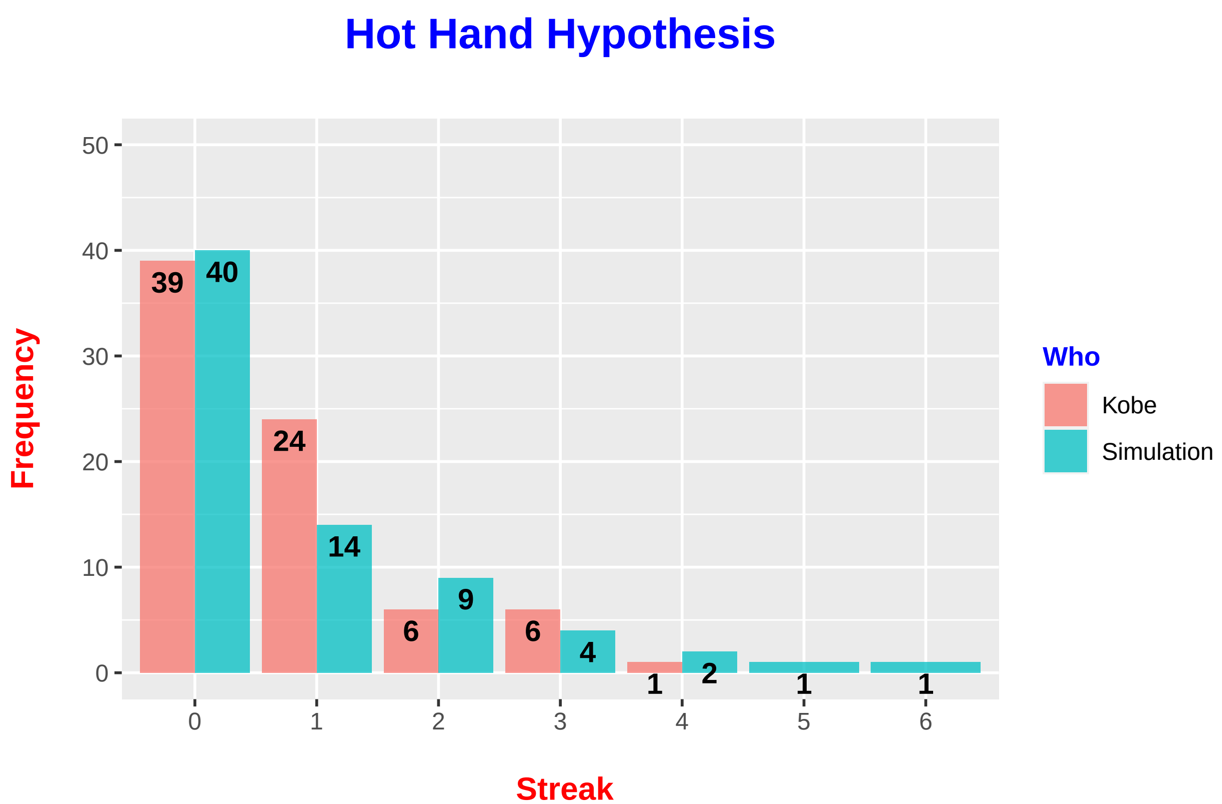
```
# simulate 133 trials, with a 45% chance of Hitting the shot
sim_basket <- sample(shot_outcomes, size = 133, prob = c(0.45, 0.55), replace = TRUE)
suc_results<-table(sim_basket)

sprintf("In the 133 trials, with no seed,
        where P(H) was .45, the number of heads was, %d", suc_results["H"])
```

With the results of the simulation saved as `sim_basket`, you have the data necessary to compare Kobe to our independent shooter.

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.

```
# kobe_streak <- calc_streak(kobe_basket$shot)
kobe_streak_tbl<-table(kobe_streak_df)


set.seed(35797)
sim_basket <- sample(shot_outcomes, size = 133, prob = c(0.45, 0.55), replace = TRUE)

sim_streak <- calc_streak(sim_basket)
sim_streak_tbl<-table(sim_streak)


# Example of side by side bar plot


# create a column of a vector of instances of "Kobe" and "Simulation",
#               the number of instances match the number of bars to plot

column_who <- c(rep("Kobe", nrow(as.data.frame(kobe_streak_tbl))),
                rep("Simulation", nrow(as.data.frame(sim_streak_tbl))))

# create a column for each frequency
column_freq <- c(as.vector(kobe_streak_tbl), as.vector(sim_streak_tbl))

# create a column for each streak
column_streak <- c(names(kobe_streak_tbl),names(sim_streak_tbl))


# matrix array
```

```r
outcome_data <- cbind(column_who, column_streak, column_freq)


outcome_data_df<-data.frame(Who = factor(column_who, levels = c("Kobe", "Simulation")),
                            Freq = column_freq, Streak=column_streak)

# create a side by side bar chart
#  the fill parameter controls the Kobe/Simulation bars
# x and y are the Streak and Count values
ggplot(data = outcome_data_df, aes(x = Streak, y = Freq, fill = Who)) +
  geom_bar(stat = "identity", position = position_dodge(), alpha = 0.75)  +
  ylim(0,50) +
  geom_text(aes(label = Freq), fontface = "bold", vjust = 1.5,
             position = position_dodge(.9), size = 4) +         # controls the number at top of bar
  labs(x = "\n Streak", y = "Frequency\n", title = "\n Hot Hand Hypothesis \n") +
  theme(plot.title = element_text(hjust = 0.5, face="bold", colour="blue", size = 16),      # center th
        axis.title.x = element_text(face="bold", colour="red", size = 12),
        axis.title.y = element_text(face="bold", colour="red", size = 12),
        legend.title = element_text(face="bold", colour="blue", size = 10))
```



The side by side bar plot suggests that a purely random distributin of a 45 percent success rate can produce streaks exceeding that of a player who shoots at a 50 percent rate.

Note: I feel like I should give credit to **this presentation** which explained how to create a side by side bar plot very well.

More Practice

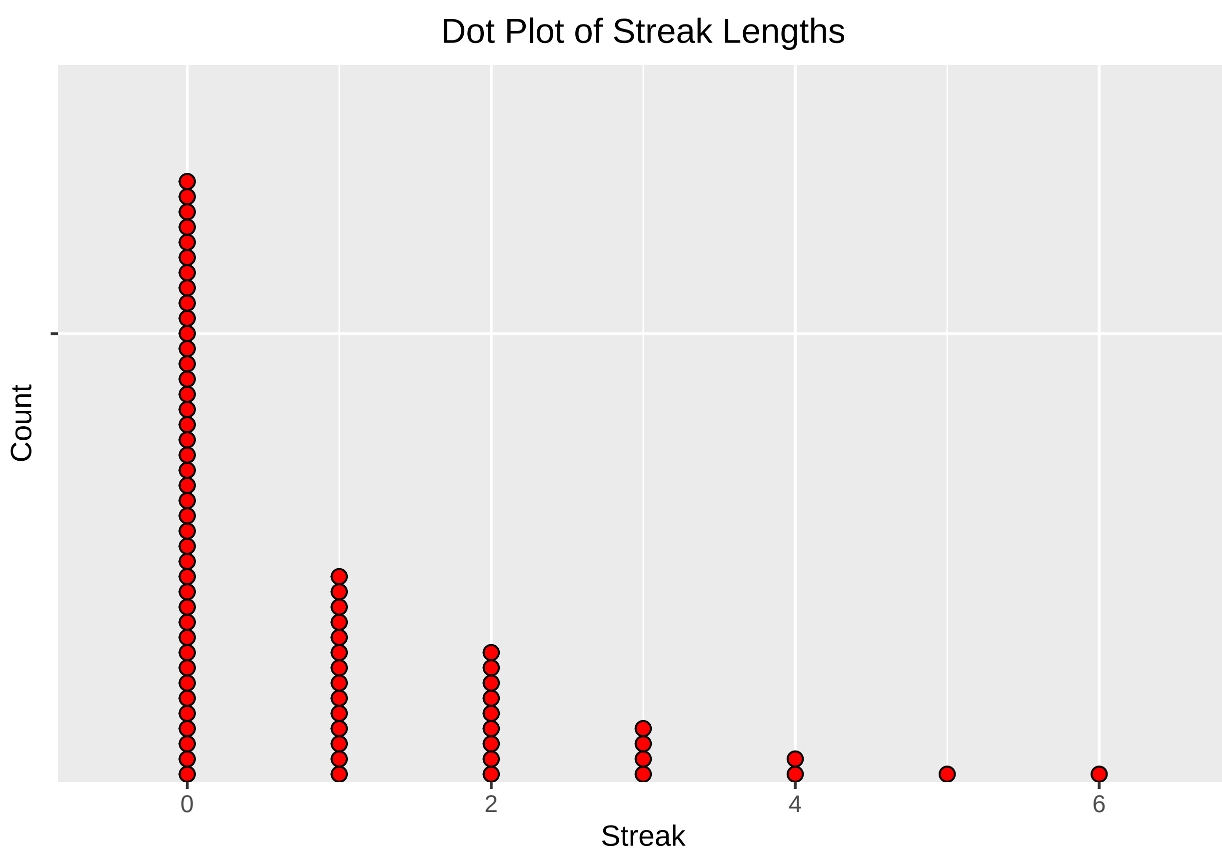
Comparing Kobe Bryant to the Independent Shooter

1. Using `calc_streak`, compute the streak lengths of `sim_basket`, and save the results in a data frame called `sim_streak`.

```
sim_streak_df<-as.data.frame(calc_streak(sim_basket))
names(sim_streak_df)<-c("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? Make sure to include a plot in your answer.

```
ggplot(sim_streak_df, aes(x = Streak, y="")) +
  geom_dotplot(binwidth = 1, width = 0.2, dotsize = .1, fill="red") +
  labs(title="Dot Plot of Streak Lengths", y="Count") +
   theme(plot.title = element_text(hjust = 0.5))          # center the title
```



1. 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.

In theory, the software is returning an independent result every time. so streaks are to be expected. The odds of getting 6 successes in a row are 1 in 64. I think in 133 trials you would not expect to get 6 success in a row.

The Law of Large Numbers states that as more observations of collected, the proportion $\hat{p}$ of occurrences with a particular outcome converges to the probability of P of that outcome.

1. 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.

As a sports fan, I can point out several issues with this study. But as an exercise, I would conclude that Kobe Bryant did not have the "hot hand" during the 2009 NBA Finals.