# Hypothesis Testing

**Your reproducible lab report:** Before you get started, download the R Markdown template for this lab. Remember all of your code and answers go in this document:

```
download.file("https://dyurovsky.github.io/85309/post/rmd/lab5.Rmd",
              destfile = "lab5.Rmd")
```

## The Hot Hands Phenomenon

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 (http://www.sciencedirect.com/science/article/pii/0010028585900106) 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*.

We do not expect to resolve this controversy today. However, in this lab we'll apply one approach to answering questions like this. The goals for this lab are to (1) think about the effects of independent and dependent events, (2) learn how to simulate shooting streaks in R, and (3) to compare a simulation to actual data in order to determine if the hot hand phenomenon appears to be real.

## Getting Started

Our investigation will focus on the performance of one player: Stephen Curry of the Gold State Warriors. Curry was the NBA *Most Valuable Player* for both the 2014 and 2015 seasons–the second time by unanimous vote. Maybe that's because of hot hands? We'll be looking at some data on Curry that I pulled from NBA Stats (http://stats.nba.com/).

Let's read in the data and see what we have.

```
curry_data <- read_csv("https://dyurovsky.github.io/85309/data/lab5/curry_data.csv")
```

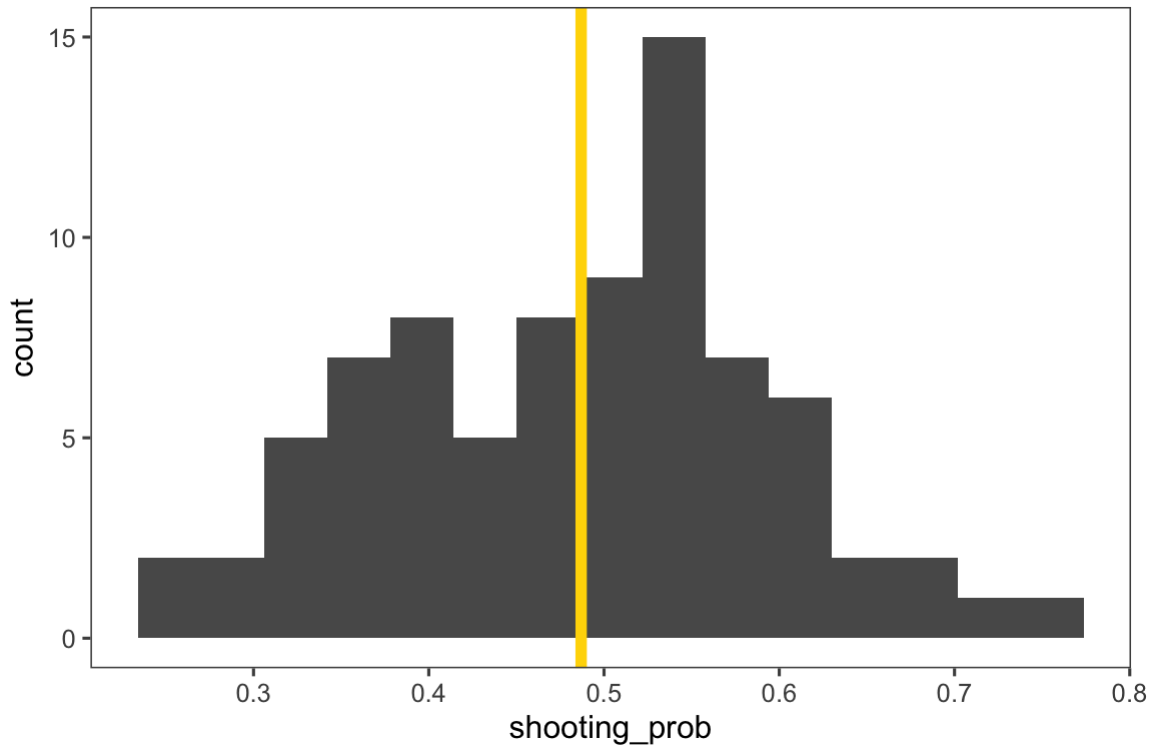
Let's take a look at this new data set: `curry_data` . Try using `View(curry_data)` in your console (or clicking on the data frame in the Environment tab). This tibble has information about every shot that Steph Curry took during the 2014 regular season. There are 1341 observations and 14 variables, where every row records a single shot taken. The `SHOT_MADE` variable in this dataset indicates whether the shot scored ( `TRUE` ) or was a miss ( `FALSE` ).

As always, let's do a little bit of exploratory data analysis to make sure we understand our data. Let's do two things. First, let's compute Curry's average shooting success. Second, let's use the `group_by` and `summarise` functions from the `tidyverse` to look at his shooting success from game to game.

```
mean_shooting_prob <- curry_data %>%
  summarise(mean = mean(SHOT_MADE)) %>%
  pull()

by_game_data <- curry_data %>%
  group_by(GAME_ID) %>%
  summarise(shooting_prob = mean(SHOT_MADE))

ggplot(by_game_data, aes(x = shooting_prob)) +
  geom_histogram(bins = 15) +
  geom_vline(aes(xintercept = mean_shooting_prob), color = "gold", size = 2)
```




**Exercise 1** Describe the distribution that we're seeing. What proportion of his shots does Curry make on average, and how does this vary from game to game?

Now let's use this data to ask whether Curry's shots show evidence of hot hands. One way we can approach this question is to look at whether Curry is more likely to make his next shot *if he has just made his previous shot*.

---

**Exercise 2** Come up with a Null Hypothesis and Alternative Hypothesis that we can use to ask whether Curry's shots provide evidence for the hot hands phenomenon.

We can use the `lag` function to make our life a lot easier. This will let us compare each row in the data frame with the row that came just before it. Because the data are already arranged chronologically, each shot in our tibble comes right after the shot it followed.

Let's try this out.

```
lag_data <- curry_data %>%
  group_by(GAME_ID) %>%
  mutate(lag_shot = lag(SHOT_MADE))
```

I'll `select` just the three relevant variables to make looking at the data easier, and then print out the first 10 rows

```
lag_data %>%
  select(GAME_ID, SHOT_MADE, lag_shot) %>%
  head(10)
```

```
## # A tibble: 10 × 3
## # Groups:   GAME_ID [1]
##    GAME_ID    SHOT_MADE lag_shot
##    <chr>      <lgl>     <lgl>
##  1 0021400014 TRUE      NA
##  2 0021400014 FALSE     TRUE
##  3 0021400014 TRUE      FALSE
##  4 0021400014 TRUE      TRUE
##  5 0021400014 FALSE     TRUE
##  6 0021400014 FALSE     FALSE
##  7 0021400014 FALSE     FALSE
##  8 0021400014 FALSE     FALSE
##  9 0021400014 FALSE     FALSE
## 10 0021400014 TRUE      FALSE
```

You can see that the `lag_shot` column contains exactly the value that is in `SHOT_MADE` in the previous row.

---

**Exercise 3** Why did I `group_by` game first before calling the `lag` function? Hint: The `NA` you see in the first row means that there is no value that came before.

We can now use the new `lag_shot` variable to group shots by whether they followed a successful or missed shot. And thus we can test our Alternative Hypothesis. Let's do one more quick exploratory data analysis to see what the difference in Curry's shooting percentage is like following successful vs. unsuccessful shots.

```
eda_hothands <- lag_data %>%
  group_by(lag_shot) %>%
  summarise(shooting_prob = mean(SHOT_MADE))
```

**Exercise 4** What do you see when you run this code? Does it look like we see evidence for the hot hands phenomenon? Also, what does the value for `shooting_prob` for `NA` tell us?

# Simulations in R

Now we're ready to test our hypothesis formally. What we need to do is generate the Null Distribution for what kind of differences in shots following successful vs. unsuccessful shots we should see if there is no hot hands phenomenon–if Curry shoots identically following successful vs. unsucessful shots. We'll do this through sampling just like we did in lecture using the `sample` function.

So what we want to do, is keep everything about the shooting data identical–e.g. how many shots Curry took, how many of them were successful, etc. The only thing we want to do is randomly determine whether these shots came after other shots that were successful.

```
lag_data %>%
  group_by(GAME_ID) %>%
  group_by(lag_shot, GAME_ID) %>%
  summarise(mean = mean(SHOT_MADE))
```

```
# Number of shots taken after shots that were made
hot_shots <- lag_data %>%
  filter(lag_shot) %>%
  nrow() # count the number of rows

# Number of shots made after shots that were made
hot_made <- lag_data %>%
  filter(lag_shot & SHOT_MADE) %>%
  nrow()

# Number of shots taken after shots that were missed
not_shots <- lag_data %>%
  filter(!lag_shot) %>%
  nrow()

# Number of shots made after shots that were missed
not_made <- lag_data %>%
  filter(!lag_shot & SHOT_MADE) %>%
  nrow()

simulate_null <- function() {

  # Make a list with the right number of shots of each type
  shots <- c(rep("Hot", hot_shots), rep("Not", not_shots))

  # randomly select the made shots from this list
  made <- sample(shots, hot_made + not_made)

  # Compute the difference shot success between hot and not shots
  random_hot_made <- sum(made == "Hot") / hot_shots
  random_not_made <- sum(made == "Not") / not_shots

  random_hot_made - random_not_made
}
```

**Exercise 5** Based on the code (https://dyurovsky.github.io/85309/post/rmd/cardiac.Rmd) we looked at during the Unit 2 lectures, generate a Null distribution using this sampling function, and compare it to the empirical difference we saw in Curry's shots. Do we see any evidence of hot hands? Or maybe cold hands?

# More Practice

**Exercise 6**

Where does the empirical difference between shots following success and failures fall on the null distribution (what percentile?) What does this mean for our alternative hypothesis?

**Exercise 7**

Come up with your own hypothesis to test using Steph Curry's shooting data. Feel free to use any of the columns already in the tibble, or to make a new column by mutating the current columns in some way (like we did with lag_shot).

**Exercise 8**

Test your hypothesis by modifying the `simulate_null` function to generate your own null distribution, and compare it to the empirical mean. In your response, make both a plot like you did in Exercise 5, and tell me what percentile the empirical data falls in like you did in Exercise 6.

**Exercise 9**

Do you think that there are any reasons that this is not a compelling test of the hot hands hypothesis? If not, why not? The debate has raged on, and some recent papers have purported to show stronger evidence in favor of hot hands. Feel free to look these up if you want inspiration for your answer, but make sure to cite them in your response if you do.