# Lab 5 - Hypothesis Testing

Sanaz Saadatifar

3/1/2022

---

# Lab report

**Load data**

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

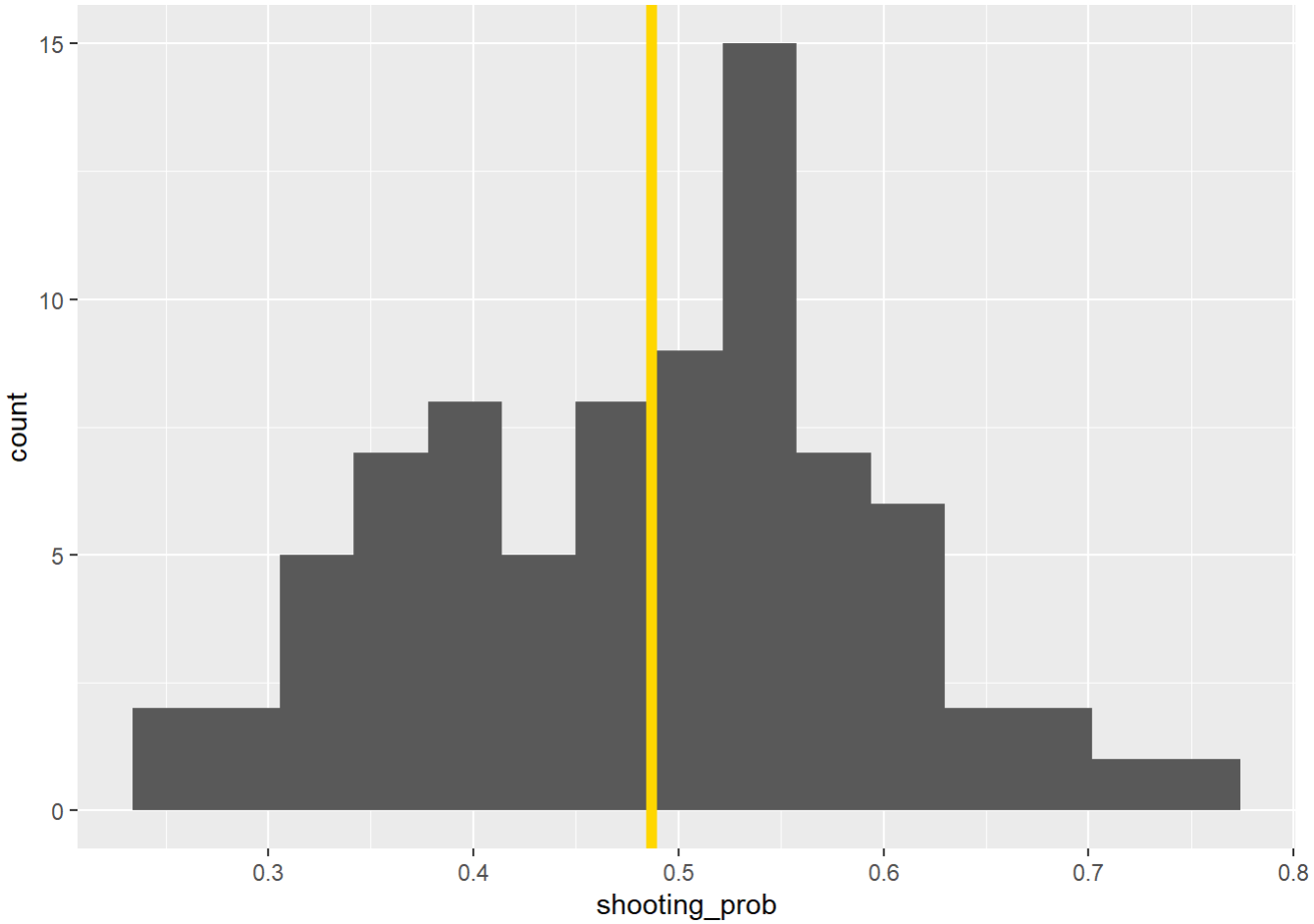
## Exercise 1:

```
# enter your code for Exercise 1 here
mean_shooting_prob <- curry_data %>%
  summarise(mean = mean(SHOT_MADE)) %>%
  pull()

sd_shooting_prob <- curry_data %>%
  summarise(mean = sd(SHOT_MADE)) %>%
  pull()

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




The distribution looks roughly normal and symmetric. The mean is around .5, and the the spread is pretty broad– there look to be games where Steph makes around 60-70% of his shots, and also games where he makes 30%

## Exercise 2:

H0: percent of shots made after making a shot is not different from the percent of shots made after missing a shot HA: percent of shots made after making a shot is different from the percent of shots made after missing a shot

## Exercise 3:

```
# enter your code for Exercise 3 here
lag_data <- curry_data %>%
  group_by(GAME_ID) %>%
  mutate(lag_shot = lag(SHOT_MADE))

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

```
## # A tibble: 10 x 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
```

I don't want the first shot of a game to count as coming after the last shot of the previous game–I want to throw it out from my analysis and not count it as a shot after a missed shot or a made shot

## Exercise 4:

```
# enter your code for Exercise 4 here
eda_hothands <- lag_data %>%
  group_by(lag_shot) %>%
  summarise(shooting_prob = mean(SHOT_MADE))

eda_hothands
```

```
## # A tibble: 3 x 2
##   lag_shot shooting_prob
##   <lgl>            <dbl>
## 1 FALSE            0.512
## 2 TRUE             0.460
## 3 NA               0.488
```

This data doesn't look consistent with the hot hands hypothesis because Steph is *less* likely to make a shot after he made his last shot relative to missing his last shot. The NA values tell me about the first shot in the game that Steph takes.

## Exercise 5:

```
# enter your code for Exercise 5 here
lag_data %>%
  group_by(GAME_ID) %>%
  group_by(lag_shot, GAME_ID) %>%
  summarise(mean = mean(SHOT_MADE))
```

```
## `summarise()` has grouped output by 'lag_shot'. You can override using the `.groups` argumen
t.
```

```
## # A tibble: 240 x 3
## # Groups:   lag_shot [3]
##    lag_shot GAME_ID     mean
##    <lgl>    <chr>      <dbl>
##  1 FALSE    0021400014 0.444
##  2 FALSE    0021400038 0.667
##  3 FALSE    0021400042 0.364
##  4 FALSE    0021400068 0.556
##  5 FALSE    0021400087 0.667
##  6 FALSE    0021400095 0.556
##  7 FALSE    0021400108 0.4
##  8 FALSE    0021400121 0.667
##  9 FALSE    0021400140 0.571
## 10 FALSE    0021400145 0.5
## # ... with 230 more rows
```

```
# 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
}

simulate_null()
```

```
## [1] 0.03363379
```

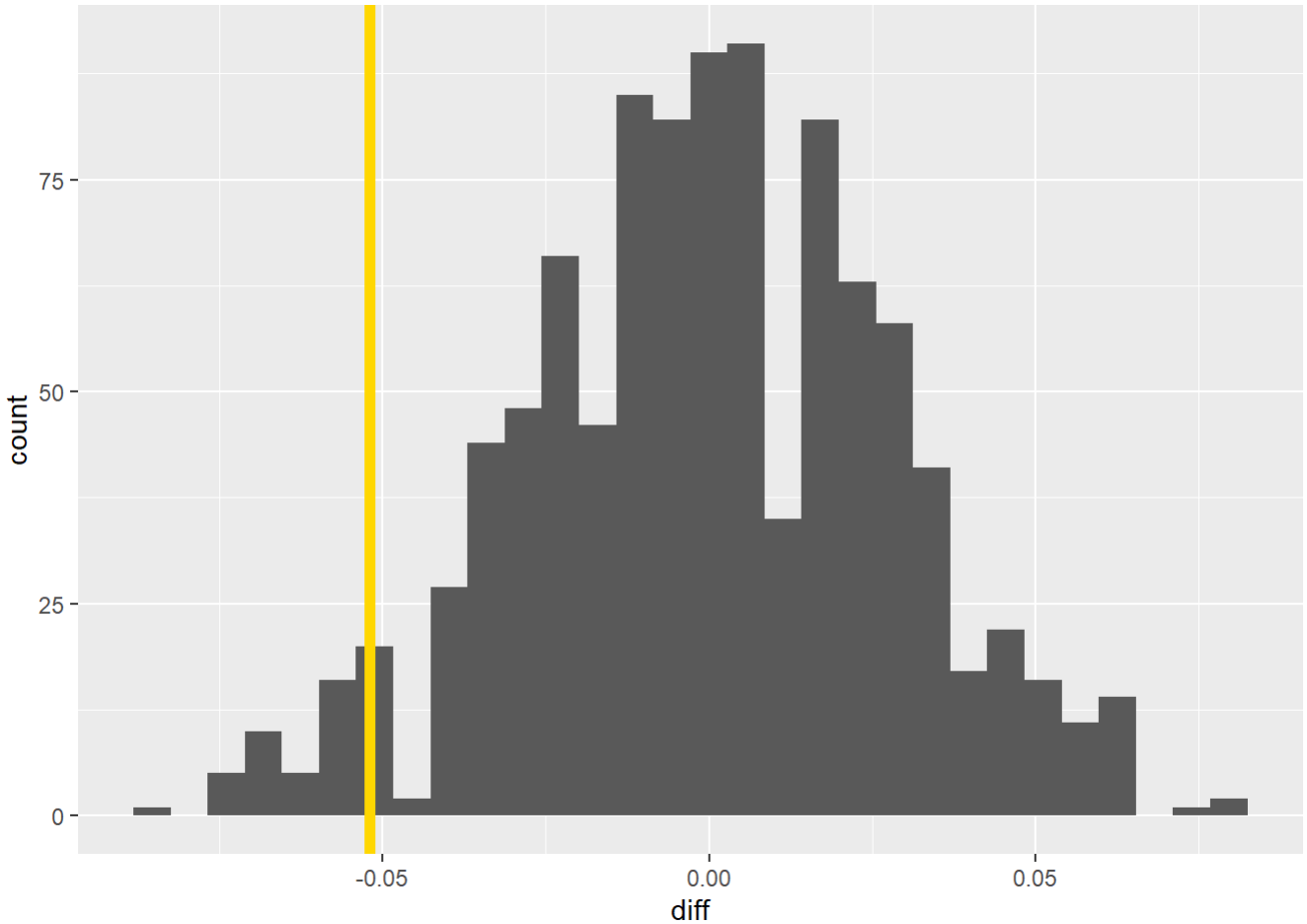
```
null_diffs <- tibble(sample = 1:1000,
                     diff = replicate(1000, simulate_null()))


empirical_diff <- (hot_made / hot_shots) - (not_made / not_shots)
empirical_diff
```

```
## [1] -0.05204103
```

```
ggplot(null_diffs, aes(x = diff)) +
  geom_histogram() +
  geom_vline(aes(xintercept = empirical_diff), color = "gold", size = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




It's hard to tell exactly from the plot, but it looks like Steph might be reliably worse after he made his last shot. So maybe some evidence for cold hands.

## Exercise 6:

```
# enter your code for Exercise 6 here
mean(empirical_diff > pull(null_diffs, diff))
```

```
## [1] 0.037
```

The empirical data fall at around the 2nd percentile, which suggests that we should reject the null hypothesis (because the data are in the bottom 2.5% percent of the null distribution). But it's pretty borderline.

# More practice:

## Exercise 7:

```
# enter your code for Exercise 7 here

# Number of shots taken in last 10 seconds
ten_shots <- curry_data %>%
  filter(SECONDS_REMAINING<11) %>%
  nrow() # count the number of rows


# Number of shots made in last 10 seconds
ten_made <- curry_data %>%
  filter(SECONDS_REMAINING<11 & SHOT_MADE) %>%
  nrow()


# Number of shots taken in last 20 seconds - last 10 seconds
twenty_shots <- curry_data %>%
  filter(SECONDS_REMAINING<21, SECONDS_REMAINING>10) %>%
  nrow()

# Number of shots made in last 20 seconds -  last 10 seconds
twenty_made <- curry_data %>%
  filter(SECONDS_REMAINING<21, SECONDS_REMAINING>10 & SHOT_MADE) %>%
  nrow()
```

I am going to see whether he made more of his shots in the last 10 seconds of the game compared to the last 20 seconds (less than 20, more than 10 seconds remaining) H0: percent of shots made in the last 10 seconds of the game is not different from the percent of shots made between the last 20 seconds and 10 seconds HA: percent of shots made in the last 10 seconds of the game is different from the percent of shots made between the last 20 seconds and 10 seconds

## Exercise 8:

```
# enter your code for Exercise 8 here

simulate1_null <- function() {

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

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

  # Compute the difference shot success between hot and not shots
  random_ten_made <- sum(made == "Hot") / ten_shots
  random_twenty_made <- sum(made == "Not") / twenty_shots

  random_ten_made - random_twenty_made
}

simulate1_null()
```

```
## [1] 0.04661256
```

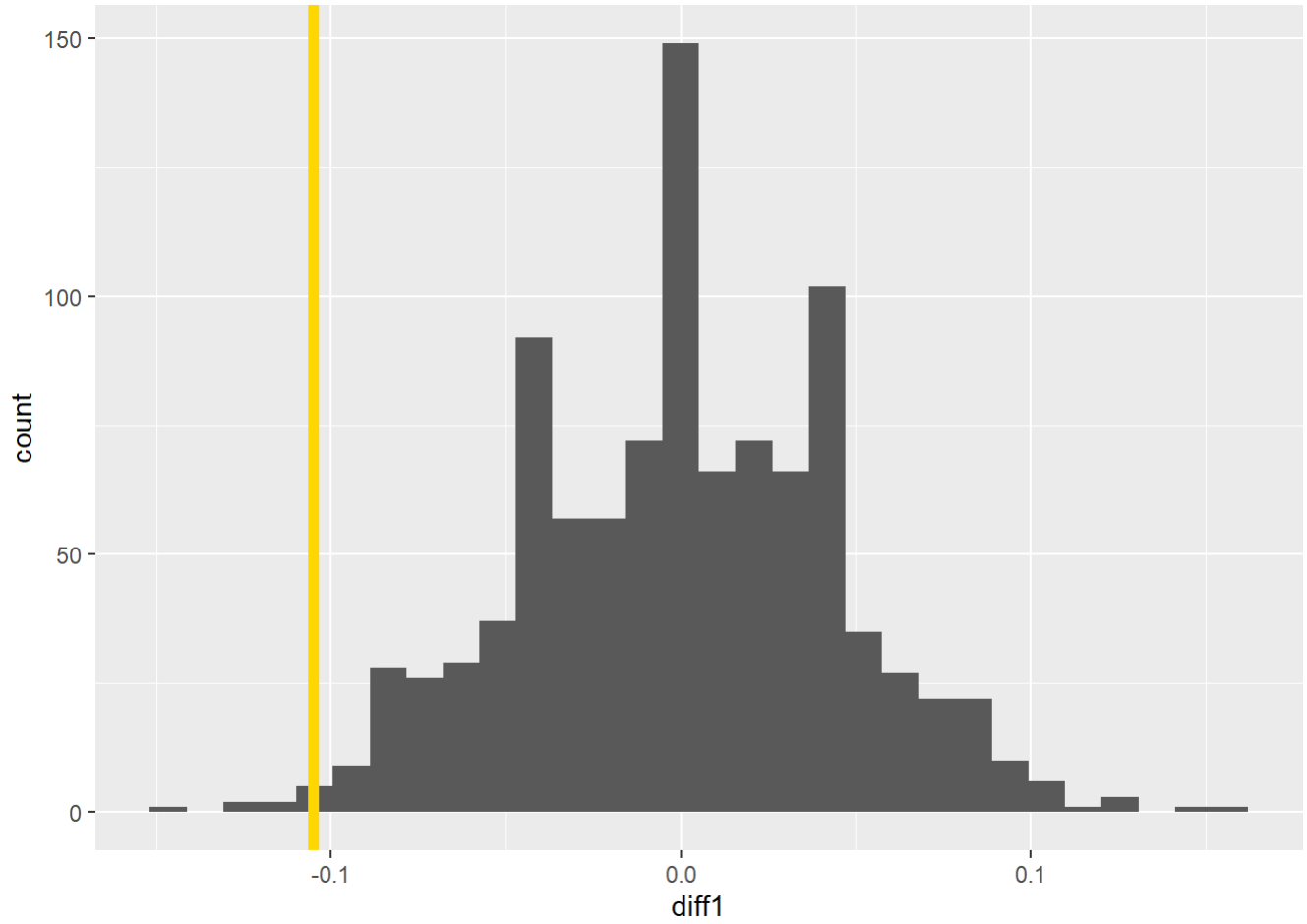
```
null_diffs1 <- tibble(sample = 1:1000,
                      diff1 = replicate(1000, simulate1_null()))


empirical_diff1 <- (ten_made / ten_shots) - (twenty_made / twenty_shots)
empirical_diff1
```

```
## [1] -0.105187
```

```
ggplot(null_diffs1, aes(x = diff1)) +
  geom_histogram() +
  geom_vline(aes(xintercept = empirical_diff1), color = "gold", size = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mean(empirical_diff1 > pull(null_diffs1, diff1))
```

```
## [1] 0.005
```

```
mean(empirical_diff1 > pull(null_diffs1, diff1))
```

```
## [1] 0.005
```

The plot shows that we can reject the null hypothesis, and percent of shots made by Steph in the last 10 seconds of the game is different (is less) from the percent of shots made by him between the last 20 seconds and 10 seconds

also, The empirical data fall at around the 1st percentile, which suggests that we should reject the null hypothesis (because the data are in the bottom 2.5% percent of the null distribution).

## Exercise 9:

I do not think this is a compelling test of the hot hands hypothesis, because we limited our test only on the success or failure of the two shots taken on a row. But in order to make a solid hypothesis we need to broaden our data set in terms of shots that we are observing. For example, instead of just one short, we can focus on two or more. Like, observing whether shots taken after two successful shots in a row, are more inclined to be successful or not (like whether the third, or fourth shots are also successful as the first two or not)