Homework 7

Tania Ommer

2023-12-08

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(moderndive)
library(skimr)
library(infer)

set.seed(20009345)

boxoffice <- read.csv("movie_boxoffice.csv", header=T) %>% distinct()
glimpse(boxoffice)
```

```
## Rows: 4,869
## Columns: 7
## $ Movie           <chr> "Raise the Titanic", "Flash Gordon", "Popeye", "The Fo~
## $ Month           <chr> "Aug", "Dec", "Dec", "Feb", "Jan", "Jan", "Jan", "Jan"~
## $ Day             <int> 1, 5, 12, 1, 1, 1, 1, 1, 4, 25, 6, 13, 20, 20, 20, 7, ~
## $ Year            <int> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, ~
## $ Budget          <dbl> 40.00, 35.00, 20.00, 1.00, 6.50, 0.35, 3.50, 35.00, 3.~
## $ Domestic_Gross  <dbl> 7.000000, 27.107960, 49.823037, 21.378361, 31.899000, ~
## $ Worldwide_Gross <dbl> 7.000000, 27.107960, 49.823037, 21.378361, 31.899000, ~
```

```
movie_sample <- boxoffice %>% rep_sample_n(size=200) %>%
  mutate(summer=ifelse(Month %in% c("Jun", "Jul", "Aug"), "1", "0"),
         yearperiod=ifelse(Year<=1999, "1", "2"),
         highglobal=ifelse(Worldwide_Gross>=Budget, "1", "0"))
```

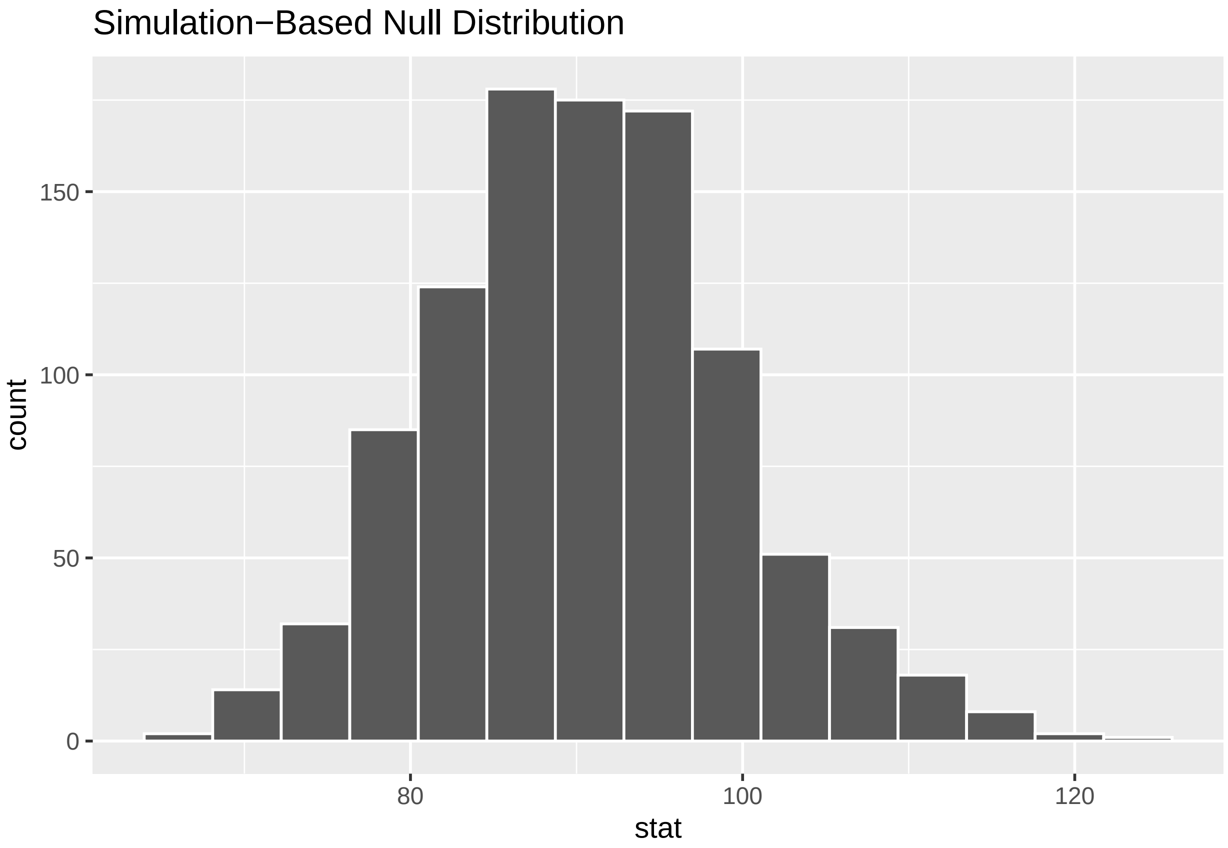
#Use the random sample with 200 movies you had from homework 5 to test

##1) Whether the average global box office earnings of all movies is different than $90 million.

###Replication/Simulation Method: ####Null Hypothesis (H0): The average global box office earnings is $90 million, mu = 90 million ####Alternative Hypothesis (H1): The average global box office earnings is different that $90 million, mu does not equal 90 million

#####One-sample mean test (two-tailed)

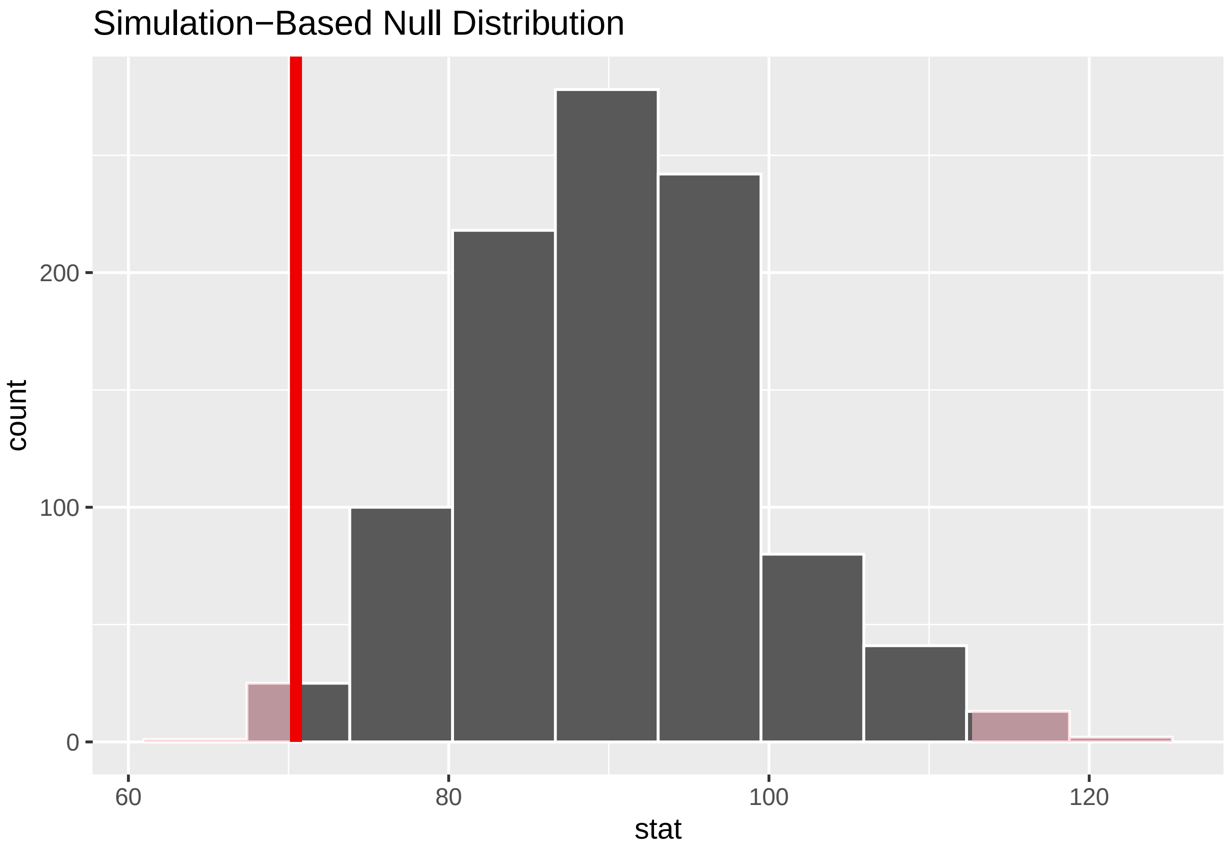
```
null_movie <- movie_sample %>%
  specify(response = Worldwide_Gross) %>%
  hypothesize(null = "point", mu = 90) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

visualize(null_movie)
```



```
sample_mean <- movie_sample %>%
  specify(response=Worldwide_Gross) %>%
  calculate(stat = "mean")

sample_mean
```

```
## Response: Worldwide_Gross (numeric)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  70.5
```

```
# visualize rejection region
visualize(null_movie, bins=10) +
  shade_p_value(obs_stat = sample_mean, direction = "two-sided")
```



```
p_value_movie <- null_movie %>%
  get_p_value(obs_stat = sample_mean, direction = "two-sided")

p_value_movie
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.024
```

The p-value is 0.024 which is less than the deafult significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is evidence that the average box office earnings are not $90 million.

###Theoretical Method, t-test: ####Null Hypothesis (H0): The average global box office earnings is $90 million, mu = 90 million ####Alternative Hypothesis (H1): The average global box office earnings is different that $90 million, mu does not equal 90 million

#####One-sample mean test (two-tailed)

```
t_test <- movie_sample %>%
  specify(response = Worldwide_Gross) %>%
  hypothesize(null = "point", mu = 90) %>%
  calculate(stat = "t")

t_test
```

```
## Response: Worldwide_Gross (numeric)
## Null Hypothesis: point
## # A tibble: 1 x 1
##    stat
##    <dbl>
## 1 -2.26
```

```
p_value_theoretical <- 2 * pnorm(abs(t_test$stat), lower.tail = FALSE)

p_value_theoretical
```

```
##          t
## 0.02399884
```

The p-value is 0.02399884 which is less than the default significance level of 0.05, therefore we reject the null hypothesis. We can conclude that there is evidence that the average box office earnings are not $90 million.

##2) Whether the average global box office earnings in the summer (June, July and August) is higher than the average global box office earnings in the rest of the year

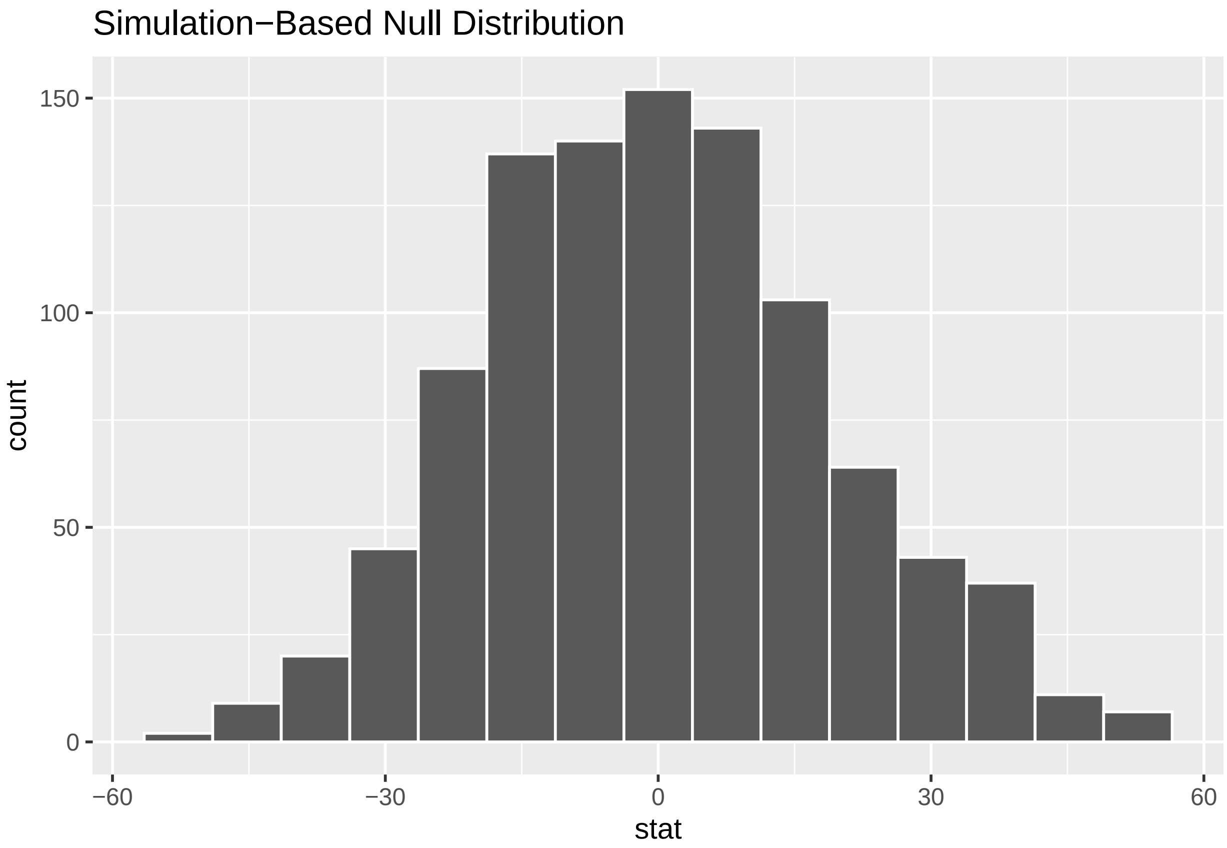
###Replication/Simulation Method:

#####Two-sample mean test (one-tailed)

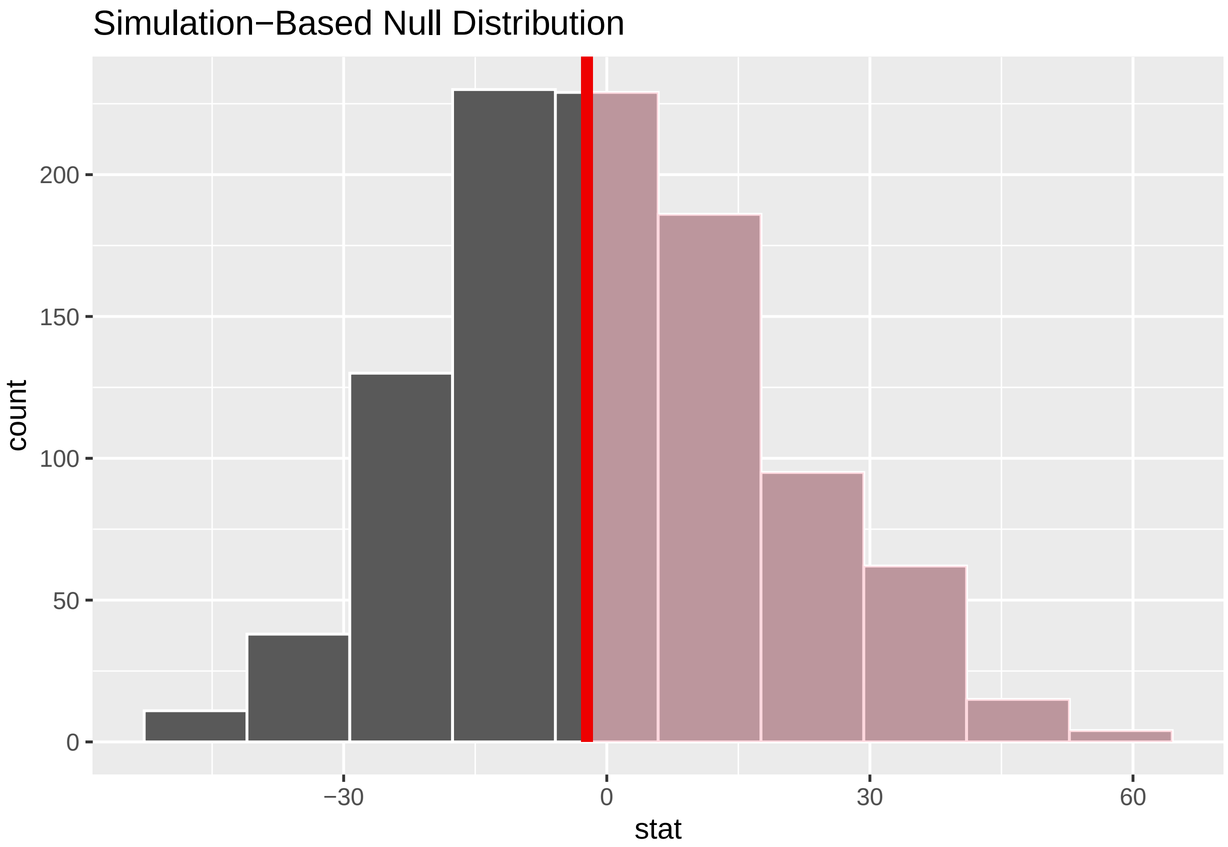
```
summer_rest <- movie_sample %>%
  specify(formula = Worldwide_Gross ~ summer) %>%
  hypothesize (null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("1", "0"))

diff_means <- movie_sample %>%
  specify(formula = Worldwide_Gross ~ summer) %>%
  calculate(stat = "diff in means", order = c("1", "0"))

diff_means
```

```
## Response: Worldwide_Gross (numeric)
## Explanatory: summer (factor)
## # A tibble: 1 x 1
##     stat
##    <dbl>
## 1 -2.26
```

```
visualize(summer_rest)
```



```
visualize(summer_rest, bins=10) +
  shade_p_value(obs_stat = diff_means, direction = "greater")
```



```
p_value_summer_rest <- summer_rest %>%
  get_p_value(obs_stat = diff_means, direction = "greater")

p_value_summer_rest
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.528
```

The p-value is 0.528 which is greater than the default significance value of 0.05. Therefore, we fail to reject the null hypothesis. There is not enough evidence to conclude a difference in average box office earning between the summer months and rest of the year.

###Theoretical Method, t-test: ####Null Hypothesis (H0): mu1 - mu2 = 0 ####Alternative Hypothesis (H1): mu1 - mu2 > 0 ####mu1 is summer months, mu2 is the rest of the months of the year

#####Two-sample mean test

```
t_test <- movie_sample %>%
  specify(formula = Worldwide_Gross ~ summer) %>%
  hypothesize (null = "independence") %>%
  calculate(stat = "t", order = c("1", "0"))

t_test
```

```
## Response: Worldwide_Gross (numeric)
## Explanatory: summer (factor)
## Null Hypothesis: independence
## # A tibble: 1 x 1
##      stat
##     <dbl>
## 1 -0.125
```

```
p_value_theoretical_2 <- pnorm(t_test$stat, 0,1, lower.tail = FALSE)

p_value_theoretical_2
```

```
##         t
## 0.5495617
```

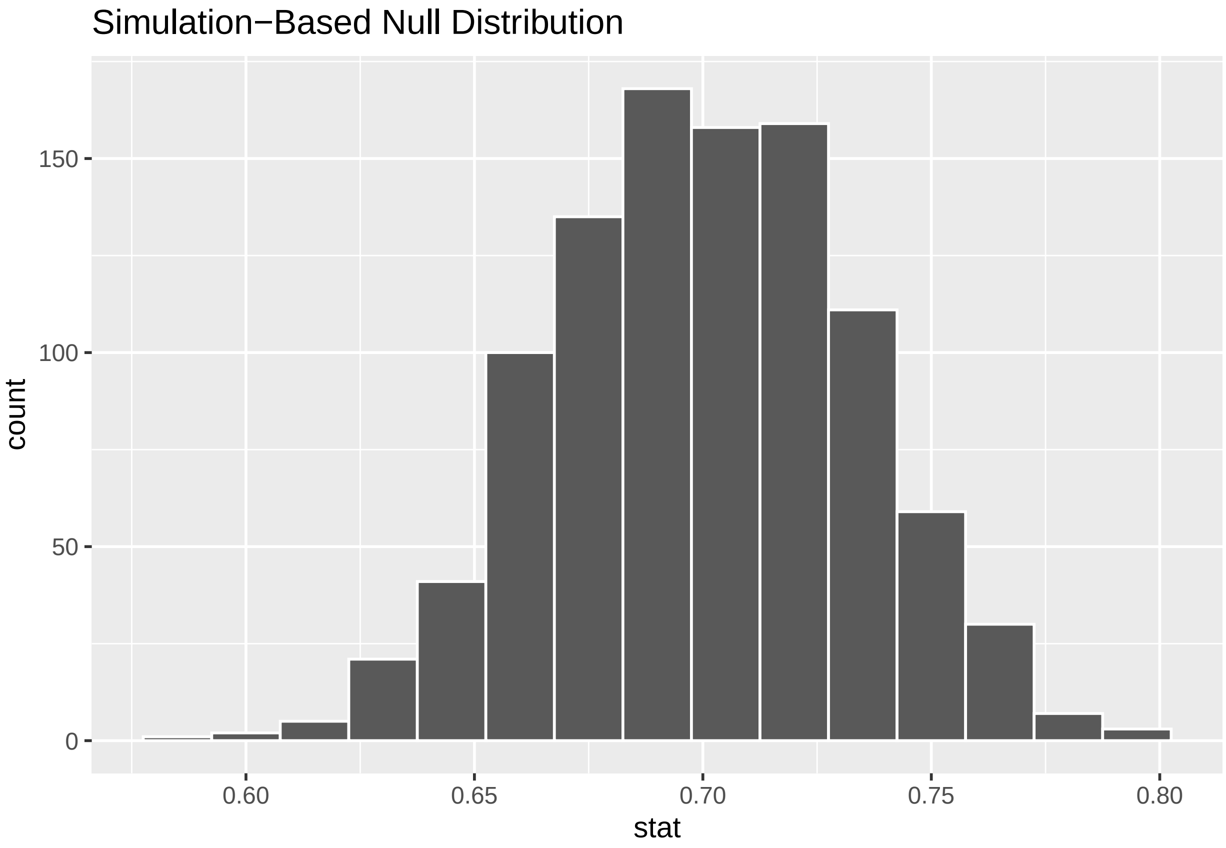
The p-value is 0.54956 which is greater than the default significance level of 0.05. Therefore, we fail to reject the null hypothesis (sample size > 30, allowing us to fail to reject). There is not enough evidence to suggest a difference in average box office eanrings between the summer months and the rest of the year.

##3) Whether the proportion of all movies whose global box office earning exceeds budget is different than 70%.

####Null Hypothesis (H0): The prop. is equal to 70%, mu = 0.7 ####Alternative Hypothesis (H1): The prop. is different than 70%, mu does not equal 0.7

#####One-sample proportion hypothesis testing (two-tailed)

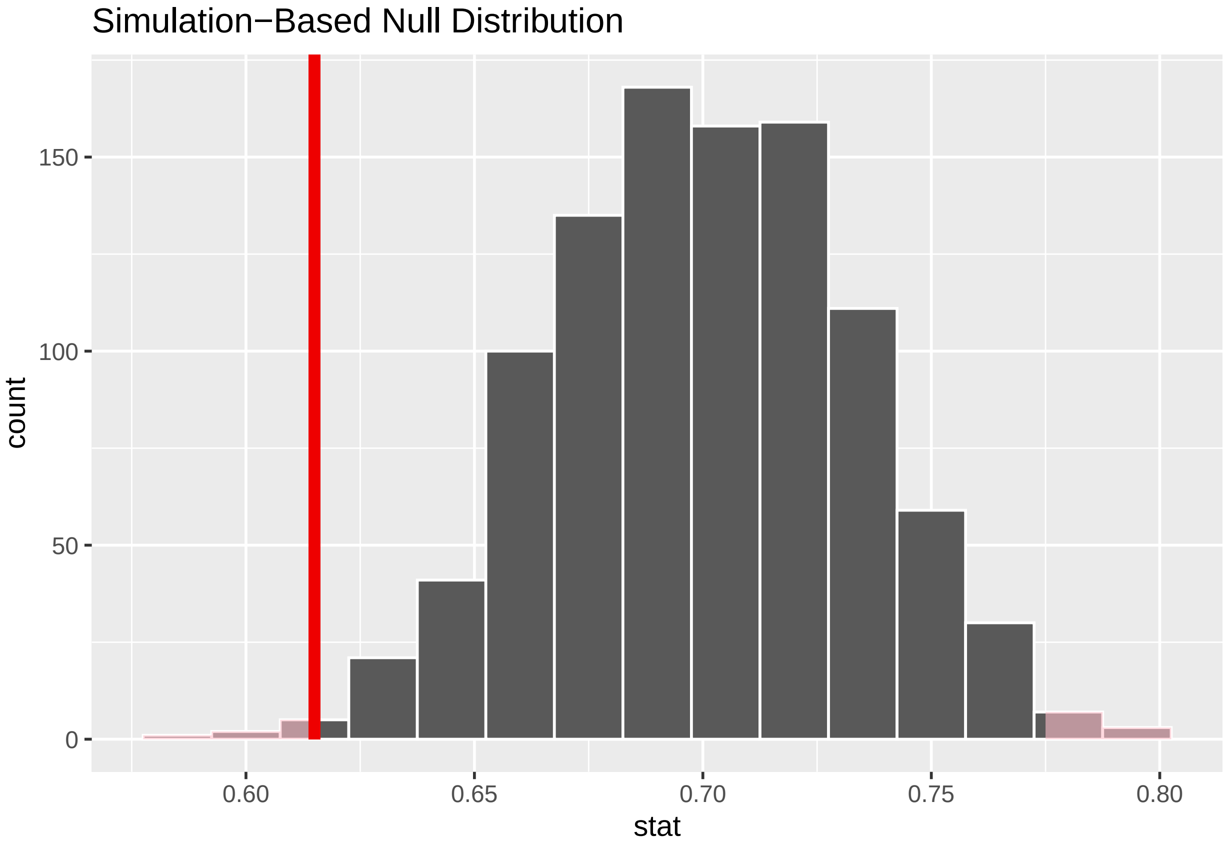
```
null_prop_test <- movie_sample %>%
  specify(response = highglobal, success = "1") %>%
  hypothesize(null = "point", p = 0.7) %>%
  generate(reps = 1000, type = "draw") %>%
  calculate(stat = "prop")

visualize(null_prop_test)
```



```
observed_prop <- movie_sample %>%
  specify(response = highglobal, success = "1") %>%
  calculate(stat = "prop")

observed_prop
```

```
## Response: highglobal (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.615
```

```
visualize(null_prop_test) +
  shade_p_value(obs_stat = observed_prop, direction = "two-sided")
```



```
p_value_prop_test <- null_prop_test %>%
  get_p_value(obs_stat = observed_prop, direction = "two-sided")

p_value_prop_test
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.014
```

The p-value is 0.014 which is less than the default significance value of 0.05. Therefore, we reject the null hypothesis and can conclude that there is evidence that the proportion is different from the given value 70%.

###Theoretical Method, z-test: ####Null Hypothesis (H0): The prop. is equal to 70%, mu = 0.7

####Alternative Hypothesis (H1): The prop. is different than 70%, mu does not equal 0.7

#####One-sample proportion hypothesis testing (two-tailed)

```
z_test <- movie_sample %>%
  specify(response = highglobal, success = "1") %>%
  hypothesize(null = "point", p=0.7) %>%
  calculate(stat = "z")

z_test
```

```
## Response: highglobal (factor)
## Null Hypothesis: point
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 -2.62
```

```
p_value_theoretical_3 <- 2 * pnorm(abs(z_test$stat), lower.tail = FALSE)

p_value_theoretical_3
```

```
## [1] 0.008711913
```

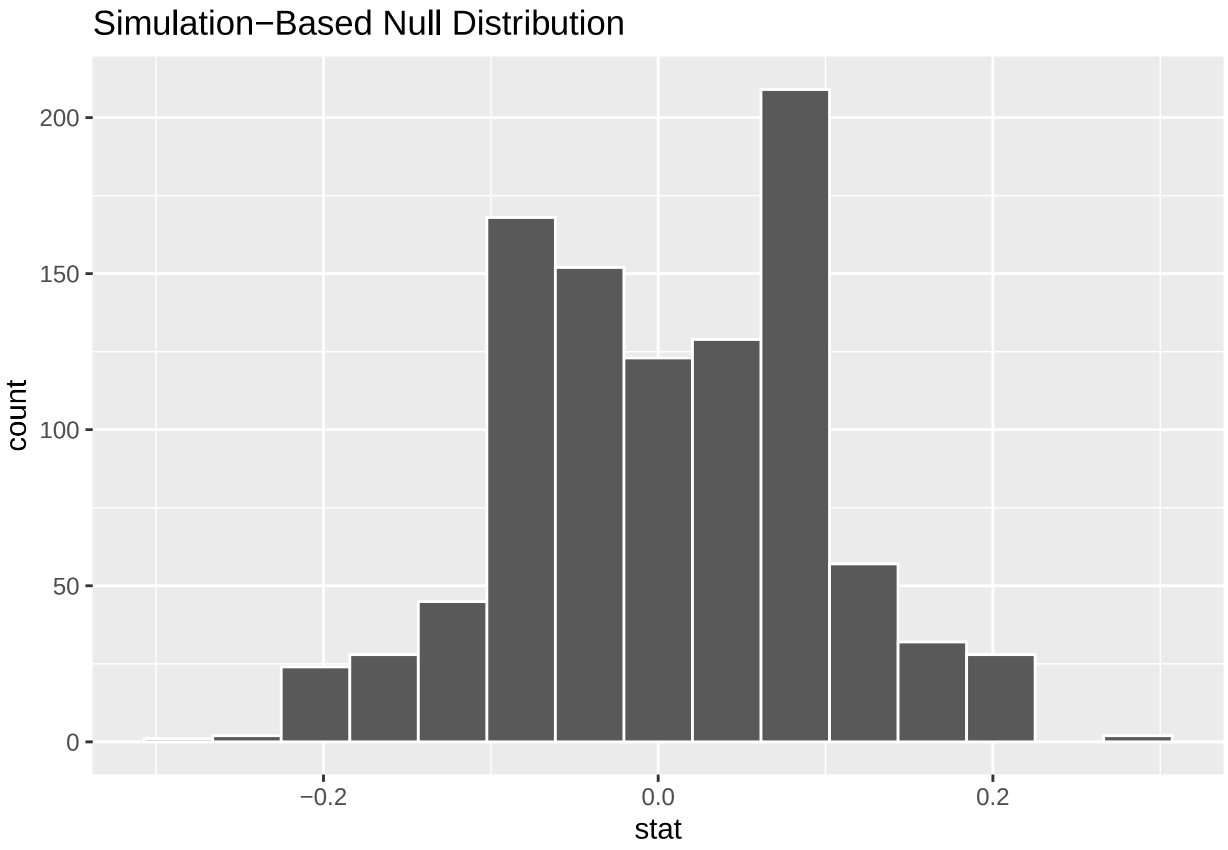
The p-value is 0.008712 which is less than the default significance value of 0.05. Therefore, we reject the null hypothesis (there are at least 15 success and failures so we can reject the null) and can conclude that there is evidence that the proportion is different from the given 70%

##4) Whether the proportion of movies released from 1980 to 1999 whose global box office earning exceeds budget is lower than the proportion of movies released from 2000 and 2018 whose global box office earning exceeds budget

###Replication/Simulation Method: ####Null Hypothesis (H0): p1 - p2 = 0, so p1 = p2 ####Alternative Hypothesis (H1): p1 - p2 < 0, so p1 < p2 ####p1 is before 2000

#####Two-sample proportion hypothesis test

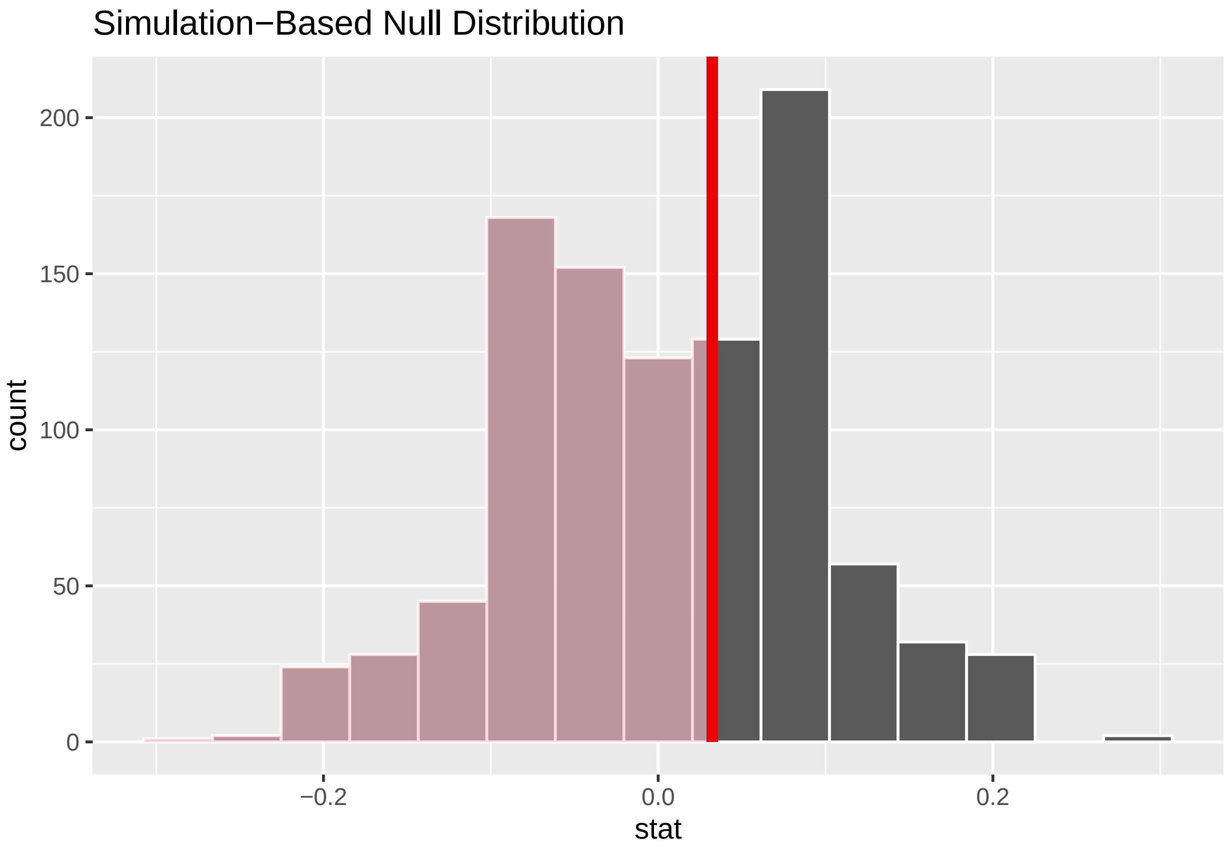
```
prop_test <- movie_sample %>%
  specify(formula = highglobal ~ yearperiod, success = "1") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("1", "2"))

visualize(prop_test)
```



```
observed_diff_prop <- movie_sample %>%
  specify(formula = highglobal ~ yearperiod, success = "1") %>%
  calculate(stat = "diff in props", order = c("1", "2"))

observed_diff_prop
```

```
## Response: highglobal (factor)
## Explanatory: yearperiod (factor)
## # A tibble: 1 x 1
##     stat
##    <dbl>
## 1 0.0323
```

```
visualize(prop_test) +
  shade_p_value(obs_stat = observed_diff_prop, direction = "left")
```



```
p_value_prop_test <- prop_test %>%
  get_p_value(obs_stat = observed_diff_prop, direction = "left")

p_value_prop_test
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.672
```

The p-value is 0.672 which is greater than the default significance level of 0.05, so we fail to reject the null hypothesis. It can be concluded that there is no evidence that the proportion of movies released from 1980 to 1999 whose global box office earnings exceeds budget is lower than the proportion of movies released from 2000 to 2018 whose global box office earnings exceeds budget.

###Theoretical Method, z-test: ####Null Hypothesis (H0): p1 - p2 = 0, so p1 = p2 ####Alternative Hypothesis (H1): p1 - p2 < 0, so p1 < p2 ####p1 is before 2000

#####Two-sample proportion test (left-tailed)

```
z_test <- movie_sample %>%
  specify(formula = highglobal ~ yearperiod, success = "1") %>%
  calculate(stat = "z", order = c("1", "2"))

z_test
```

```
## Response: highglobal (factor)
## Explanatory: yearperiod (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.372
```

```
p_value_theoretical_4 <- pnorm(z_test$stat, lower.tail = TRUE)

p_value_theoretical_4
```

```
## [1] 0.6451572
```

The p-value is 0.64516 which is greater than the default significance level of 0.05, therefore we fail to reject the null hypothesis. It can be concluded that there is no evidence that the proportion of movies released from 1980 to 1999 whose global box office earnings exceeds budget is lower than the proportion of movies released from 2000 to 2018 whose global box office earnings exceeds budget.