# Homework 6

2023-11-30

```
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(gapminder)
library(infer)
library(broom)

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, ~
```

For each confidence interval, use all percentile method, standard error method and theoretical method. Get the bootstrap distribution first and use the distribution to justify whether standard error method can be used. Also justify whether theoretical method can be used.

Use the random sample with 200 movies you had from homework 5 to develop:

```
# random sample with 200 movies from homework 5
movie_sample <- boxoffice %>% rep_sample_n(size=200) %>%
  mutate(summer=ifelse(Month %in% c("Jun", "Jul", "Aug"), "1", "0"),
  year_period=ifelse(Year<=1999, "1", "2"),
  highglobal=ifelse(Worldwide_Gross>=Budget, 1, 0))

movie_1 <- movie_sample %>% summarise(xbar=mean(Worldwide_Gross),
```

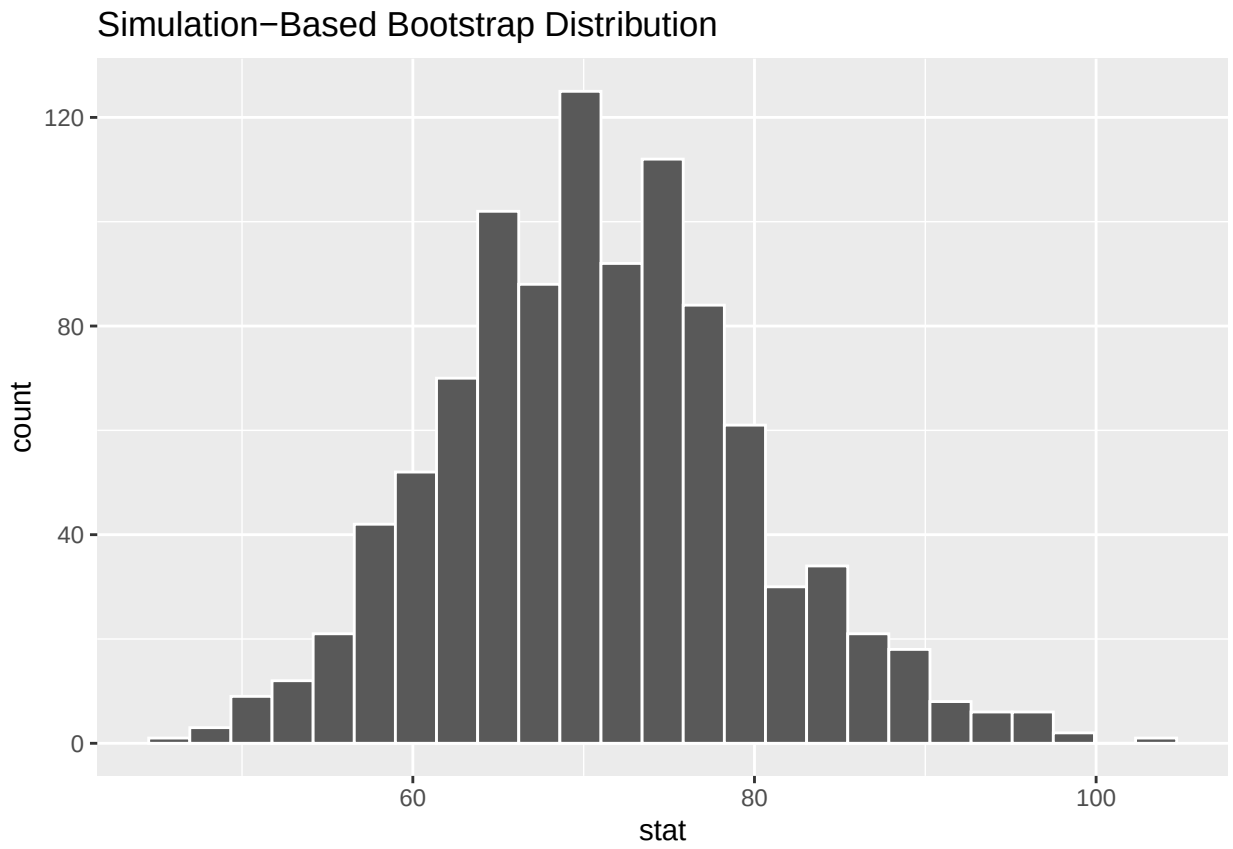
```
  sd=sd(Worldwide_Gross), sumglobal=sum(highglobal), nobs=n()) %>%
  mutate(phat=sumglobal/nobs)
```

1. The confidence interval of the average global box office earning of all movies from 1980 to 2018

```
movie_resamples <- movie_sample %>%
  specify(response=Worldwide_Gross) %>% generate(reps=1000) %>%
  calculate(stat="mean")
```

```
## Setting `type = "bootstrap"` in `generate()`.
```

```
visualize(data=movie_resamples, bins=25)
```




```
ci_percentile <- get_ci(x=movie_resamples, type="percentile")
ci_se         <- get_ci(x=movie_resamples, type="se", point_estimate=movie_1$xbar)
ci_theoretical<- movie_1 %>% mutate(lower_ci=xbar-1.96*sd/sqrt(nobs),
                  upper_ci=xbar+1.96*sd/sqrt(nobs)) %>% select(lower_ci, upper_ci)

ci_percentile
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     54.2     90.1
```

```
ci_se
```

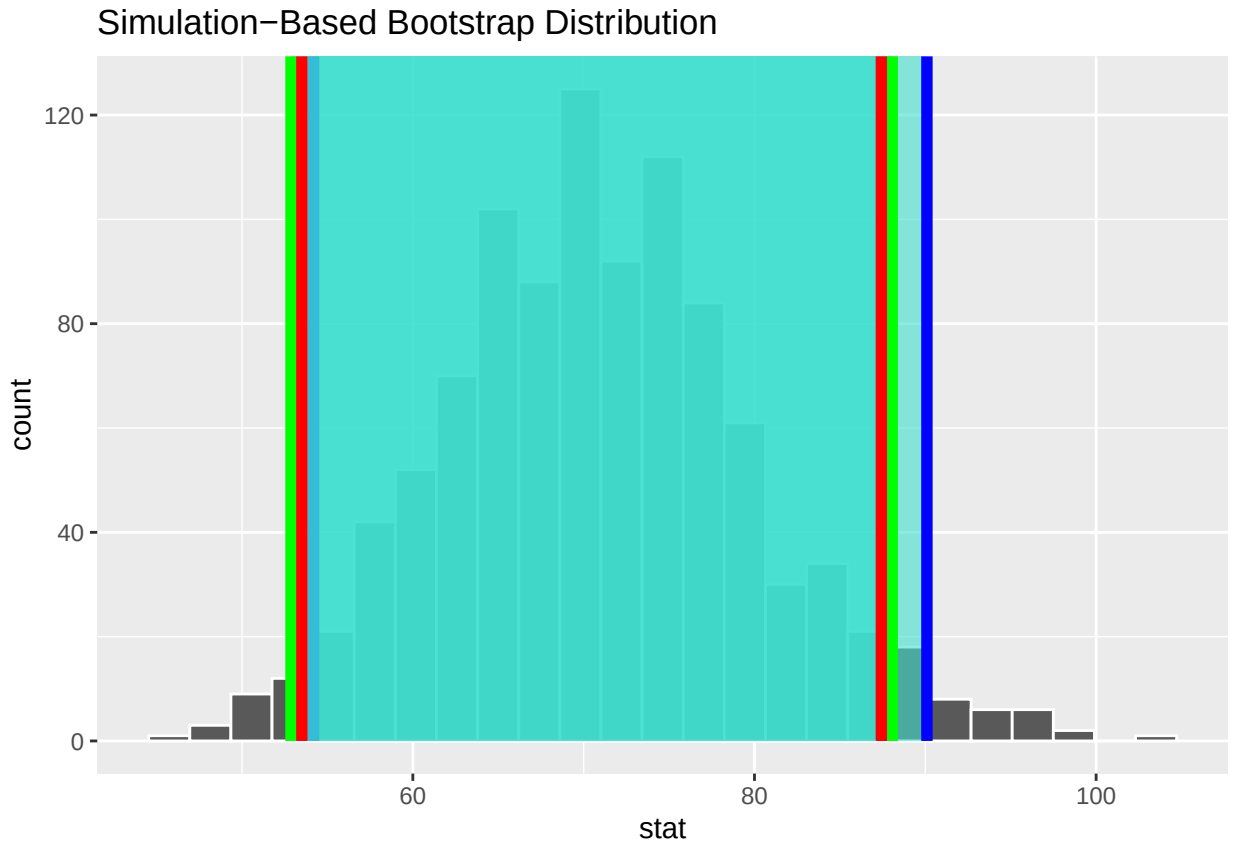
```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
```

```
## 1     52.9     88.1
```

```
ci_theoretical
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     53.5     87.4
```

```
visualize(data=movie_resamples, bins=25) + shade_ci(endpoints=ci_percentile, color="blue") +
  shade_ci(endpoints=ci_se, color="green") + shade_ci(endpoints=ci_theoretical, color="red")
```

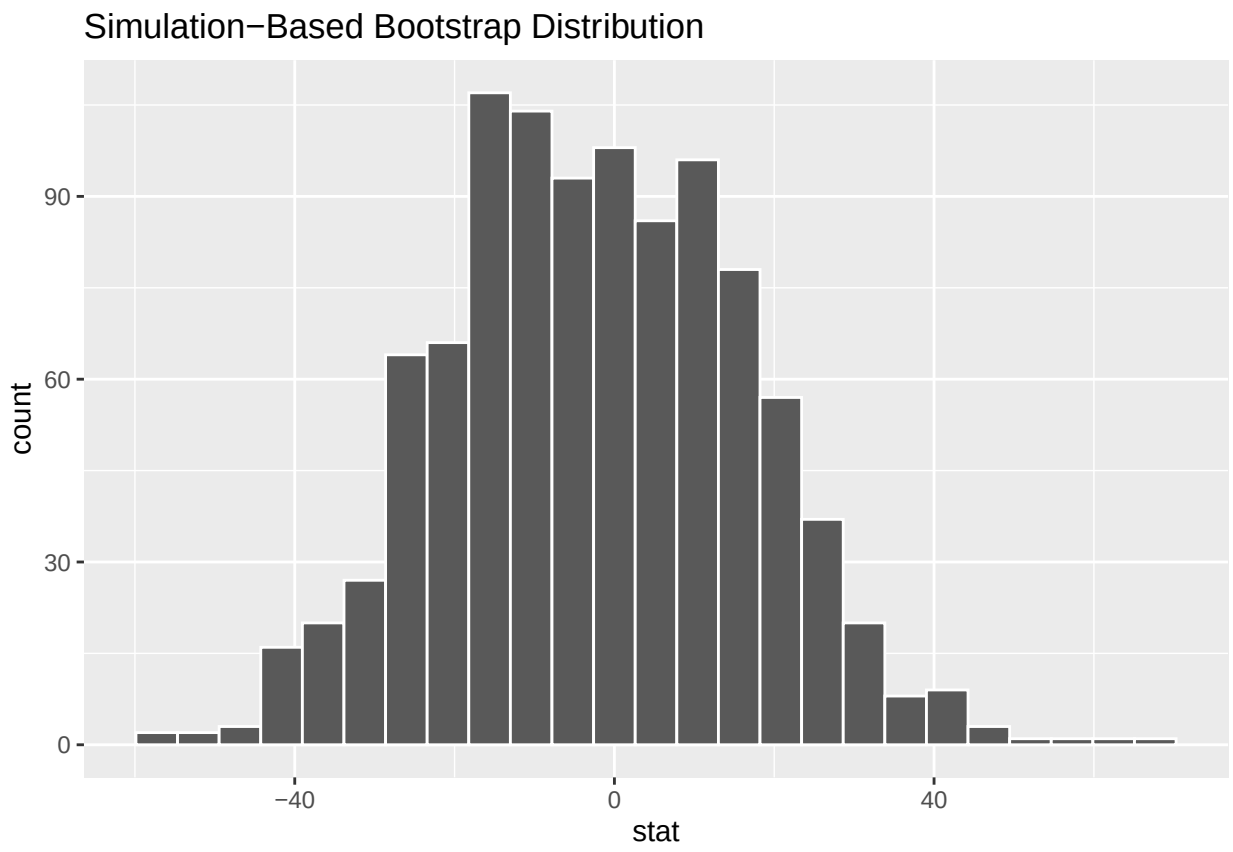



The standard error and theoretical methods can be used since the CI for the average global box office earning and the sampling distribution of sample mean is approximately normal when the sample is large enough. All three methods are used.

2. The confidence interval of the difference of average global box office earnings between movies in the summer (June, July and August) and movies in the rest of the year

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

movie_summer <- movie_sample %>% filter(summer==1) %>%
  summarise(xbar_summer=mean(Worldwide_Gross), sd_summer=sd(Worldwide_Gross),
            n_summer=n()) %>% select(xbar_summer, sd_summer, n_summer)

movie_nonsummer <- movie_sample %>% filter(summer==0) %>%
  summarise(xbar_nonsummer=mean(Worldwide_Gross), sd_nonsummer=sd(Worldwide_Gross),
```

```
            n_nonsummer=n()) %>% select(xbar_nonsummer, sd_nonsummer, n_nonsummer)

calc_summer <- cbind(movie_summer, movie_nonsummer) %>%
  mutate(xbar_diff=xbar_summer - xbar_nonsummer,
         se_diff=sqrt(sd_summer^2/n_summer+sd_nonsummer^2/n_nonsummer))

visualize(data=movie_resamples, bins=25)
```




```
ci_percentile2 <- get_ci(x=movie_resamples, type="percentile")
ci_se2         <- get_ci(x=movie_resamples, type="se",
                         point_estimate=calc_summer$xbar_diff)
ci_theoretical2<- calc_summer %>% mutate(lower_ci=xbar_diff-1.96*se_diff,
                  upper_ci=xbar_diff+1.96*se_diff) %>% select(lower_ci, upper_ci)

ci_percentile2
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -37.9     33.7
```

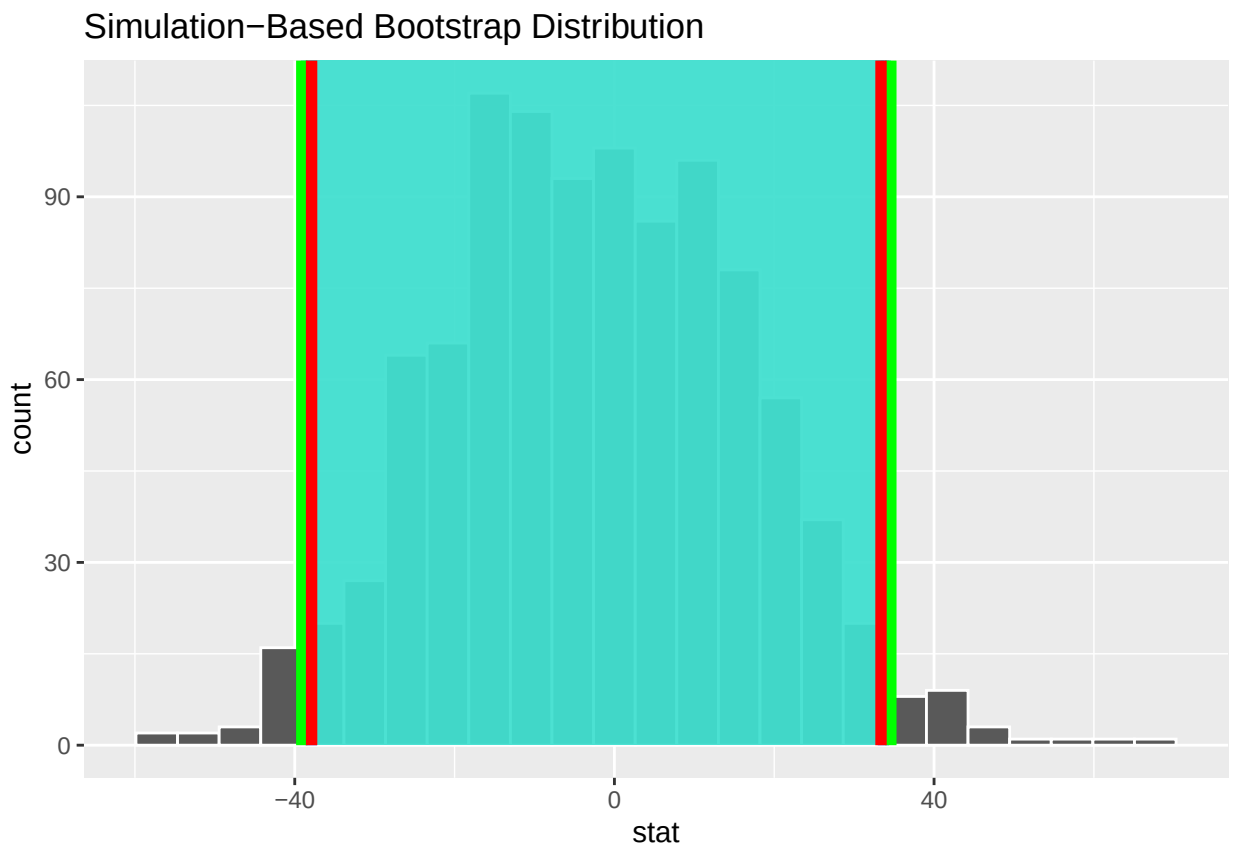
```
ci_se2
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -39.1     34.6
```

```
ci_theoretical2
```

```
##   lower_ci upper_ci
## 1 -37.8899 33.36199
```

```
visualize(data=movie_resamples, bins=25) + shade_ci(endpoints=ci_percentile2,
            color="blue") + shade_ci(endpoints=ci_se2, color="green") +
            shade_ci(endpoints=ci_theoretical2, color="red")
```



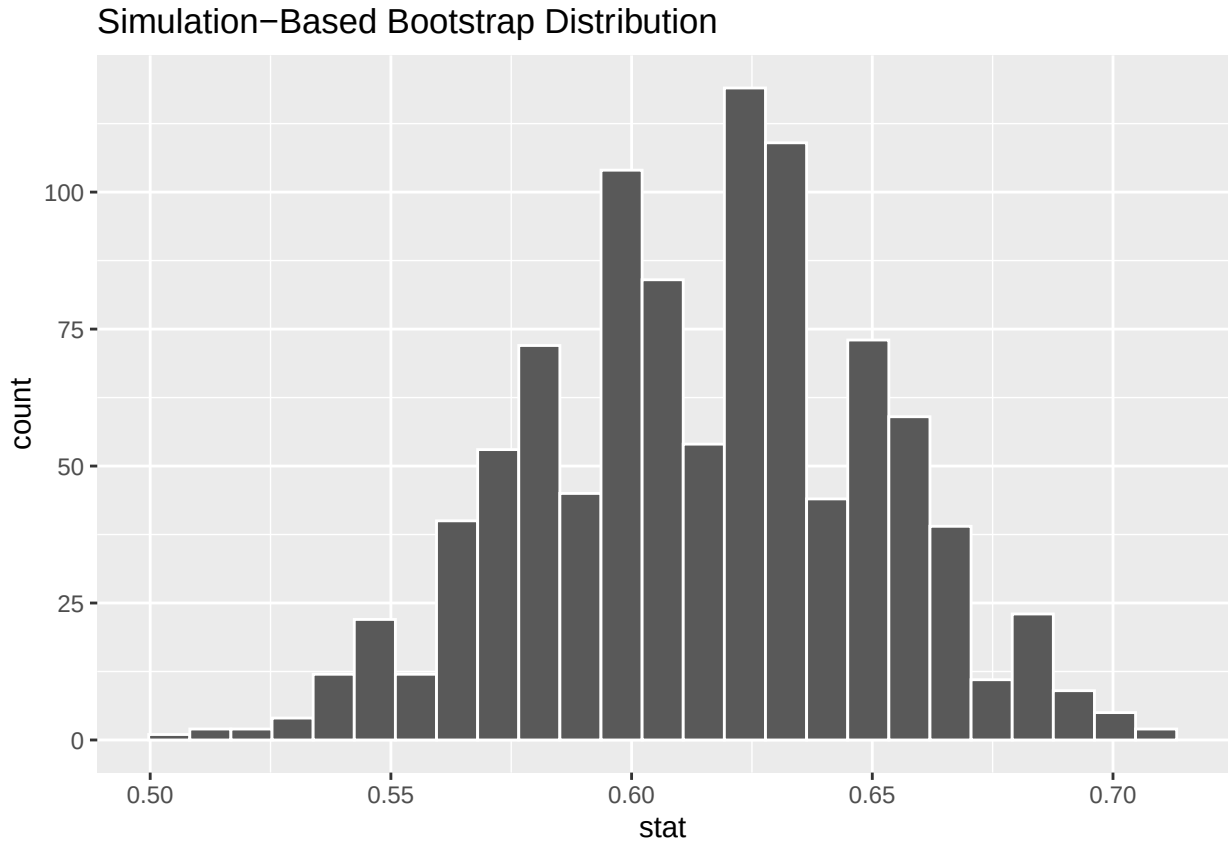

The standard error and theoretical methods can be used since the CI for the difference of average global box office earnings between movies in the summer and the rest is approximately normal when the sample is large enough. All three methods are used.

3. The confidence interval of the proportion of movies whose global box office earning exceeds budget.

```
movie_resamples <- movie_sample %>% mutate(highglobal=as.factor(highglobal)) %>%
  specify(response=highglobal, success="1") %>% generate(reps=1000) %>%
  calculate(stat="prop")
```

```
## Setting `type = "bootstrap"` in `generate()`.
```

```
visualize(data=movie_resamples, bins=25)
```




```
ci_percentile3 <- get_ci(x=movie_resamples, type="percentile")
ci_se3         <- get_ci(x=movie_resamples, type="se", point_estimate=movie_1$phat)
ci_theoretical3<- movie_1 %>% mutate(lower_ci=phat-1.96*sqrt(phat*(1-phat)/nobs),
        upper_ci=phat+1.96*sqrt(phat*(1-phat)/nobs)) %>% select(lower_ci, upper_ci)

ci_percentile3
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.545     0.68
```

```
ci_se3
```

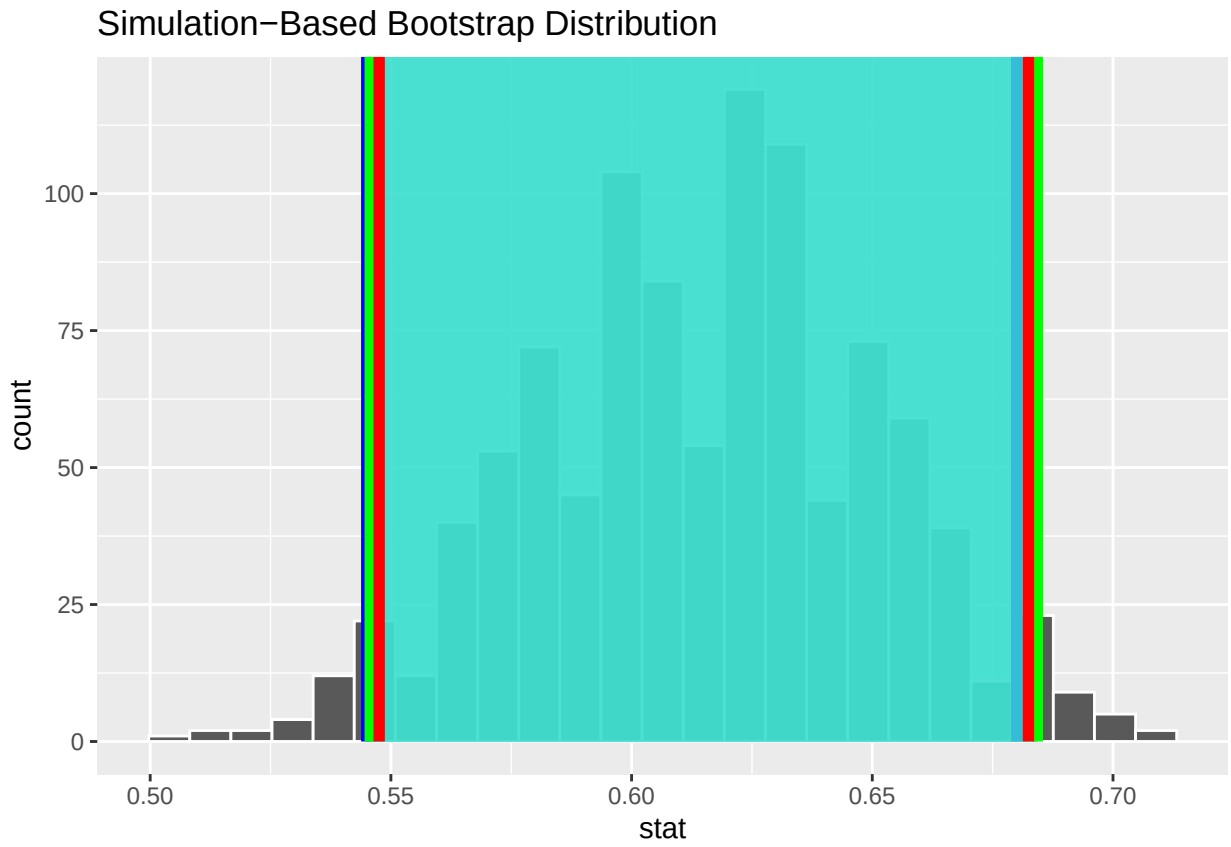
```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.546    0.684
```

```
ci_theoretical3
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.548    0.682
```

```
visualize(data=movie_resamples, bins=25) +
  shade_ci(endpoints=ci_percentile3, color="blue") +
  shade_ci(endpoints=ci_se3, color="green") +
```

```
  shade_ci(endpoints=ci_theoretical3, color="red")
```




The standard error and theoretical methods can be used since the CI for the proportion of movies whose global box office earning exceeds budget is approximately normal.

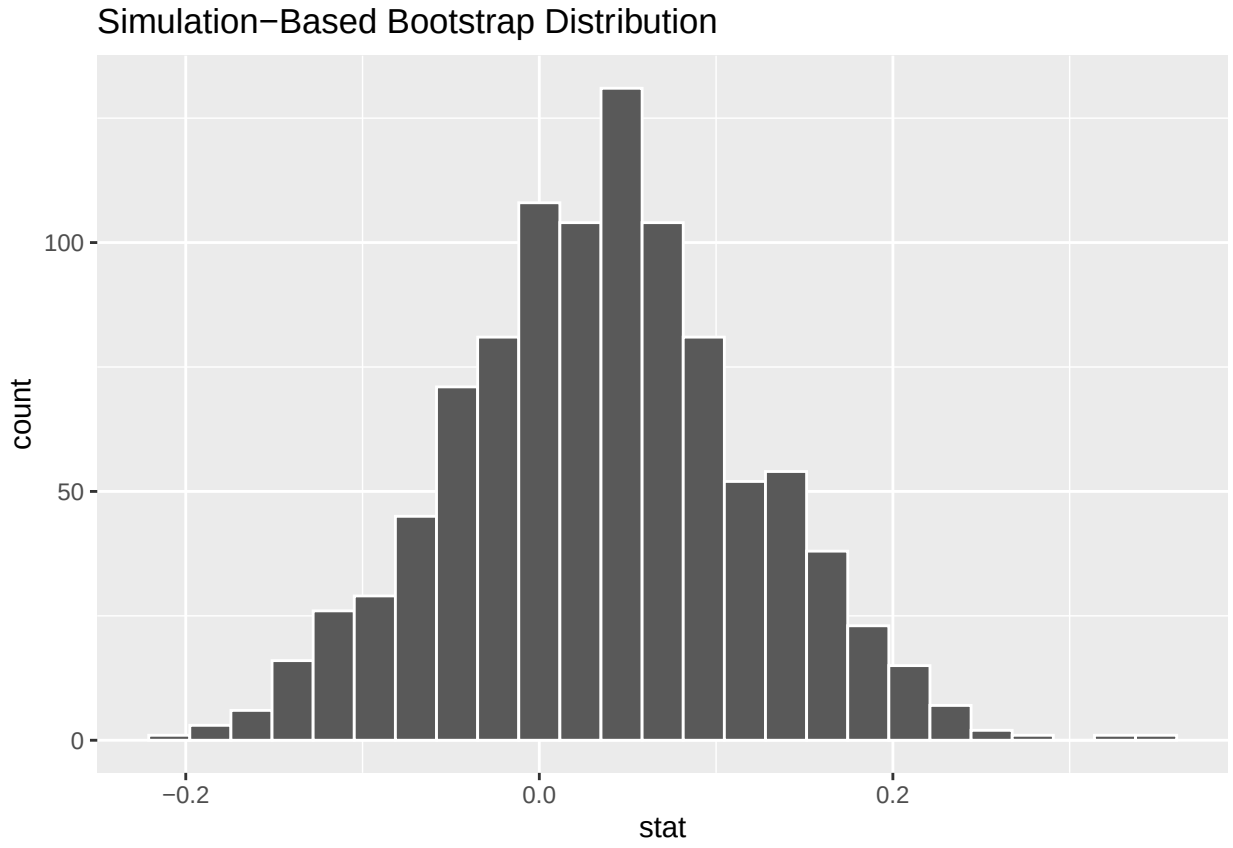
4. The confidence interval of the difference of proportions of movies whose global box office earning exceeds budget between movies released from 1980 to 1999 and those released from 2000 and 2018

```
movie_resamples <- movie_sample %>% mutate(highglobal=as.factor(highglobal)) %>%
  specify(formula = highglobal ~ year_period, success="1") %>%
  generate(reps=1000) %>% calculate(stat="diff in props", order=c(1,2))
```

```
## Setting `type = "bootstrap"` in `generate()`.
```

```
movie_b2000 <- movie_sample %>% filter(year_period==1) %>%
  summarise(sum_b2000=sum(highglobal), n_b2000=n()) %>%
  select(sum_b2000, n_b2000)

movie_a2000 <- movie_sample %>% filter(year_period==2) %>%
  summarise(sum_a2000=sum(highglobal), n_a2000=n()) %>%
  select(sum_a2000, n_a2000)

year <- cbind(movie_b2000, movie_a2000) %>% mutate(p1hat=sum_b2000/n_b2000,
                p2hat=sum_a2000/n_a2000,
                phat_diff=p1hat-p2hat,
                se_diff=sqrt((p1hat*(1-p1hat)/n_b2000)+(p2hat*(1-p2hat)/n_a2000)))

calc_summer <- cbind(movie_summer, movie_nonsummer) %>%
  mutate(xbar_diff=xbar_summer - xbar_nonsummer,
```

```
        se_diff=sqrt(sd_summer^2/n_summer+sd_nonsummer^2/n_nonsummer))

visualize(data=movie_resamples, bins=25)
```




```
ci_percentile4 <- get_ci(x=movie_resamples, type="percentile")
ci_se4         <- get_ci(x=movie_resamples, type="se", point_estimate=year$phat_diff)
ci_theoretical4<- year %>% mutate(lower_ci=phat_diff-1.96*se_diff,
                 upper_ci=phat_diff+1.96*se_diff) %>% select(lower_ci, upper_ci)

ci_percentile4
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.128    0.199
```

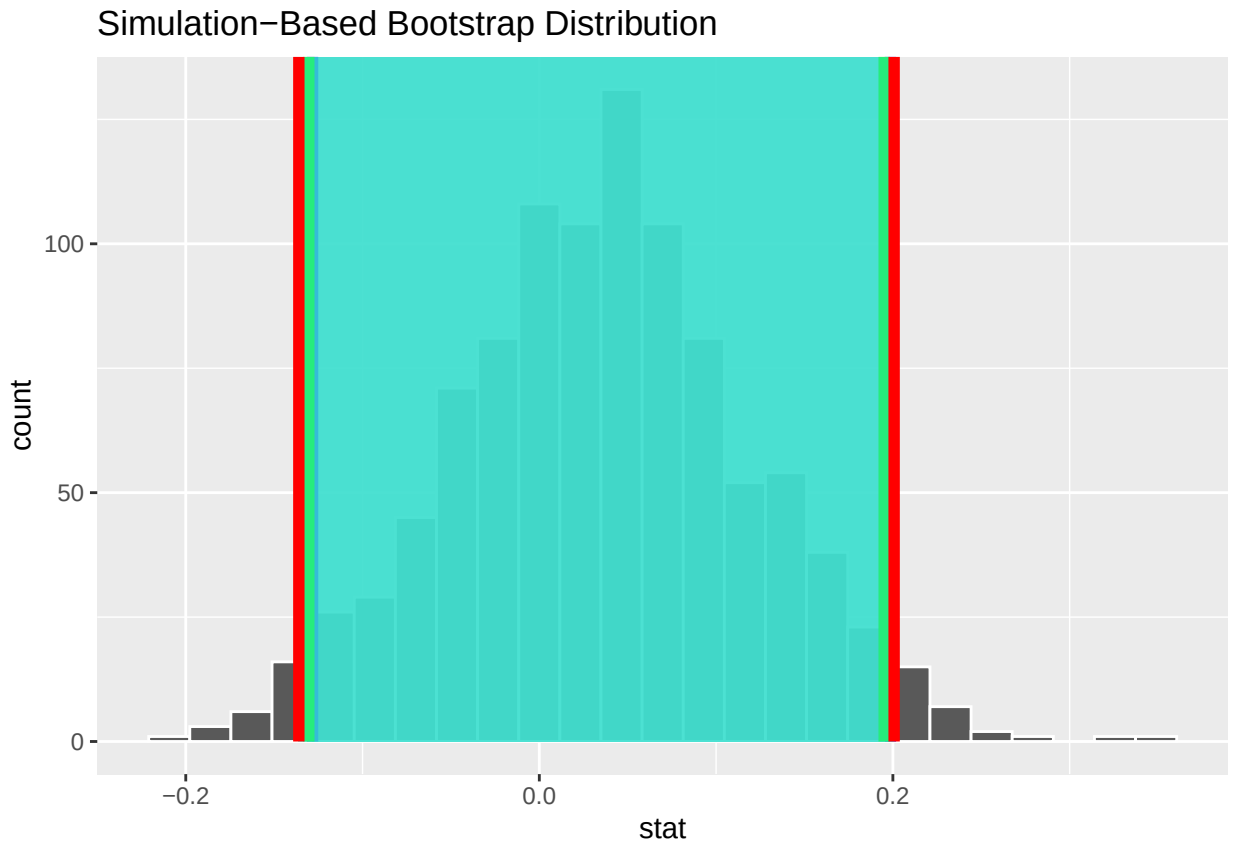
```
ci_se4
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.130    0.195
```

```
ci_theoretical4
```

```
##     lower_ci  upper_ci
## 1 -0.1360443 0.2007043
```

```
visualize(data=movie_resamples, bins=25) + shade_ci(endpoints=ci_percentile4, color="blue") + shade_ci
```




The standard error and theoretical methods can be used since the CI for the difference of proportions of movies whose global box office earning exceeds budget between movies released from 1980 to 1999 and those released from 2000 and 2018 is approximately normal when the sample is large enough. All three methods are used.