HW6 - Akshaj Kammari

Due: 04/18/2024

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(boot)

movies <- read_csv("movie_boxoffice.csv")
```

```
## Rows: 4969 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Movie, Month
## dbl (5): Day, Year, Budget, Domestic_Gross, Worldwide_Gross
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
set.seed(123)
movie_sample <- movies %>% sample_n(200)
```

1. Confidence Interval for Average Global Box Office Earnings

```
bootstrap_mean <- function(data, indices) {
  d <- data[indices, ]  # resample data
  mean(d$Worldwide_Gross)
}

set.seed(123)
boot_mean <- boot(data=movie_sample, statistic=bootstrap_mean, R=1000)

ci_mean_percentile <- quantile(boot_mean$t, probs=c(0.025, 0.975))
ci_mean_se <- boot.ci(boot_mean, type="bca")
```

2. Confidence Interval for Difference in Earnings Between Summer and Other Months

```
bootstrap_diff <- function(data, indices) {
  d <- data[indices, ]
  summer_mean <- mean(d$Worldwide_Gross[d$Month %in% c("Jun", "Jul", "Aug")])
  other_mean <- mean(d$Worldwide_Gross[!d$Month %in% c("Jun", "Jul", "Aug")])
```

```
  summer_mean - other_mean
}

boot_diff <- boot(data=movie_sample, statistic=bootstrap_diff, R=1000)

ci_diff_percentile <- quantile(boot_diff$t, probs=c(0.025, 0.975))
ci_diff_se <- boot.ci(boot_diff, type="bca")
```

3. Confidence Interval for the Proportion of Movies Exceeding Budget

```
bootstrap_prop <- function(data, indices) {
  d <- data[indices, ]
  mean(d$Worldwide_Gross > d$Budget)
}

boot_prop <- boot(data=movie_sample, statistic=bootstrap_prop, R=1000)

ci_prop_percentile <- quantile(boot_prop$t, probs=c(0.025, 0.975))
ci_prop_se <- boot.ci(boot_prop, type="bca")

print("Confidence Interval for Mean Earnings (Percentile):")
```

```
## [1] "Confidence Interval for Mean Earnings (Percentile):"
```

```
print(ci_mean_percentile)
```

```
##     2.5%    97.5%
##  68.6387 117.0485
```

```
print("Confidence Interval for Mean Earnings (SE):")
```

```
## [1] "Confidence Interval for Mean Earnings (SE):"
```

```
print(ci_mean_se)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_mean, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 70.93, 120.83 )
## Calculations and Intervals on Original Scale
```

```
print("Confidence Interval for Earnings Difference (Percentile):")
```

```
## [1] "Confidence Interval for Earnings Difference (Percentile):"
```

```
print(ci_diff_percentile)
```

```
##      2.5%     97.5%
##  16.87101 141.61507
```

```
print("Confidence Interval for Earnings Difference (SE):")
```

```
## [1] "Confidence Interval for Earnings Difference (SE):"
```

```r
print(ci_diff_se)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_diff, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 22.50, 151.85 )
## Calculations and Intervals on Original Scale
```

```r
print("Confidence Interval for Proportion Exceeding Budget (Percentile):")
```

```
## [1] "Confidence Interval for Proportion Exceeding Budget (Percentile):"
```

```r
print(ci_prop_percentile)
```

```
##  2.5% 97.5%
##  0.56  0.69
```

```r
print("Confidence Interval for Proportion Exceeding Budget (SE):")
```

```
## [1] "Confidence Interval for Proportion Exceeding Budget (SE):"
```

```r
print(ci_prop_se)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_prop, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.555,  0.685 )
## Calculations and Intervals on Original Scale
```

4. Confidence Interval for Difference in Proportion of Movies Exceeding budget between 1980 to 1999 and 2000 to 2018

```r
movie_sample <- movie_sample %>%
  mutate(period = ifelse(Year >= 1980 & Year <= 1999, "1980-1999", "2000-2018"))

bootstrap_diff_prop <- function(data, indices) {
  sample_data <- data[indices, ]
  prop_early <- mean(sample_data$Worldwide_Gross[sample_data$period == "1980-1999"] > sample_data$Budge
  prop_late <- mean(sample_data$Worldwide_Gross[sample_data$period == "2000-2018"] > sample_data$Budget
  prop_late - prop_early
}

set.seed(123)
boot_diff_prop <- boot(data=movie_sample, statistic=bootstrap_diff_prop, R=1000)

ci_diff_prop_percentile <- quantile(boot_diff_prop$t, probs=c(0.025, 0.975))
ci_diff_prop_se <- boot.ci(boot_diff_prop, type="bca")
```

```r
print("Confidence Interval for Difference of Proportions (Percentile):")
```

```
## [1] "Confidence Interval for Difference of Proportions (Percentile):"
```

```r
print(ci_diff_prop_percentile)
```

```
##      2.5%     97.5%
## -0.194249  0.131250
```

```r
print("Confidence Interval for Difference of Proportions (SE):")
```

```
## [1] "Confidence Interval for Difference of Proportions (SE):"
```

```r
print(ci_diff_prop_se)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_diff_prop, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   (-0.2042,  0.1189 )
## Calculations and Intervals on Original Scale
```

##The sample distribution of the statistic is assumed to be roughly normal using the standard error method The central limit theorem indicates that, with a sufficient number of resamples, the distribution of the sample means will gravitate towards normalcy since we are using bootstrapping, which creates thousands of resamples from the original data. As a result, the usual error approach is appropriate.

##The theoretical method says that the sampling distribution of the statistic approaches a known distribution (such the normal distribution) as the sample size approaches infinity. As a result of bootstrapping, we are able to simulate an endlessly huge sample size in our scenario, even though we are not working with enormous sample sizes in the conventional sense. As a result, the theoretical approach is also appropriate.