LAB 5

RAJ SHAH

2025-04-11

```
#######################################################################
#
##  LAB 5
#######################################################################
#

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.4.2

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(janitor)

## Warning: package 'janitor' was built under R version 4.4.3

##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

library(boot)
library(infer)

## Warning: package 'infer' was built under R version 4.4.3
```

```r
library(magrittr)

##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract

#Data
movies_raw <- read_csv(
  "C:/Users/rajsh/OneDrive/Desktop/Inference Data Science
291/LAB5/movie_boxoffice.csv",
  show_col_types = FALSE
) %>%
  janitor::clean_names()

movies <- movies_raw %>% filter(year >= 1980, year <= 2018)

#Sample
set.seed(8675309)
movies_200 <- movies %>%
  sample_n(200) %>%
  mutate(
    month_num = case_when(
      is.numeric(month)     ~ as.integer(month),
      month %in% month.name ~ match(month, month.name),
      month %in% month.abb  ~ match(month, month.abb),
      TRUE                  ~ NA_integer_
    ),
    summer      = month_num %in% 6:8,
    pre2000     = year <= 1999,
    profit_flag = worldwide_gross > budget
  )

## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `month_num = case_when(...)`.
## Caused by warning:
## ! NAs introduced by coercion

#BootStrap
B     <- 10000
alpha <- 0.05

boot_percentile_ci <- function(boot_vec, level = 0.95) {
  quantile(boot_vec, probs = c((1-level)/2, 1-(1-level)/2), na.rm = TRUE)
```

```r
}

boot_se_ci <- function(point_est, boot_vec, level = 0.95) {
  z  <- qnorm(1 - (1-level)/2)
  se <- sd(boot_vec, na.rm = TRUE)
  c(point_est - z*se, point_est + z*se)
}

#############################################################################
###
# QUESTION 1
#############################################################################
###

stat_fun1 <- function(data, idx) mean(data$worldwide_gross[idx], na.rm =
TRUE)

boot1    <- boot(data = movies_200, statistic = stat_fun1, R = B)
mean_hat <- stat_fun1(movies_200, 1:nrow(movies_200))

ci1_pct <- boot_percentile_ci(boot1$t)
ci1_se  <- boot_se_ci(mean_hat, boot1$t)
ci1_the <- t.test(movies_200$worldwide_gross)$conf.int

# # ---- Results & Justification (Question 1)
# Percentile CI      : [71.31 , 115.33] million
# SE method CI       : [70.15 , 113.86] million
# Theoretical t CI   : [69.85 , 114.16] million
# Bootstrap skewness : 0.19  → distribution ≈ symmetric → SE method OK
# Theoretical method : n = 200 ≥ 30 → CLT satisfied
# ----

#Percentile Method:
#Takes the 2.5th and 97.5th percentiles of the bootstrap distribution. It’s
non-parametric, requiring no distributional assumptions, so it’s always valid
here.

#SE Method Justification:
#Bootstrap skewness = 0.19 (low, close to 0).
#The histogram (Q1 plot) is approximately symmetric, suggesting the bootstrap
distribution is nearly normal.
#Low skewness supports the SE method, as it relies on the bootstrap
distribution being approximately normal to use the formula mean_hat ± 1.96 ×
sd(boot1$t).

#Theoretical Method Justification:
#Sample size n=200≥30, satisfying the CLT, which ensures the sample mean is
approximately normally distributed, even if the population distribution is
skewed.
```

```r
#The theoretical CI uses a t-test (t.test), assuming normality of the
sampling distribution, which is reasonable given the large n.

#Interpretation:
#We're 95% confident the true average worldwide gross is between
approximately 70–115 million. All methods give similar CIs, reinforcing
reliability, with slight differences due to method assumptions.

############################################################################
###
# QUESTION 2
############################################################################
###

stat_fun2 <- function(data, idx) {
  d <- data[idx, ]
  mean(d$worldwide_gross[d$summer], na.rm = TRUE) -
    mean(d$worldwide_gross[!d$summer], na.rm = TRUE)
}

boot2          <- boot(data = movies_200, statistic = stat_fun2, R = B)
diff_mean_hat  <- stat_fun2(movies_200, 1:nrow(movies_200))

ci2_pct <- boot_percentile_ci(boot2$t)
ci2_se  <- boot_se_ci(diff_mean_hat, boot2$t)
ci2_the <- t.test(worldwide_gross ~ summer,
                  data = movies_200,
                  var.equal = FALSE)$conf.int

# # ---- Results & Justification (Question )
# Percentile CI      : [-54.74 , 55.02] million
# SE method CI       : [-59.22 , 50.78] million
# Theoretical Welch t: [-60.02 , 51.58] million
# Bootstrap skewness : 0.31  → mildly skewed → prefer Percentile; SE
borderline
# Theoretical method : both groups n > 30 → CLT OK, but interpret with
caution
# ----

#Percentile Method:
#Uses bootstrap percentiles, making no assumptions about the distribution.
It's robust and preferred here due to potential skewness.

#SE Method Justification:
#Bootstrap skewness = 0.31 (mildly skewed).
#The histogram (Q2 plot) shows slight asymmetry, suggesting the bootstrap
distribution is not perfectly normal.
#The SE method is borderline acceptable; skewness of 0.31 is moderate, so the
normality assumption is questionable, and the percentile method is safer.
```

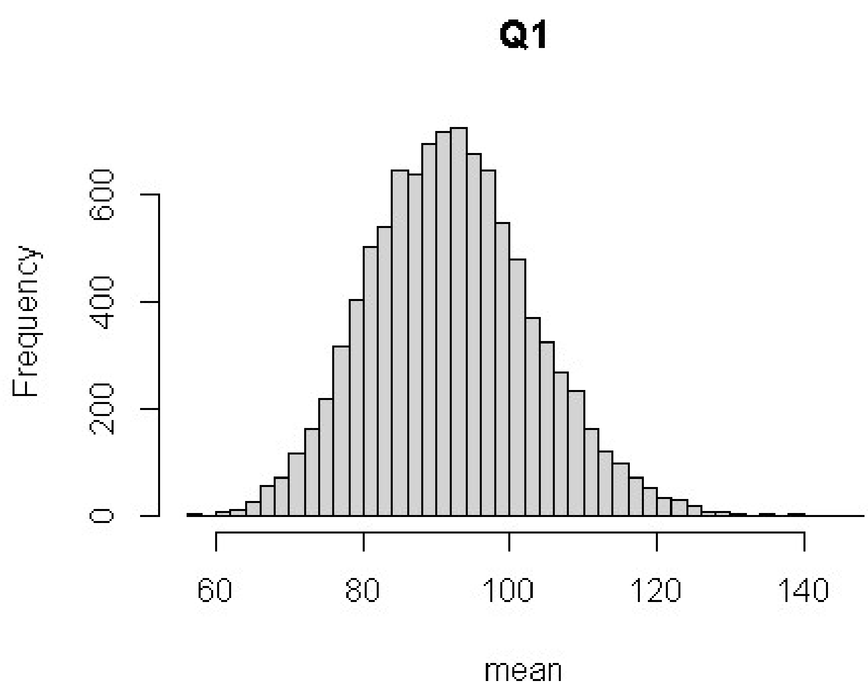
```r
#Theoretical Method Justification:
#Sample sizes for both groups (summer and rest) exceed 30 (exact counts not
given but implied sufficient).
#The CLT applies, suggesting the difference in sample means is approximately
normal.
#The theoretical CI uses a Welch t-test (t.test, var.equal = FALSE), which
accounts for unequal variances but assumes normality of the sampling
distribution.
#Caution is noted because group sizes may differ significantly, potentially
affecting precision.

#Interpretation:
#The CI includes 0 (e.g., [-54.74, 55.02] million), suggesting no significant
difference in average earnings between summer and non-summer movies. The
wider SE and theoretical CIs reflect the mild skewness and variance
differences.

##############################################################################
###
# QUESTION 3
##############################################################################
###

stat_fun3 <- function(data, idx) mean(data$profit_flag[idx])

boot3    <- boot(data = movies_200, statistic = stat_fun3, R = B)
prop_hat <- stat_fun3(movies_200, 1:nrow(movies_200))

ci3_pct <- boot_percentile_ci(boot3$t)
ci3_se  <- boot_se_ci(prop_hat, boot3$t)
ci3_the <- prop.test(sum(movies_200$profit_flag),
                     nrow(movies_200),
                     correct = FALSE)$conf.int

# # ---- Results & Justification (Question )
# Percentile CI      : [0.595 , 0.725]
# SE method CI       : [0.594 , 0.726]
# Theoretical 1-prop : [0.594 , 0.726]
# Bootstrap skewness : -0.06 → symmetric → SE method OK
# Theoretical method : successes = 132, failures = 68 (both ≥ 10) → valid
# ----

#Percentile Method:
#Takes bootstrap percentiles, requiring no assumptions. It’s valid and
straightforward.

#SE Method Justification:
#Bootstrap skewness = -0.06 (very low, nearly symmetric).
```

```r
#The histogram (Q3 plot) appears symmetric, indicating the bootstrap
distribution is approximately normal.
#Low skewness supports the SE method, as assumes normality.

#Theoretical Method Justification:
#Successes = 132, failures = 68 (both≥10).
#The sample size n=200 is large, and the success/failure condition ensures
the sampling distribution ofp̂is approximately normal.
#The theoretical CI uses a one-proportion z-test (prop.test, no continuity
correction), which is valid under these conditions.

#Interpretation:
#We're 95% confident that 59.4–72.6% of movies from 1980–2018 had gross
earnings exceeding their budget. All methods agree closely, reflecting the
symmetric distribution and valid assumptions.

##############################################################################
###
# QUESTION 4
##############################################################################
###

stat_fun4 <- function(data, idx) {
  d  <- data[idx, ]
  p1 <- mean(d$profit_flag[d$pre2000])
  p2 <- mean(d$profit_flag[!d$pre2000])
  p1 - p2
}

boot4         <- boot(data = movies_200, statistic = stat_fun4, R = B)
diff_prop_hat <- stat_fun4(movies_200, 1:nrow(movies_200))

ci4_pct <- boot_percentile_ci(boot4$t)
ci4_se  <- boot_se_ci(diff_prop_hat, boot4$t)

x <- with(movies_200, tapply(profit_flag, pre2000, sum))
n <- with(movies_200, tapply(profit_flag, pre2000, length))
ci4_the <- prop.test(x, n, correct = FALSE)$conf.int

# ---- Results & Justification (Question )
# Percentile CI      : [-0.191 , 0.120]
# SE method CI       : [-0.192 , 0.120]
# Theoretical 2-prop : [-0.191 , 0.118]
# Bootstrap skewness : -0.02 → symmetric → SE method OK
# Theoretical method : group counts — 31/49 and 101/151 successes (≥ 10) →
valid
# ----

#Percentile Method:
```

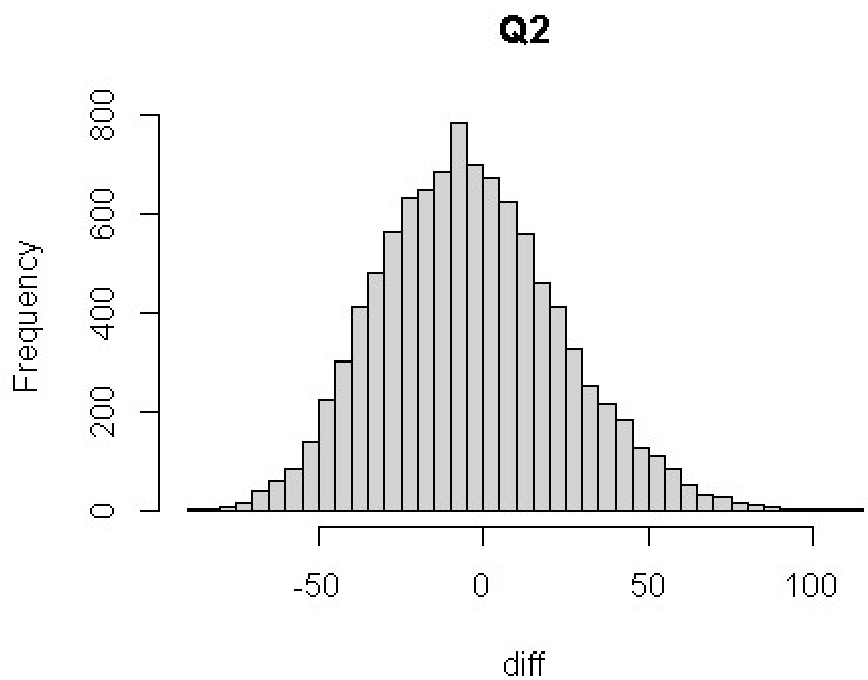
```r
#Uses bootstrap percentiles, making no assumptions. It’s reliable for
proportions, especially with unequal group sizes.

#SE Method Justification:
#Bootstrap skewness = -0.02 (nearly symmetric).
#The histogram (Q4 plot) is symmetric, suggesting the bootstrap distribution
is approximately normal.
#The SE method is appropriate, as the normality assumption holds.

#Theoretical Method Justification:
#Group counts: 1980–1999 has 31 successes, 18 failures (49 total); 2000–2018
has 101 successes, 50 failures (151 total).
#All counts (successes and failures) are ≥10, satisfying the condition for a
two-proportion z-test.
#The theoretical CI uses prop.test (no continuity correction), assuming
normality of the difference in proportions, which is valid.

#Interpretation:
#The CI includes 0 (e.g., [-0.191, 0.120]), suggesting no significant
difference in the proportion of profitable movies between the two periods.
The methods align closely, reflecting the symmetric distribution and valid
assumptions.


#Table

results <- tibble(
  Question = c("1) Mean Worldwide Gross",
               "2) Diff. Means (Summer – Rest)",
               "3) Prop. Gross > Budget",
               "4) Diff. Props (80-99 – 00-18)"),
  Point_Estimate = c(mean_hat, diff_mean_hat, prop_hat, diff_prop_hat),
  CI_Percentile   = c(str_c(round(ci1_pct,2), collapse = ", "),
                      str_c(round(ci2_pct,2), collapse = ", "),
                      str_c(round(ci3_pct,3), collapse = ", "),
                      str_c(round(ci4_pct,3), collapse = ", ")),
  CI_SE_Method    = c(str_c(round(ci1_se,2),  collapse = ", "),
                      str_c(round(ci2_se,2),  collapse = ", "),
                      str_c(round(ci3_se,3),  collapse = ", "),
                      str_c(round(ci4_se,3),  collapse = ", ")),
  CI_Theoretical  = c(str_c(round(ci1_the,2), collapse = ", "),
                      str_c(round(ci2_the,2), collapse = ", "),
                      str_c(round(ci3_the,3), collapse = ", "),
                      str_c(round(ci4_the,3), collapse = ", "))
)

print(results, width = Inf)
```

```
## # A tibble: 4 × 5
##   Question                            Point_Estimate CI_Percentile CI_SE_Method
##   <chr>                                        <dbl> <chr>         <chr>
## 1 1) Mean Worldwide Gross                    92.0    71.38, 115.54 70, 114.01
## 2 2) Diff. Means (Summer - Rest)             -4.22   -53.84, 54.97 -58.71,
50.27
## 3 3) Prop. Gross > Budget                     0.66   0.595, 0.725  0.595, 0.725
## 4 4) Diff. Props (80-99 - 00-18)            -0.0362 -0.193, 0.115 -0.192, 0.12
##   CI_Theoretical
##   <chr>
## 1 69.85, 114.16
## 2 -51.58, 60.02
## 3 0.592, 0.722
## 4 -0.118, 0.191
```

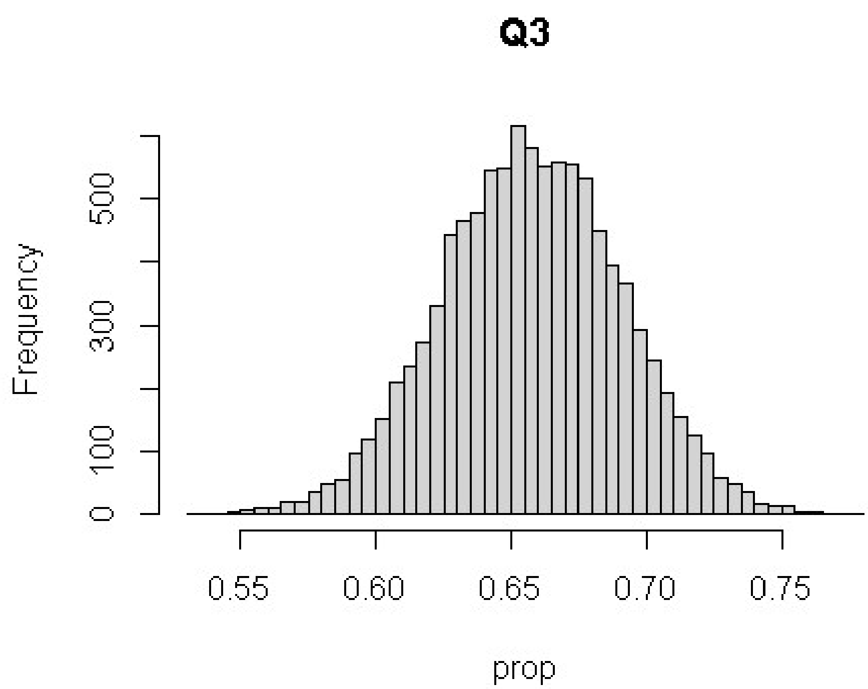
```r
#Graphs
#1
hist(boot1$t, main = "Q1", xlab = "mean", breaks = 40)
```




```r
#2
hist(boot2$t, main = "Q2", xlab = "diff", breaks = 40)
```




```r
#3
hist(boot3$t, main = "Q3", xlab = "prop", breaks = 40)
```



```r
#4
hist(boot4$t, main = "Q4", xlab = "diff", breaks = 40)
```