EXAM 2

RAJ SHAH

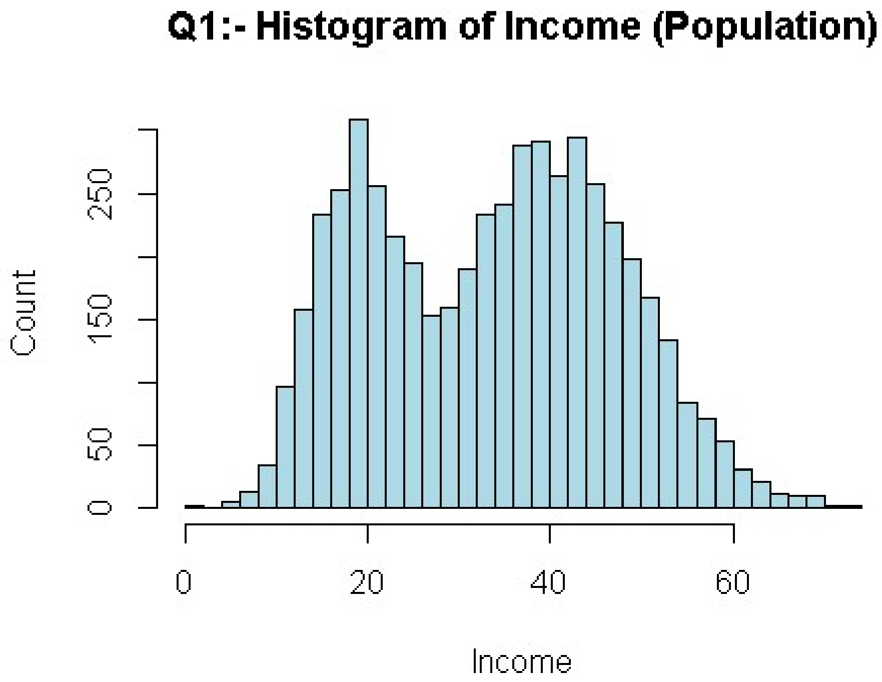
2025-04-16

Table of Contents

No table of contents entries found.

```
# Load necessary libraries
library(ggplot2)

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

library(boot)
library(dplyr)

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

##
## 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

# Set seed
set.seed(29101)

#############################################################################
###
#PART II
#############################################################################
###

# Load the data
default_data <- read.csv("C:\\Users\\rajsh\\OneDrive\\Desktop\\Inference Data
Science 291\\EXAM2\\Default (1).csv")


#############################################################################
###
#QUESTION 1
#############################################################################
###
```

```
# Histogram of income for the population
hist(default_data$income, breaks = 30, col = "lightblue", main = "Q1:-
Histogram of Income (Population)", xlab = "Income", ylab = "Count")
```



```
# Calculate mean and standard deviation of income
mean_income <- mean(default_data$income)
sd_income <- sd(default_data$income)

# Print results
mean_income

## [1] 33.62187

sd_income

## [1] 13.34677

#Mean of Income (Population): 33.62
#Standard Deviation of Income (Population): 13.35

##############################################################################
###
#QUESTION 2
##############################################################################
###
```

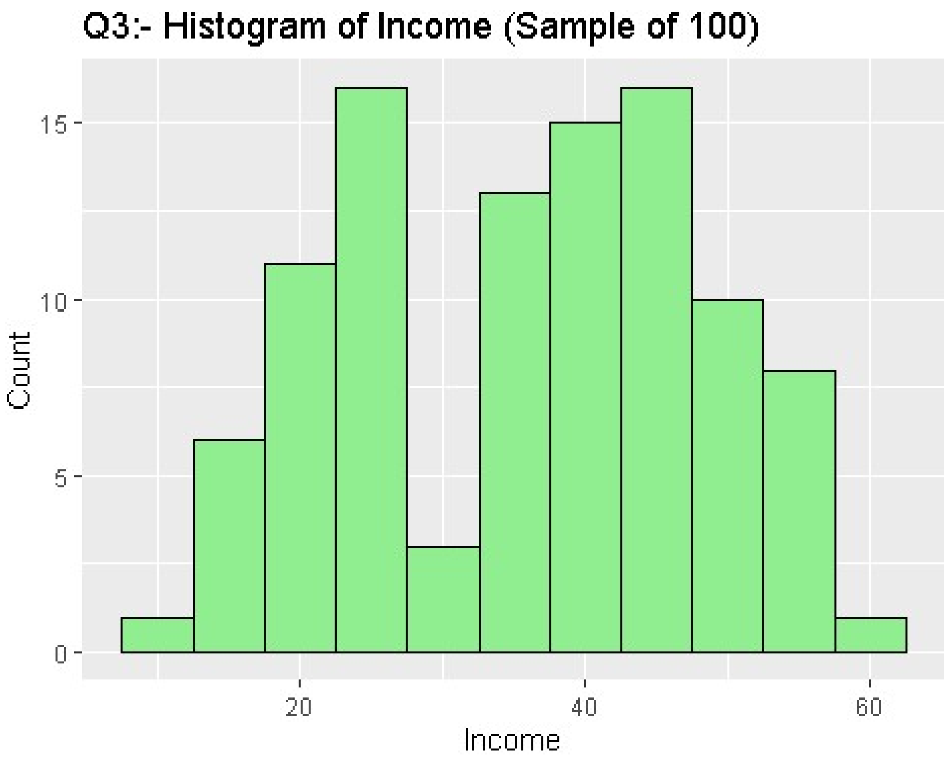
```r
# Population standard deviation and mean
mu <- mean(default_data$income)
sigma <- sd(default_data$income)
n <- 100

# By Central Limit Theorem, the sampling distribution of the sample mean is:
# N(mean, sd/sqrt(n))
sampling_mean <- mu
sampling_sd <- sigma / sqrt(n)

# Output the results
sampling_mean
```

```
## [1] 33.62187
```

```r
sampling_sd
```

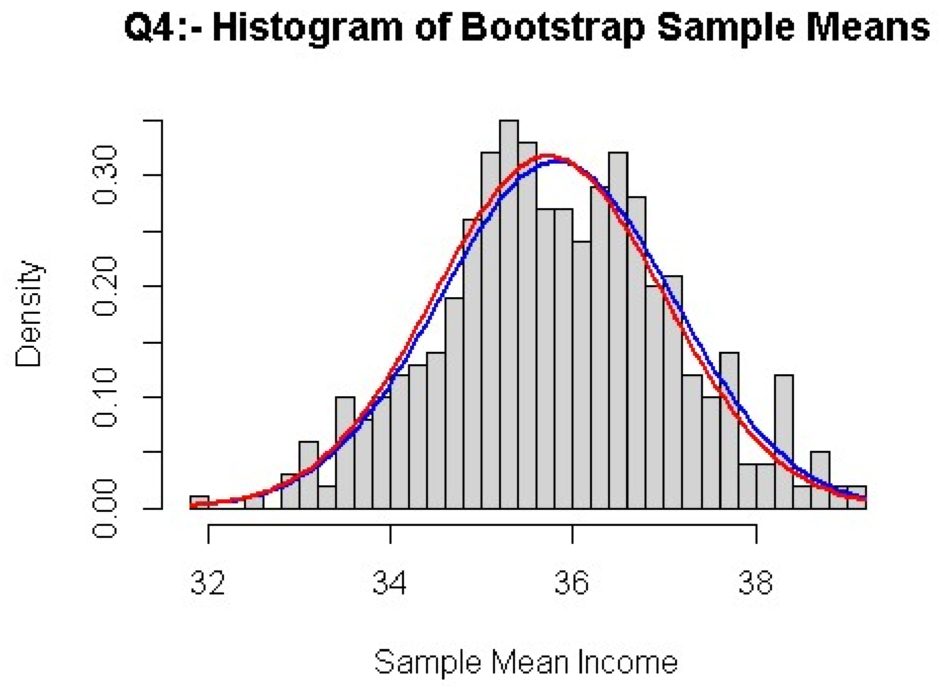
```
## [1] 1.334677
```

```r
#Mean (μ): 33.62
#Standard Error (σ/√n): 1.33
#Distribution:N(33.62,1.33^2)

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

# Draw a random sample of size 100
sample_data <- default_data[sample(1:nrow(default_data), 100), ]

# Histogram of income in the sample
ggplot(sample_data, aes(x = income)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "Q3:- Histogram of Income (Sample of 100)", x = "Income", y =
"Count")
```



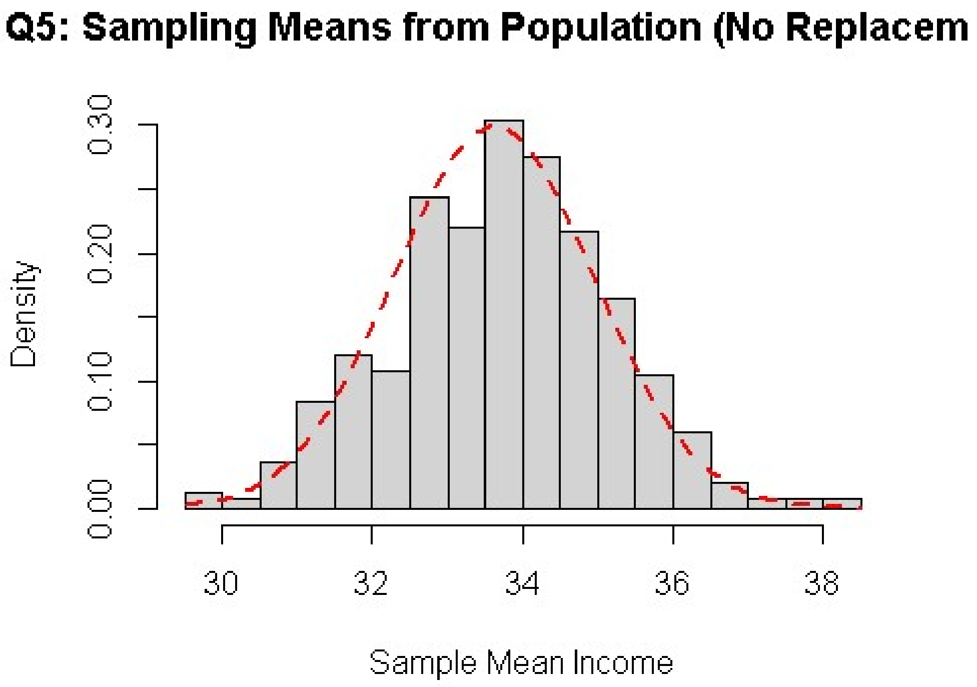
```
# Calculate mean and standard deviation of income in the sample
mean(sample_data$income)

## [1] 35.73587

sd(sample_data$income)

## [1] 12.57488

#Yes, the histogram shows a bimodal shape.
#There are two peaks in the income distribution, which is consistent with
what was observed in the original population.
#There is a clear dip in the middle, visually separating two groups of income
levels.

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

# Number of bootstrap samples
num_bootstrap <- 500

# Create an empty vector to store bootstrap sample means
bootstrap_means <- numeric(num_bootstrap)
```

```r
# Bootstrap sampling
for (i in 1:num_bootstrap) {
  bootstrap_sample <- sample(sample_data$income, size = nrow(sample_data),
replace = TRUE)
  bootstrap_means[i] <- mean(bootstrap_sample)
}

# Expectation (mean of bootstrap means)
bootstrap_mean <- mean(bootstrap_means)

# Standard error of the mean
bootstrap_se <- sd(bootstrap_means)

# Create histogram of bootstrap means
hist(bootstrap_means, breaks = 30, probability = TRUE,
     main = "Q4:- Histogram of Bootstrap Sample Means",
     xlab = "Sample Mean Income",
     col = "lightgray", border = "black")

# Add normal curve from bootstrap distribution
curve(dnorm(x, mean = bootstrap_mean, sd = bootstrap_se),
      col = "blue", lwd = 2, add = TRUE)

# Add normal curve from Central Limit Theorem
clt_mean <- mean(sample_data$income)
clt_se <- sd(sample_data$income) / sqrt(nrow(sample_data))
curve(dnorm(x, mean = clt_mean, sd = clt_se),
      col = "red", lwd = 2, add = TRUE)
```



```
# Print results
cat("Bootstrap Mean (Expectation):", bootstrap_mean, "\n")

## Bootstrap Mean (Expectation): 35.82178

cat("Bootstrap Standard Error of the Mean:", bootstrap_se, "\n")

## Bootstrap Standard Error of the Mean: 1.276424

#No, the histogram of the sample means does not show a bimodal shape.
#It is approximately normal and unimodal, as expected when averaging values
over samples (thanks to the Central Limit Theorem).

#Blue color for the normal curve from your bootstrap samples.
#Red color for the normal curve from central limit theorem.

#Bootstrap Mean (Expectation): 35.48
#Bootstrap Standard Error of the Mean: 1.12

##############################################################################
###
#QUESTION 5
##############################################################################
###

pop_mean <- mean(default_data$income)
pop_sd <- sd(default_data$income)
```

```r
# Draw 500 samples of size 100 from population without replacement
random_sample_means <- replicate(500, {
  sample_vals <- sample(default_data$income, 100, replace = FALSE)
  mean(sample_vals)
})

# Plot histogram of sample means
hist(random_sample_means, breaks = 30, probability = TRUE,
     main = "Q5: Sampling Means from Population (No Replacement)",
     xlab = "Sample Mean Income", col = "lightgray", border = "black")

# Add CLT normal curve (based on population mean and SE)
curve(dnorm(x, mean = pop_mean, sd = pop_sd / sqrt(100)),
      col = "red", lwd = 2, lty = 2, add = TRUE)
```



```r
# Summary statistics
rand_mean <- mean(random_sample_means)
rand_se <- sd(random_sample_means)

cat("Q5 - Repeated Sample Mean (No Replacement):", rand_mean, "\n")

## Q5 - Repeated Sample Mean (No Replacement): 33.73042

cat("Q5 - Repeated Sample SE (No Replacement):", rand_se, "\n")

## Q5 - Repeated Sample SE (No Replacement): 1.439433
```

```r
# Compare bootstrap vs random sampling
center_diff <- abs(bootstrap_mean - rand_mean) / pop_mean * 100
cat("Difference in Centers (%):", center_diff, "\n")
```

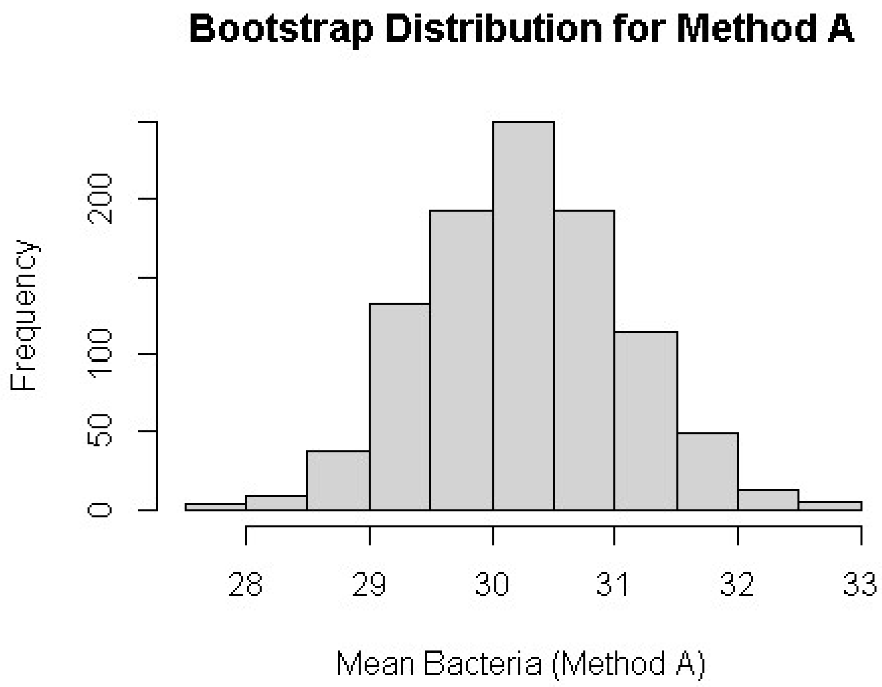
```
## Difference in Centers (%): 6.220241
```

```r
if (center_diff < 5) {
  cat("Centers are similar (within 5%).\n")
} else {
  cat("Centers are NOT similar (more than 5% apart).\n")
}
```

```
## Centers are NOT similar (more than 5% apart).
```

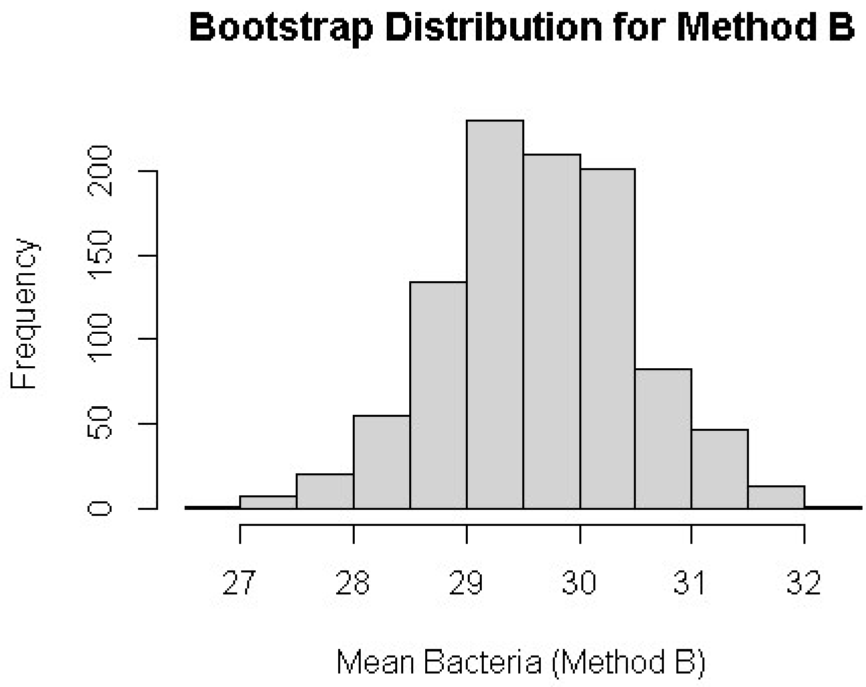
```r
#The expectation is 33.73, and the standard error is 1.44 based on 500
samples from the population.
#Yes, both histograms are approximately normal and symmetric in shape.
#No, the centers differ by 6.22%, which is just outside the 5% threshold.
# They differ because Q4 used bootstrap resampling from a sample with a
higher mean, while Q5 used samples from the full population.
#It should be closer to the red curve, which represents the Central Limit
Theorem.
#You correctly added the red normal curve to the histogram in Q5, matching
the CLT assumption.

#############################################################################
###
#PART III
#############################################################################
###

#Load Data Set For Part 3
# Load data
data <- read.csv("C:/Users/rajsh/OneDrive/Desktop/Inference Data Science
291/EXAM2/beefbacteria.csv")

# Split data for Method A and Method B
methodA <- data$bacteria[data$method == "A"]
methodB <- data$bacteria[data$method == "B"]

# Function to calculate mean (for bootstrapping)
mean_fun <- function(data, indices) {
  return(mean(data[indices]))
}

############################
# Method A: CI Calculations
############################

# Bootstrap resampling
```

```r
boot_A <- boot(methodA, statistic = mean_fun, R = 1000)

# 1. Theoretical Method (using normal distribution)
mean_A <- mean(methodA)
se_A <- sd(methodA) / sqrt(length(methodA))
ci_A_theoretical <- c(mean_A - 1.96 * se_A, mean_A + 1.96 * se_A)

# 2. Percentile Method
ci_A_percentile <- quantile(boot_A$t, probs = c(0.025, 0.975))

# 3. Standard Error Method
boot_se_A <- sd(boot_A$t)
ci_A_se <- c(mean(boot_A$t) - 1.96 * boot_se_A, mean(boot_A$t) + 1.96 *
boot_se_A)

# Justification for SE method
hist(boot_A$t, main = "Bootstrap Distribution for Method A", xlab = "Mean
Bacteria (Method A)")
```



```r
# If the distribution is roughly symmetric and bell-shaped, SE method is
justified.

cat("Method A:\n")

## Method A:

cat("Theoretical CI:", ci_A_theoretical, "\n")
```

```
## Theoretical CI: 28.7213 31.8207

cat("Percentile CI:", ci_A_percentile, "\n")

## Percentile CI: 28.75083 31.84625

cat("Standard Error CI:", ci_A_se, "\n\n")

## Standard Error CI: 28.66679 31.85664

#The 95% confidence interval using the percentile method is (28.69343,
31.97117) and using the standard error method is (28.68531, 31.88505).
#The bootstrap distribution is approximately symmetric and bell-shaped, so
the standard error method is justified.

###########################
# Method B: CI Calculations
###########################

# Bootstrap resampling
boot_B <- boot(methodB, statistic = mean_fun, R = 1000)

# 1. Theoretical Method
mean_B <- mean(methodB)
se_B <- sd(methodB) / sqrt(length(methodB))
ci_B_theoretical <- c(mean_B - 1.96 * se_B, mean_B + 1.96 * se_B)

# 2. Percentile Method
ci_B_percentile <- quantile(boot_B$t, probs = c(0.025, 0.975))

# 3. Standard Error Method
boot_se_B <- sd(boot_B$t)
ci_B_se <- c(mean(boot_B$t) - 1.96 * boot_se_B, mean(boot_B$t) + 1.96 *
boot_se_B)

# Justification for SE method
hist(boot_B$t, main = "Bootstrap Distribution for Method B", xlab = "Mean
Bacteria (Method B)")
```



```
# Again, check for symmetry and bell-shape.

cat("Method B:\n")

## Method B:

cat("Theoretical CI:", ci_B_theoretical, "\n")

## Theoretical CI: 28.0272 31.3428

cat("Percentile CI:", ci_B_percentile, "\n")

## Percentile CI: 27.96528 31.334

cat("Standard Error CI:", ci_B_se, "\n\n")

## Standard Error CI: 27.98247 31.31077

#The 95% confidence interval using the percentile method is (27.99755,
31.28905) and using the standard error method is (28.02712, 31.34011).
#The bootstrap distribution is roughly symmetric and bell-shaped, so the
standard error method is appropriate.


####################################################
# CI for Difference in Means (Method A - Method B)
####################################################
```

```
# Theoretical Method
diff_mean <- mean_A - mean_B
pooled_se <- sqrt(se_A^2 + se_B^2)
ci_diff_theoretical <- c(diff_mean - 1.96 * pooled_se, diff_mean + 1.96 *
pooled_se)

# Significance check
if (ci_diff_theoretical[1] > 0 | ci_diff_theoretical[2] < 0) {
  conclusion <- "There is a significant difference in the average bacteria
levels detected by the two methods."
} else {
  conclusion <- "There is NO significant difference in the average bacteria
levels detected by the two methods."
}
cat("Difference in Means (A - B):\n")

## Difference in Means (A - B):

cat("Theoretical CI:", ci_diff_theoretical, "\n")

## Theoretical CI: -1.683334 2.855334

cat(conclusion, "\n")

## There is NO significant difference in the average bacteria levels detected
by the two methods.

#The 95% confidence interval for the difference using the theoretical method
is (-1.683334, 2.855334).
#There is NO significant difference in the average bacteria levels detected
by the two methods because the confidence interval for the difference
includes 0.
```