Math 536, HW 1

Navi Chawla

2026-02-03

Problem 1: Gender Pay Discrimination Analysis

Introduction

This analysis investigates potential gender-based pay discrimination at a Northern California tech company using salary data from random samples of 104 female and 115 male employees.

Data Loading and Summary

```
# Load data
data = read.csv("C:/Users/navic/Downloads/Categorical Data Analysis/HW1P1.csv", h=T)

# Obtain salary data and remove NAs
females = data$Females[!is.na(data$Females)]
males = data$Males[!is.na(data$Males)]

# Sample sizes
n_f = length(females)
n_m = length(males)

# Summary statistics
cat("Female employees: n =", n_f, "\n")
```

```
## Female employees: n = 104
```

```
cat("  Mean: $", round(mean(females), 2), "  |  SD: $", round(sd(females), 2),
    "  |  Median: $", round(median(females), 2), "\n\n", sep="")
```

```
##   Mean: $63753.81  |  SD: $39517.39  |  Median: $52135
```

```
cat("Male employees: n =", n_m, "\n")
```

```
## Male employees: n = 115
```

```
cat("  Mean: $", round(mean(males), 2), "  |  SD: $", round(sd(males), 2),
    "  |  Median: $", round(median(males), 2), "\n", sep="")
```

```
##   Mean: $78034.08  |  SD: $55630.88  |  Median: $58341
```

Part A: Classical Two-Sample T-Test (Using CLT)

Hypotheses:

- H_0: $\mu_{males} \leq \mu_{females}$ (no gender discrimination)
- H_A: $\mu_{males} > \mu_{females}$ (males earn more - discrimination exists)

```
# Welch's t-test (not assuming equal variances)
t_test = t.test(males, females, alternative = "greater")

cat("Difference in means (Male - Female): $", round(mean(males) - mean(females), 2), "\n", sep="")
```

```
## Difference in means (Male - Female): $14280.27
```

```
cat("t-statistic:", round(t_test$statistic, 4), "  |  df:", round(t_test$parameter, 2),
    "  |  P-value:", format(t_test$p.value, scientific = TRUE), "\n")
```

```
## t-statistic: 2.2054   |  df: 205.8   |  P-value: 1.426607e-02
```

Decision: Reject H_0 at $\alpha = 0.05$ (p-value = 0.01427)

Interpretation: The p-value of 0.01427 means there is only a 1.43% chance of observing a salary difference this large (or larger) if there were truly no gender discrimination. This provides statistically significant evidence that males earn more than females.

```
ci_classical = t.test(males, females)$conf.int
cat("95% Confidence Interval: [$", round(ci_classical[1], 2), ", $",
    round(ci_classical[2], 2), "]\n", sep="")
```

```
## 95% Confidence Interval: [$1514.24, $27046.3]
```

Part B: Bootstrap Hypothesis Test

```
# Observed difference
obs_diff = mean(males) - mean(females)

# Pool all data (assume H0 is true)
pooled_data = c(females, males)

# Bootstrap under null hypothesis
set.seed(536)
n_bootstrap = 10000
bootstrap_diffs = rep(0, n_bootstrap)

for(i in 1:n_bootstrap) {
  boot_females = sample(pooled_data, size = n_f, replace = TRUE)
  boot_males = sample(pooled_data, size = n_m, replace = TRUE)
```

```r
  bootstrap_diffs[i] = mean(boot_males) - mean(boot_females)
}

# Calculate p-value
pvalue_bootstrap = length(bootstrap_diffs[bootstrap_diffs >= obs_diff]) / n_bootstrap

cat("Observed difference: $", round(obs_diff, 2), "\n", sep="")
```

```
## Observed difference: $14280.27
```

```r
cat("Bootstrap P-value:", round(pvalue_bootstrap, 4), "\n")
```

```
## Bootstrap P-value: 0.0136
```

Decision: Reject H_0 at $\alpha = 0.05$ (p-value = 0.0136)

Interpretation: The p-value of 0.0136 means only 1.36% of bootstrap samples (generated under the assumption of no discrimination) showed a difference as large as what we observed. This provides strong evidence of gender discrimination.

```r
hist(bootstrap_diffs, breaks = 50,
     main = "Bootstrap Distribution Under H0",
     xlab = "Difference in Means (Male - Female, $)",
     col = "lightgreen",
     border = "white")
abline(v = obs_diff, col = "red", lwd = 2, lty = 2)
abline(v = 0, col = "black", lwd = 2)
legend("topright",
       legend = c("Observed", "H0: zero"),
       col = c("red", "black"),
       lty = c(2, 1),
       lwd = 2)
```

Summary: Both methods agree (both p-values < 0.05). Classical p-value = 1.426607e-02, Bootstrap p-value = 0.0136.

Report to Company

Our statistical analysis provides compelling evidence of gender-based pay discrimination within your tech company. Male employees earn an average of **$14,280.27** more than female employees. Both the classical t-test (p = 0.0143) and bootstrap test (p = 0.0136) reject the null hypothesis at $\alpha = 0.05$, indicating this salary gap is highly unlikely to occur by random chance alone. The 95% confidence interval ($1,514 to $27,046) suggests the discrimination is both statistically significant and economically meaningful. We strongly recommend conducting a comprehensive audit of compensation practices, reviewing promotion and hiring procedures for potential bias, and developing a remediation plan to address these disparities.

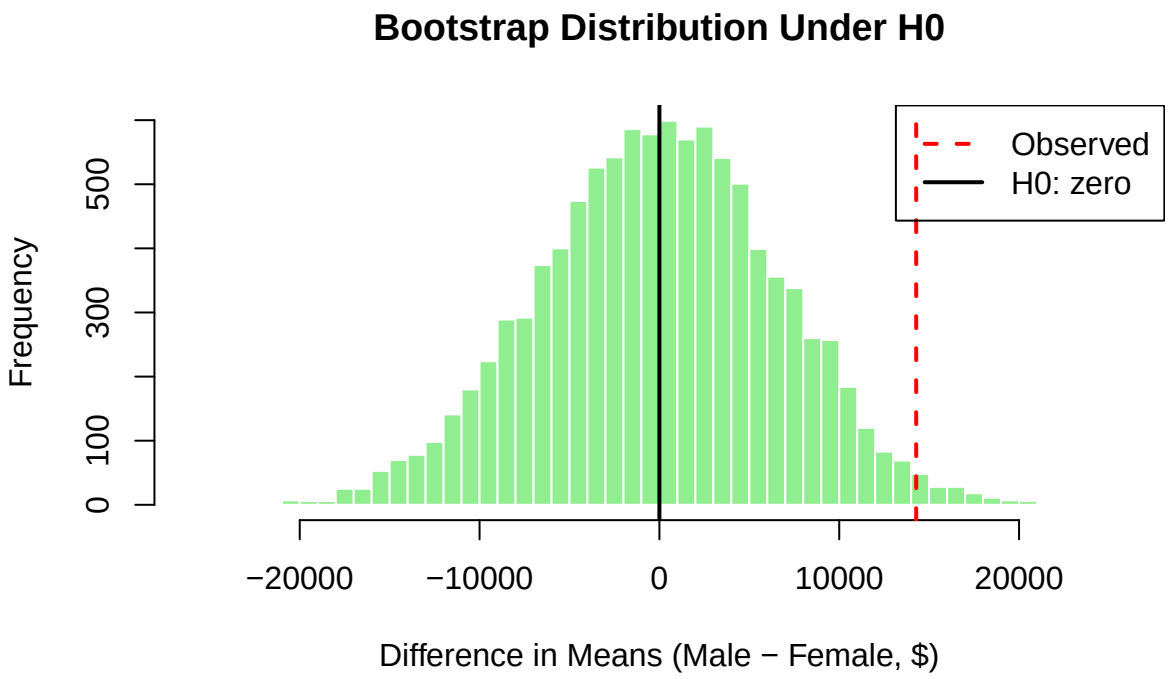

Figure 1: Bootstrap distribution of mean differences under the null hypothesis

Problem 2: Does Bootstrapping Always Work?

Introduction

We investigate whether bootstrap estimates of the 90th percentile are unbiased, using the adult income dataset as our population (n = 32,561).

```
# Load adult dataset
cols <- c("age", "workclass", "fnlwgt", "education", "education_num",
          "marital_status", "occupation", "relationship", "race", "sex",
          "capital_gain", "capital_loss", "hours_per_week", "native_country", "income")

adult_data <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data",
                       header = FALSE, col.names = cols,
                       na.strings = " ?", strip.white = TRUE)

population = adult_data$age
true_90tile = quantile(population, 0.90)

cat("Population size:", length(population), "\n")
```

```
## Population size: 32561
```

```
cat("Population 90th percentile:", true_90tile, "\n")
```

```
## Population 90th percentile: 58
```

Part A: Is the Sample 90th Percentile Unbiased?

We draw 10,000 samples of size 20 from the population and examine whether the sample 90th percentile is an unbiased estimator.

```
set.seed(536)
n_samples = 10000
sample_size = 20
sample_90tiles = rep(0, n_samples)

for(i in 1:n_samples) {
  sample_i = sample(population, size = sample_size, replace = FALSE)
  sample_90tiles[i] = quantile(sample_i, 0.90)
}

cat("Mean of sample 90th percentiles:", round(mean(sample_90tiles), 2), "\n")
```

```
## Mean of sample 90th percentiles: 55.62
```

```
cat("True population 90th percentile:", true_90tile, "\n")
```

```
## True population 90th percentile: 58
```

```
cat("Bias:", round(mean(sample_90tiles) - true_90tile, 2), "\n")
```

```
## Bias: -2.38
```

Answer: No, the 90th percentile from a sample of size 20 is **NOT an unbiased estimator** of the population 90th percentile.

Finding: The 90th percentile estimator is biased — it underestimates the true population 90th percentile by about 2.38 years on average.

```
hist(sample_90tiles, breaks = 50,
     main = "Distribution of Sample 90th Percentiles (n=20)",
     xlab = "90th Percentile (Age)",
     col = "lightblue",
     border = "white")
abline(v = true_90tile, col = "red", lwd = 2, lty = 2)
abline(v = mean(sample_90tiles), col = "blue", lwd = 2, lty = 2)
legend("topleft",
       legend = c(paste("True =", true_90tile),
                  paste("Mean =", round(mean(sample_90tiles), 1)),
                  paste("Bias =", round(mean(sample_90tiles) - true_90tile, 2))),
       col = c("red", "blue", "black"),
       lty = c(2, 2, 1), lwd = 2, cex = 0.9)
```

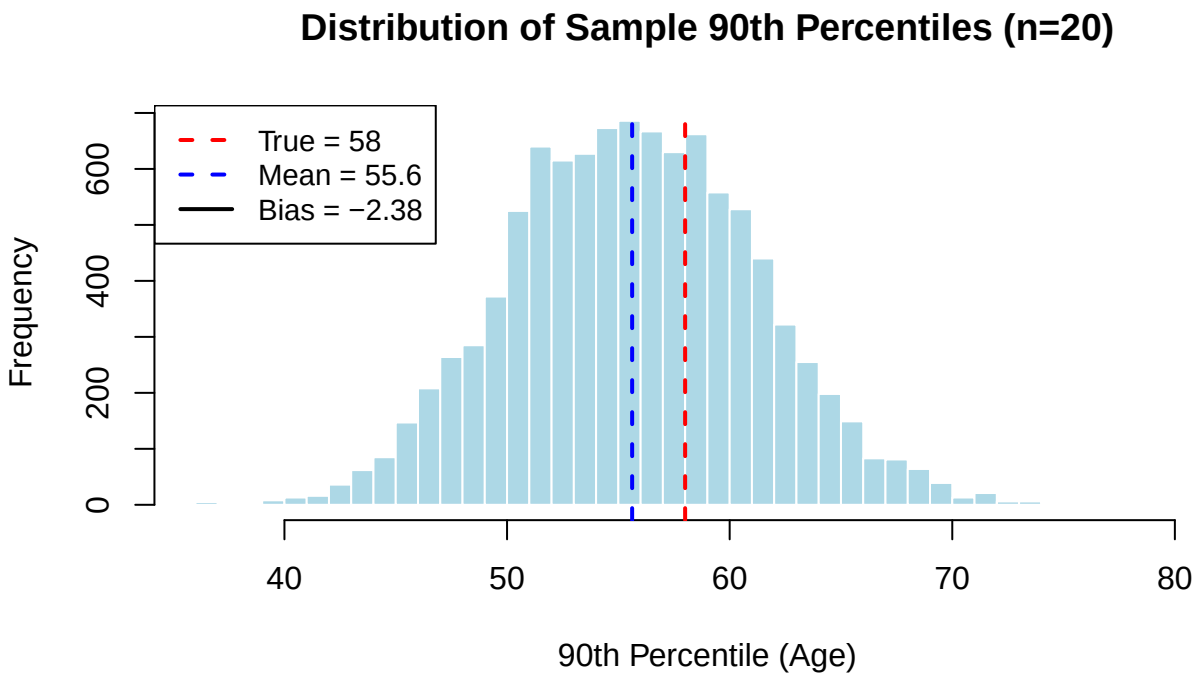

Figure 2: Distribution showing bias in sample 90th percentile estimates

Part B: Does Bootstrap Replicate the Bias?

We take one sample and bootstrap 10,000 times to see if bootstrap also shows bias.

```
set.seed(123)
original_sample = sample(population, size = 20, replace = FALSE)
original_90tile = quantile(original_sample, 0.90)

n_bootstrap = 10000
bootstrap_90tiles = rep(0, n_bootstrap)

for(i in 1:n_bootstrap) {
  boot_sample = sample(original_sample, size = 20, replace = TRUE)
  bootstrap_90tiles[i] = quantile(boot_sample, 0.90)
}

bootstrap_bias = mean(bootstrap_90tiles) - original_90tile

cat("Original sample 90th percentile:", original_90tile, "\n")
```

```
## Original sample 90th percentile: 51.3
```

```
cat("Mean of bootstrap estimates:", round(mean(bootstrap_90tiles), 2), "\n")
```

```
## Mean of bootstrap estimates: 50.75
```

```r
cat("Bootstrap bias:", round(bootstrap_bias, 2), "\n")
```

```
## Bootstrap bias: -0.55
```

Answer: Yes, bootstrap estimates of the 90th percentile **are biased**.

Finding: Bootstrap underestimates the sample 90th percentile by 0.55 years. This demonstrates that bootstrap replicates the bias inherent in the original estimator.

Quantification: The bias = -0.547 (E[bootstrap estimate] - original sample estimate).

```r
hist(bootstrap_90tiles, breaks = 50,
     main = "Bootstrap Distribution of 90th Percentile",
     xlab = "90th Percentile (Age)",
     col = "lightgreen",
     border = "white")
abline(v = original_90tile, col = "blue", lwd = 2, lty = 2)
abline(v = mean(bootstrap_90tiles), col = "darkgreen", lwd = 2)
abline(v = true_90tile, col = "red", lwd = 2, lty = 3)
legend("topleft",
       legend = c(paste("Original =", round(original_90tile, 1)),
                  paste("Bootstrap mean =", round(mean(bootstrap_90tiles), 1)),
                  paste("True =", true_90tile)),
       col = c("blue", "darkgreen", "red"),
       lty = c(2, 1, 3), lwd = 2, cex = 0.85)
```

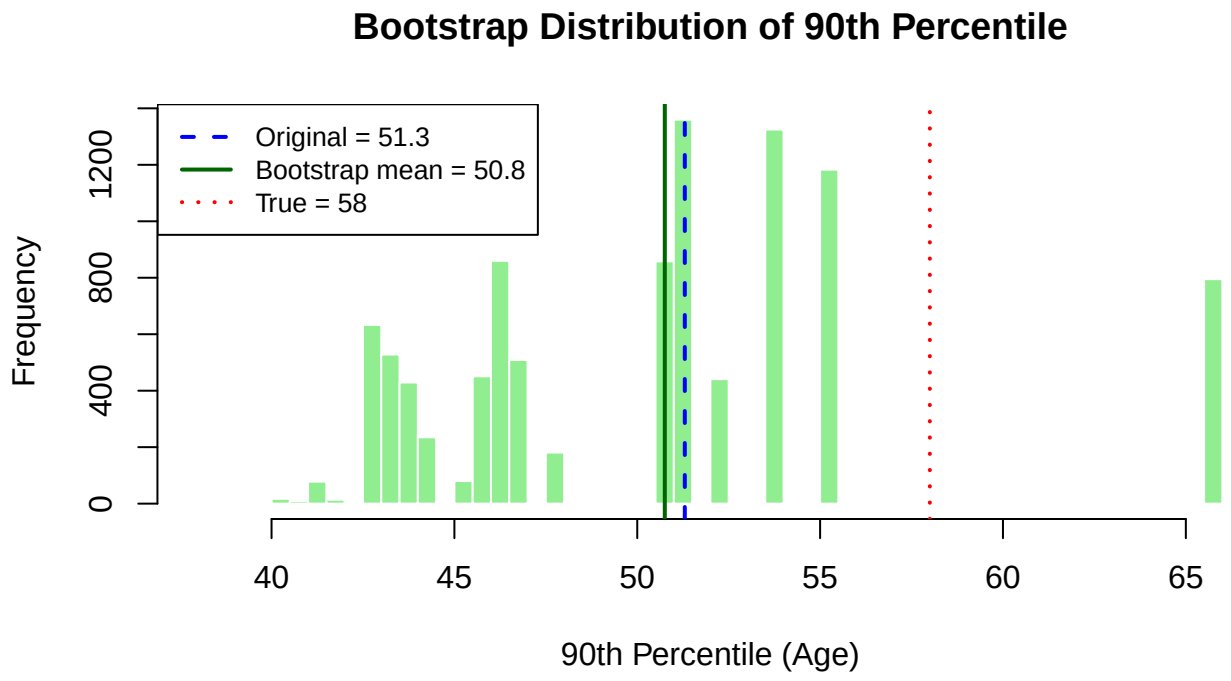

Figure 3: Bootstrap distribution showing bias relative to original sample

Part C: Bias-Corrected Confidence Interval

We create two confidence intervals: standard bootstrap and bias-corrected.

```
# Standard Bootstrap CI
ci_standard = quantile(bootstrap_90tiles, c(0.025, 0.975))

# Bias-Corrected CI (using Part A bias)
bias_from_partA = mean(sample_90tiles) - true_90tile
bootstrap_90tiles_corrected = bootstrap_90tiles - bias_from_partA
ci_corrected = quantile(bootstrap_90tiles_corrected, c(0.025, 0.975))

cat("Standard Bootstrap CI:", round(ci_standard[1], 2), "to", round(ci_standard[2], 2), "\n")
```

```
## Standard Bootstrap CI: 43 to 66
```

```
cat("  Contains true value (", true_90tile, ")? ",
    ifelse(true_90tile >= ci_standard[1] & true_90tile <= ci_standard[2], "YES", "NO"), "\n\n", sep="")
```

```
##   Contains true value (58)? YES
```

```
cat("Bias-Corrected CI:", round(ci_corrected[1], 2), "to", round(ci_corrected[2], 2), "\n")
```

```
## Bias-Corrected CI: 45.38 to 68.38
```

```
cat("  Contains true value (", true_90tile, ")? ",
    ifelse(true_90tile >= ci_corrected[1] & true_90tile <= ci_corrected[2], "YES", "NO"), "\n", sep="")
```

```
##   Contains true value (58)? YES
```

```
# Standard CI
hist(bootstrap_90tiles, breaks = 50,
     main = "Standard Bootstrap CI (No Bias Correction)",
     xlab = "90th Percentile (Age)",
     col = "lightgreen",
     border = "white")
abline(v = ci_standard, col = "darkgreen", lwd = 2)
abline(v = true_90tile, col = "red", lwd = 2, lty = 2)
legend("topleft", legend = c("95% CI", "True value = 58"),
       col = c("darkgreen", "red"), lty = c(1, 2), lwd = 2, cex = 0.9)
```

```
# Bias-corrected CI
hist(bootstrap_90tiles_corrected, breaks = 50,
     main = "Bias-Corrected Bootstrap CI (Using Part A)",
     xlab = "90th Percentile (Age)",
     col = "lightcoral",
     border = "white")
abline(v = ci_corrected, col = "darkred", lwd = 2)
abline(v = true_90tile, col = "red", lwd = 2, lty = 2)
legend("topleft", legend = c("95% CI", "True value = 58"),
       col = c("darkred", "red"), lty = c(1, 2), lwd = 2, cex = 0.9)
```

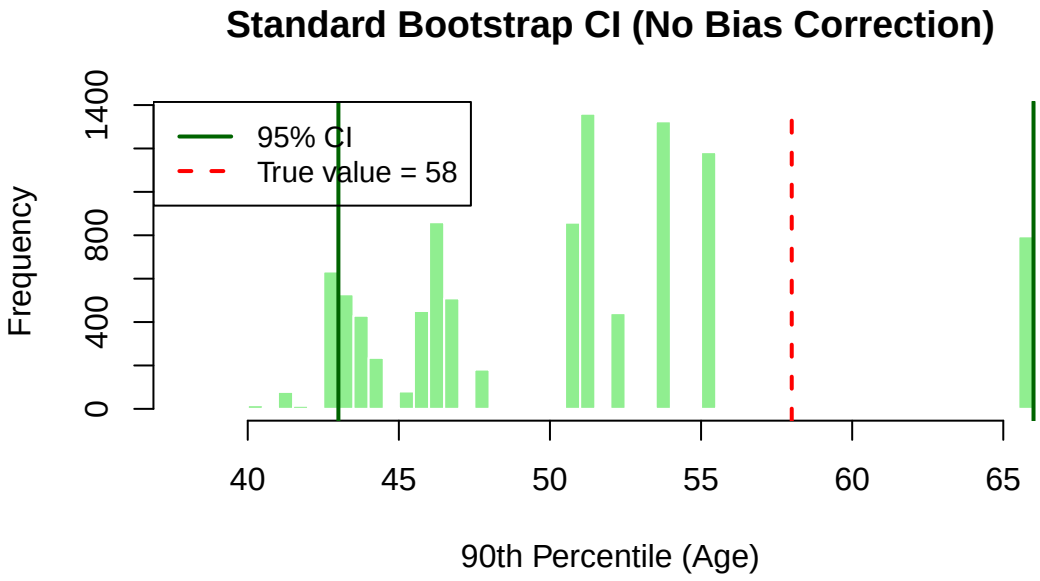

Figure 4: Standard bootstrap confidence interval (no bias correction)

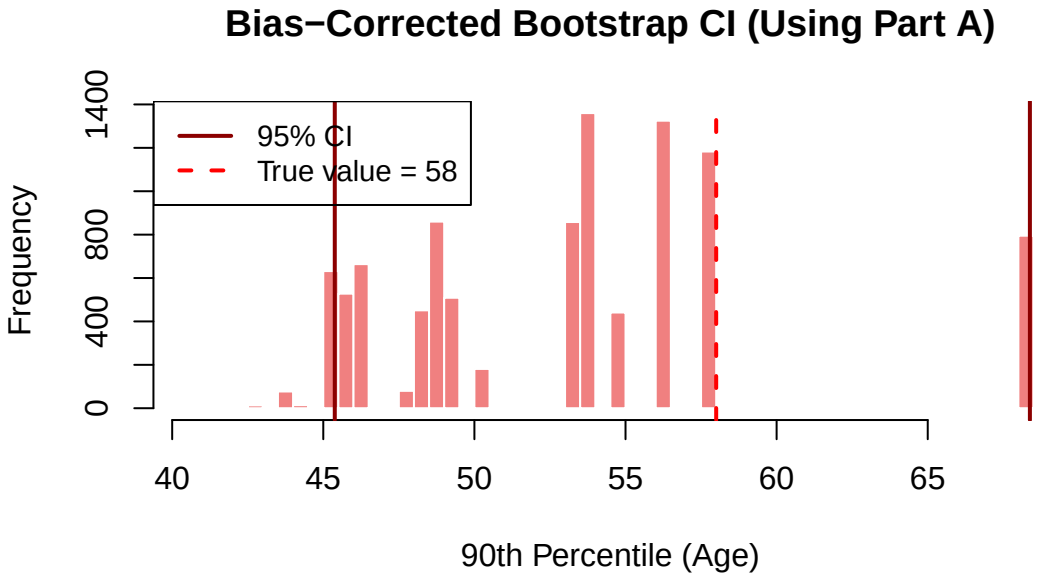

Figure 5: Bias-corrected bootstrap confidence interval using Part A bias

Answer: We used **Part A** (bias correction).

Why? Part A revealed the inherent bias of the 90th percentile estimator (bias = -2.38). This bias is a property of the estimator itself, not specific to our one sample. By correcting for this systematic bias, we obtain a more accurate confidence interval for the true population parameter.

Summary: Does Bootstrapping Always Work?

Short Answer: No. Bootstrap is not always unbiased, especially for extreme quantiles with small sample sizes.

Key Findings:

1. Sample 90th percentile has bias = -2.38
2. Bootstrap also shows bias = -0.55

3. Bias correction improves CI accuracy

Takeaway: Bootstrap mimics the sampling distribution, but if the original estimator is biased, bootstrap will also be biased. For extreme quantiles with small samples, bias correction is essential for accurate inference.