Homework 3

Due date: 09/26/2024

September 19, 2024

# Introduction

This exercise sheet contains a series of problems designed to test and enhance your understanding of the topics covered in the course. Please ensure that you attempt all problems and provide detailed solutions where necessary. If you have any questions or need clarification, feel free to reach out your TA.

# Exercises

## Exercise 1: New Estimator: *Y*1

For the following, consider the estimator *Y*1, the value of the first sampled item for *Y*1, *Y*2,... *Yn* sampled with replacement from a population with mean *µ* and variance *σ*2. The estimand is the population mean *µ.*

1. Is this estimator biased? Write a proof showing it is unbiased or deriving the bias. [5pts]

A close up of a board

Description automatically generated

1. Derive the variance of this estimator. [5pts]

A close up of a paper

Description automatically generated

1. Using the provided Pokémon sample and variable HP (hit points), create a bootstrap estimated sampling distribution of this estimator, plot a histogram of your bootstrap estimated sampling distribution, and report your 95% confidence interval. [5pts]

A screenshot of a computer code

Description automatically generated

A graph of a pyramid

Description automatically generated with medium confidence

1. If you had the population and could plot the sampling distribution of your estimator, how would it compare to the estimand? What information are you using to make that determination? [5pts]

If I had the population and could plot the sampling distribution of my estimator, they would be very similar, if not the same. According to part (a), this is an unbiased estimator, so the estimator should be completely representative of the estimand.

## Exercise 2: New Estimator: *Y*˜

For the following, consider the estimator  for *Y*1, *Y*2,... *Yn* sampled with replacement from a population with mean *µ* and variance *σ*2. The estimand is the population mean *µ*.

1. Is this estimator biased? Write a proof showing it is unbiased or deriving the bias. [5pts]

A close up of a paper

Description automatically generated

1. Derive the variance of this estimator. [5pts]

A close-up of a notebook

Description automatically generated

1. Using the provided Pokémon sample and variable HP (hit points), create a bootstrap estimated sampling distribution of this estimator, plot a histogram of your bootstrap estimated sampling distribution, and report your 95% confidence interval. [5pts]

A screenshot of a computer code

Description automatically generated

A graph of a pyramid

Description automatically generated with medium confidence

1. If you had the population and could plot the sampling distribution of your estimator, how would it compare to the estimand? What information are you using to make that determination? [5pts]

If I had the population and could plot the sampling distribution of my estimator, they would be different. This would be a very inaccurate estimator since it introduces bias and features a different variability from the population.

## Exercise 3: A bimodal situation

Use the following commands to generate your population: population = c(rnorm(n = 1000, mean = 1, sd = 2), rnorm(n = 1000, mean = 10, sd = 2)).

1. Compute the population mean by using the command ”mean” and plot an histogram for the population. [5pts]

A white background with numbers and symbols

Description automatically generated



A graph of a number of columns

Description automatically generated

1. Generate *sample*1 by sampling only 50 elements of the population lower than 1, and *sample*2 with 50 elements of the population greater than 10. Plot a histogram of *sample*1 and *sample*2. Compare the two histograms with the one of the populations obtained in (a). Is this the usual way we sample? [5pts]

A screenshot of a computer

Description automatically generated

A graph with numbers and a line

Description automatically generated with medium confidence

A graph with numbers and a line

Description automatically generated

No, this is not the typical way we sample. Typically, we would take a random sample that is representative of the entire population.

1. Obtain and plot the sampling distribution from the population for the mean using samples with sample size 50. Use the distribution to calibrate a 95% confidence interval.

A white rectangular object with a black border

Description automatically generated

A graph of a graph

Description automatically generated

1. Obtain and plot the bootstrapped sampling distribution from *sample*1 for the mean. Use the distribution to calibrate a 95% confidence interval. [5pts]

A screenshot of a computer

Description automatically generated

A graph of a graph

Description automatically generated

1. Obtain and plot the bootstrapped sampling distribution from *sample*2 for the mean. Use the distribution to calibrate a 95% confidence interval. [5pts]

Numbers and symbols on a white background

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

1. Compare the histograms obtained in (c), (d), and (e). Explain if they’re centered with respect to the population mean and explain why. [5pts]

Only the sample obtained in (c) is centered with respect to the population mean. This is because we did not take random samples. This introduced bias into our analysis of sample 1 and sample 2 which explains why each 95% CI is skewed rather than centered with respect to the population mean.

## Exercise 4: Difference in Means

For the following, use the provided Chihuahua sample (for a and b) and the alternative Chihuahua sample (part c).

1. Calculate the difference in mean weight between male and female chihuahuas in your sample. How many chihuahuas are male? how many are female? [5pts]

A screenshot of a computer code

Description automatically generated

1. Create a 95% confidence interval for the difference in means, calibrated using the bootstrap estimated sampling distribution. Create a histogram of your bootstrap estimated sampling distribution and annotate it with a vertical line to indicate your difference in means estimate from part (a). [5pts]



A graph of a graph

Description automatically generated with medium confidence

1. Using the alternate sample, report the sample difference in mean weight and how many male and female Chihuahuas there are. Repeat (b) with the alternate sample and compare your resulting histogram and 95% CI to part (b). [10 pts]

A number with a number on it

Description automatically generated with medium confidence

A graph of a pyramid

Description automatically generated with medium confidence

In this alternative sample, the sample difference in means is closer to the population difference in means than in the original sample. This means that the alternative sample is more representative of the overall population than the original sample.

## Exercise 5: Two estimators for *θ*

Given a sample *X*1*,...,Xn* of i.i.d. with *Xi* for *i* = 1*,...,n* having probability mass function of:

, and.

1. Calculate *E*[*Xi*] and *V ar*[*Xi*]. [10 pts]

A close-up of a math equation

Description automatically generated

1. Given the estimand *θ* and its estimator, show if this estimator is biased or not. [5pts]

A close up of a paper

Description automatically generated

1. Derive the variance of the estimator *θ*ˆ. [5pts]

A notebook with writing on it

Description automatically generated

1. Plot using R how *V ar*[*θ*ˆ] changes for 30 ≤ *n* ≤ 100 when

A graph with a blue line

Description automatically generated

1. Given the estimand *θ* and its estimator *θ*ˆ∗ = |*Xn*|, show if this estimator is biased or not.

[5pts]

A close-up of a paper

Description automatically generated

1. Derive the variance of the estimator *θ*ˆ∗. [5pts]

A math equations on a piece of paper

Description automatically generated

1. Plot using R how *V ar*[*θ*ˆ∗] changes for 30 ≤ *n* ≤ 100 when 

A graph with a line and numbers

Description automatically generated

1. Which of the two estimators do you prefer? Why? Consider the bias and the variance and how they change with *n* to answer. [5pts]

I prefer the second estimator because it is unbiased and features a variance of 0. No matter how large a sample I choose to include, the variance is always zero. Therefore, I would prefer the second estimator.

# Submission Instructions

Please submit your completed exercises by **September 26** through **gradescope**. Ensure that your solutions are well-organized, clearly written, and include all necessary calculations and explanations. Questions about submission should be directed to your TA.

# Helpful Resources

To better assist you in the completion of this exercise sheet, we suggest you to review the following material:

* **Lecture 3** - covering estimated sampling distributions, bootstrapping, and calibration of confidence intervals;
* **Lecture 4** - covering random variables, probability, expected values, and properties of expected values;
* **Lecture 5** - covering variance and properties of variance;
* **Lecture 6** - covering sampling as random variables, bias and variance of estimators;
* **Lab** - practicing all of the above