Veronica Vargas

Homework 3

Due date: 09/26/2024

September 19, 2024

1 Introduction

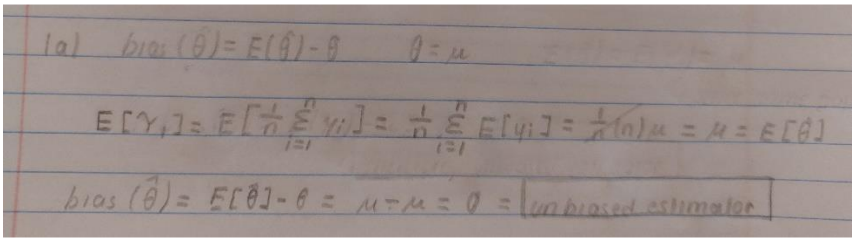
This exercise sheet contains a series of problems designed to test and enhance your understanding of the topics covered in the course. Please ensure that you attempt all problems and provide detailed solutions where necessary. If you have any questions or need clarification, feel free to reach out your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

2 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,... Y_n sampled with replacement from a population with mean μ and variance σ^2. The estimand is the population mean μ.

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

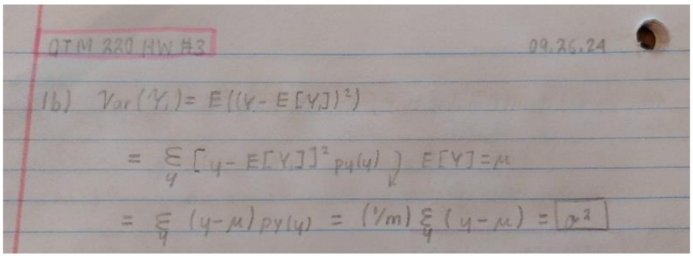


1a) $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta \qquad \theta = \mu$

$$E[Y_1] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[y_i] = \frac{1}{\cancel{n}}(\cancel{n})\mu = \mu = E[\hat{\theta}]$$

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta = \mu - \mu = 0 = \boxed{\text{unbiased estimator}}$$

(b) Derive the variance of this estimator. [5pts]



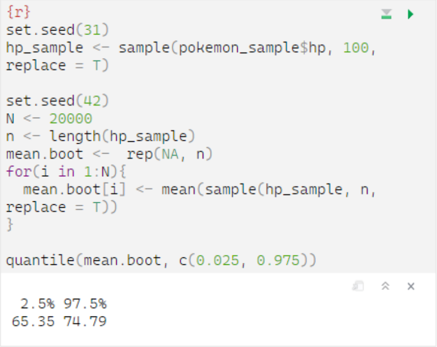
QTM 220 HW #3 09.26.24

1b) $\text{Var}(Y_1) = E((Y - E[Y_1])^2)$

$$= \sum_{y} [y - E[Y_1]]^2 p_Y(y) \quad E[Y] = \mu$$

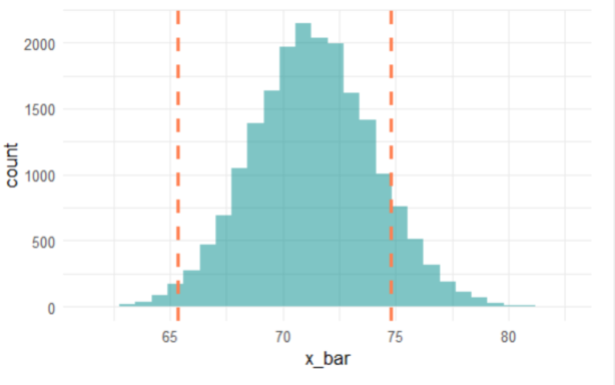
$$= \sum_{y} (y-\mu) p_Y(y) = (1/m) \sum_{y} (y-\mu) = \boxed{\sigma^2}$$

(c) 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]



```
{r}
set.seed(31)
hp_sample <- sample(pokemon_sample$hp, 100,
replace = T)

set.seed(42)
N <- 20000
n <- length(hp_sample)
mean.boot <-  rep(NA, n)
for(i in 1:N){
  mean.boot[i] <- mean(sample(hp_sample, n,
replace = T))
}

quantile(mean.boot, c(0.025, 0.975))
```

```
 2.5% 97.5%
65.35 74.79
```



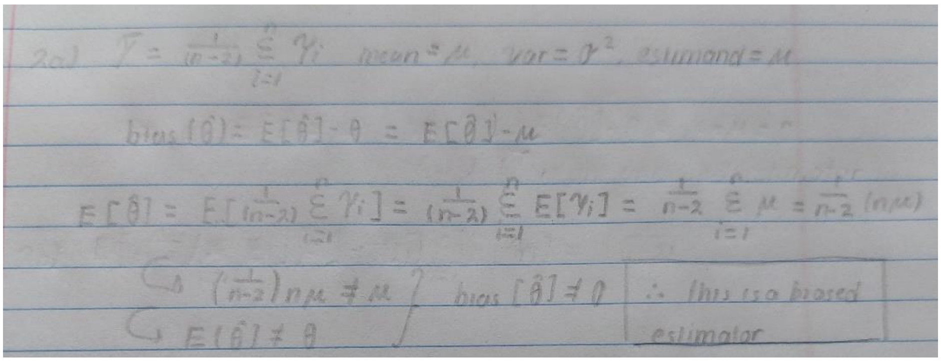
(d) If you had the population and could plot the sampling distribution of your estimator, how would it compare to the estimand? What information are you using to make that determination? [5pts]

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: $\tilde{Y}$

For the following, consider the estimator $\tilde{Y} = \frac{1}{n-2}\sum_{i=1}^{n} Y_i$ for Y_1, Y_2,... Y_n sampled with replacement from a population with mean μ and variance σ^2. The estimand is the population mean μ.

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



2a) $\tilde{Y} = \frac{1}{(n-2)} \sum_{i=1}^{n} Y_i$ mean $= \mu$, var $= \sigma^2$, estimand $= \mu$

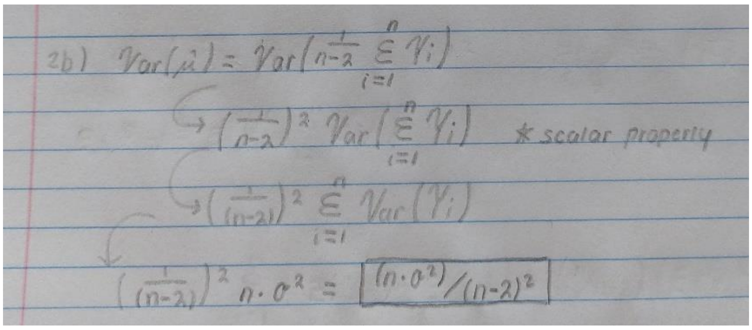
bias$(\hat{\theta}) = E[\hat{\theta}] - \theta = E[\hat{\theta}] - \mu$

$E[\hat{\theta}] = E[\frac{1}{(n-2)} \sum_{i=1}^{n} Y_i] = \frac{1}{(n-2)} \sum_{i=1}^{n} E[Y_i] = \frac{1}{n-2} \sum_{i=1}^{n} \mu = \frac{1}{n-2}(n\mu)$

$\hookrightarrow (\frac{1}{n-2}) n\mu \neq \mu$ } bias$[\hat{\theta}] \neq 0$ $\boxed{\therefore \text{This is a biased estimator}}$

$\hookrightarrow E[\hat{\theta}] \neq \theta$

(b) Derive the variance of this estimator. [5pts]



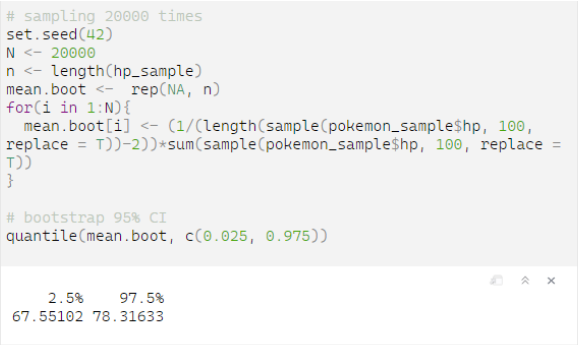
2b) $Var(\hat{\mu}) = Var(\frac{1}{n-2} \sum_{i=1}^{n} Y_i)$

$\hookrightarrow (\frac{1}{n-2})^2 \, Var(\sum_{i=1}^{n} Y_i)$ ✱ scalar property

$\hookrightarrow (\frac{1}{(n-2)})^2 \sum_{i=1}^{n} Var(Y_i)$

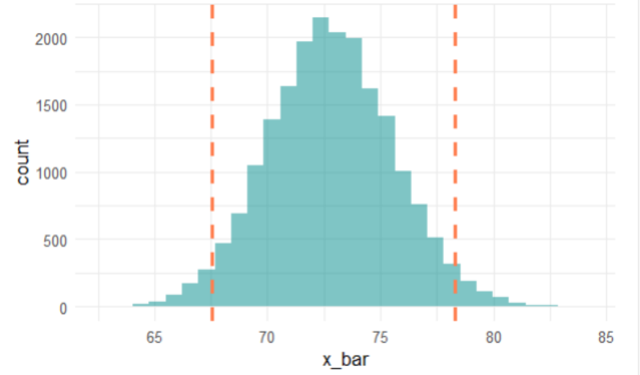
$(\frac{1}{(n-2)})^2 \, n \cdot \sigma^2 = \boxed{(n \cdot \sigma^2)/(n-2)^2}$

(c) 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]



```
# sampling 20000 times
set.seed(42)
N <- 20000
n <- length(hp_sample)
mean.boot <-  rep(NA, n)
for(i in 1:N){
  mean.boot[i] <- (1/(length(sample(pokemon_sample$hp, 100,
replace = T))-2))*sum(sample(pokemon_sample$hp, 100, replace =
T))
}

# bootstrap 95% CI
quantile(mean.boot, c(0.025, 0.975))
```

```
    2.5%    97.5%
67.55102 78.31633
```



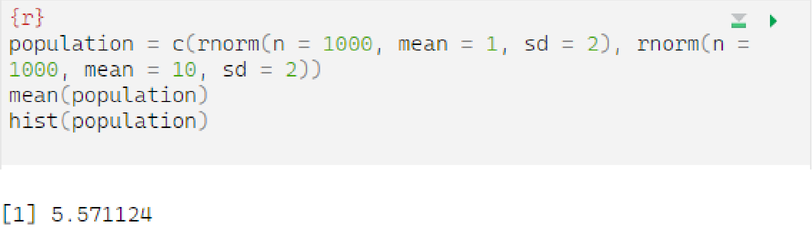
(d) If you had the population and could plot the sampling distribution of your estimator, how would it compare to the estimand? What information are you using to make that determination? [5pts]

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

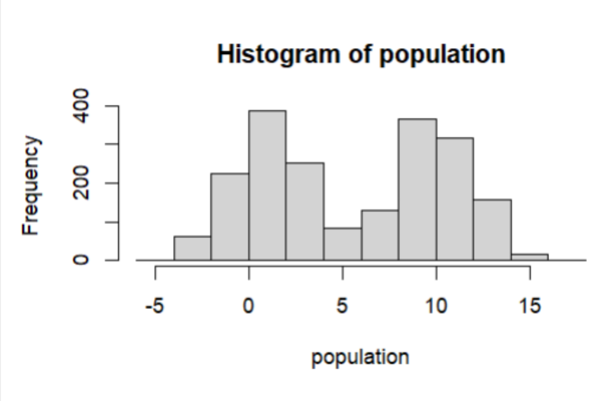
Use the following commands to generate your population: population = c(rnorm(n = 1000, mean = 1, sd = 2), rnorm(n = 1000, mean = 10, sd = 2)).

(a) Compute the population mean by using the command "mean" and plot an histogram for the population. [5pts]

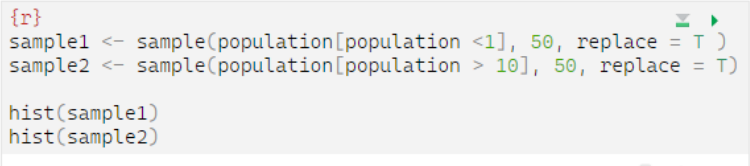


```
{r}
population = c(rnorm(n = 1000, mean = 1, sd = 2), rnorm(n =
1000, mean = 10, sd = 2))
mean(population)
hist(population)
```

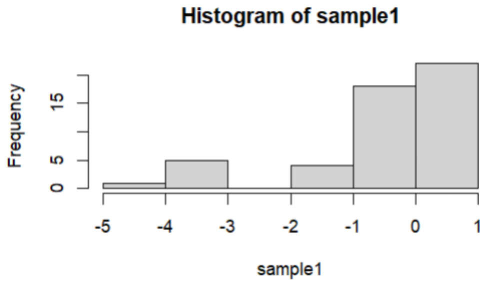
```
[1] 5.571124
```

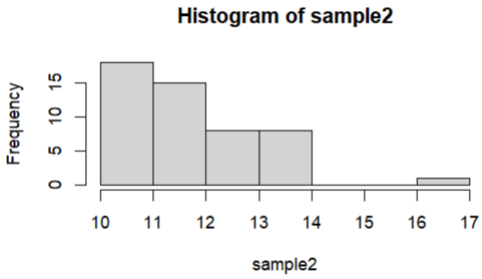


(b) Generate $sample_1$ by sampling only 50 elements of the population lower than 1, and $sample_2$ with 50 elements of the population greater than 10. Plot a histogram of $sample_1$ and $sample_2$. Compare the two histograms with the one of the populations obtained in (a). Is this the usual way we sample? [5pts]



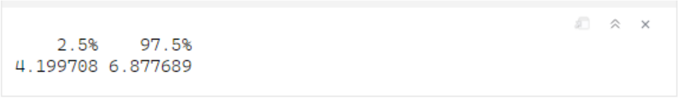
```
{r}
sample1 <- sample(population[population <1], 50, replace = T )
sample2 <- sample(population[population > 10], 50, replace = T)

hist(sample1)
hist(sample2)
```



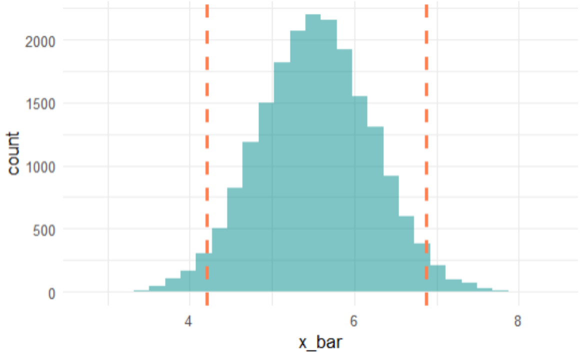


No, this is not the typical way we sample. Typically, we would take a random sample that is representative of the entire population.

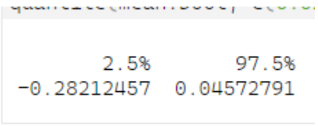
(c) Obtain and plot the sampling distribution from the population for the mean using samples with sample size 50. Use the distribution to calibrate a 95% confidence interval.



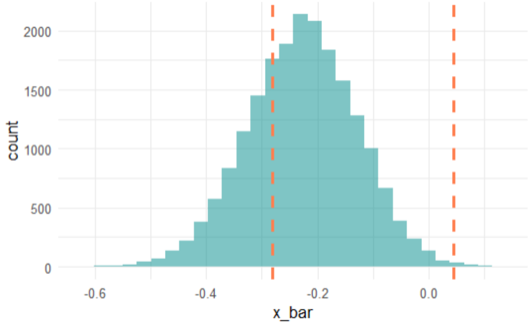
```
    2.5%    97.5% 
4.199708 6.877689 
```



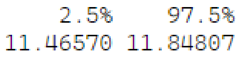
(d) Obtain and plot the bootstrapped sampling distribution from $sample_1$ for the mean. Use the distribution to calibrate a 95% confidence interval. [5pts]



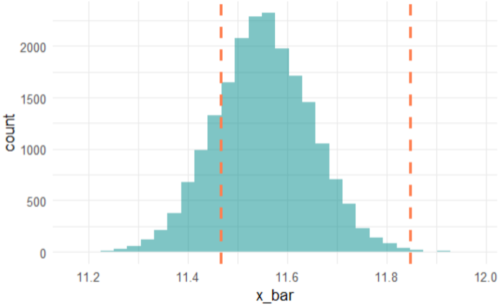
```
quantile(mean.boot, c(0.0

        2.5%        97.5%
-0.28212457  0.04572791
```



(e) Obtain and plot the bootstrapped sampling distribution from *sample*$_2$ for the mean. Use the distribution to calibrate a 95% confidence interval. [5pts]



```
    2.5%    97.5%
11.46570 11.84807
```



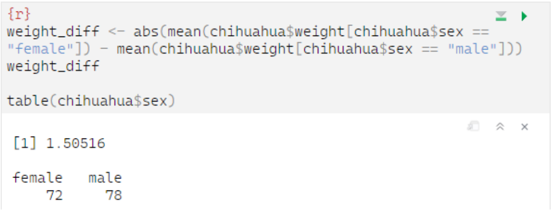
(f) Compare the histograms obtained in (c), (d), and (e). Explain if they're centered with respect to the population mean and explain why. [5pts]

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).

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

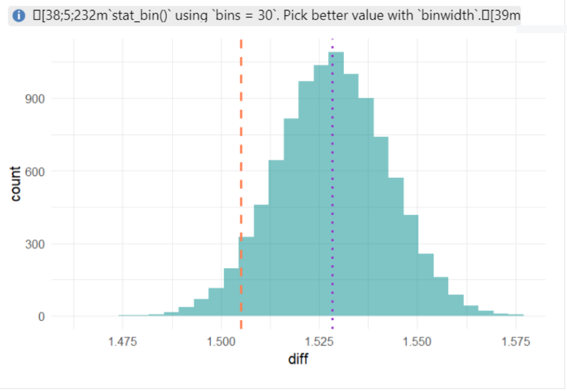


```
{r}
weight_diff <- abs(mean(chihuahua$weight[chihuahua$sex ==
"female"]) - mean(chihuahua$weight[chihuahua$sex == "male"]))
weight_diff

table(chihuahua$sex)
```

```
[1] 1.50516

female   male
    72     78
```

(b) 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]

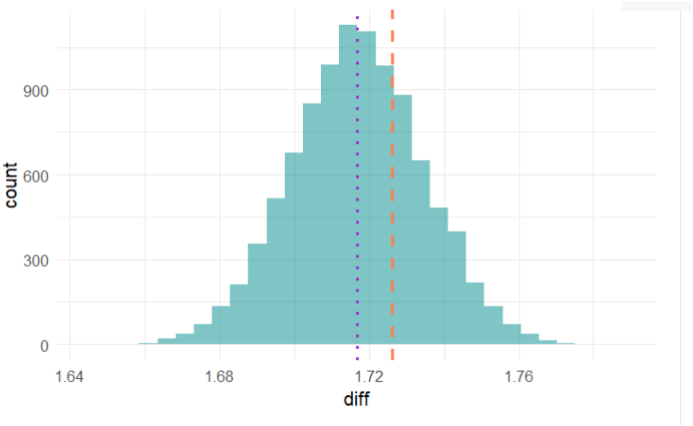
```
    2.5%    97.5% 
1.500773 1.555775 
```



ℹ □[38;5;232m`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.□[39m

(c) 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]

```
    2.5%    97.5% 
1.681966 1.751324 
```

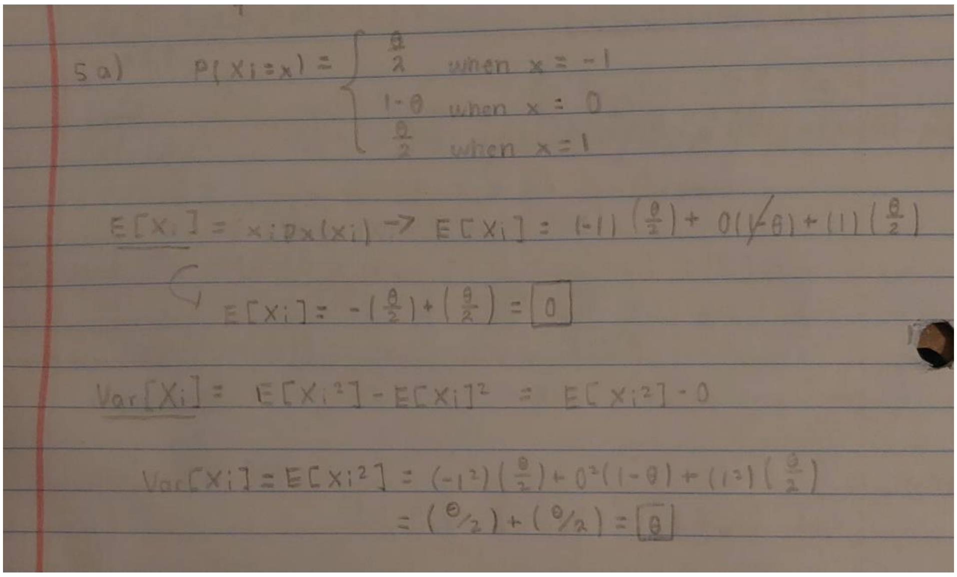


In this alternative sample, the sample difference in means is closer to the population difference in means than in the original sample. This means that the alternative sample is more representative of the overall population than the original sample.

Exercise 5: Two estimators for θ

Given a sample $X_1,...,X_n$ of i.i.d. with X_i for $i = 1,...,n$ having probability mass function of: $p_{X_i}(-1) = \frac{\theta}{2}$, $p_{X_i}(0) = 1 - \theta$, and $p_{X_i}(1) = \frac{\theta}{2}$.

(a) Calculate $E[X_i]$ and $Var[X_i]$. [10 pts]



5 a) $P(X_i=x) = \begin{cases} \frac{\theta}{2} & \text{when } x=-1 \\ 1-\theta & \text{when } x=0 \\ \frac{\theta}{2} & \text{when } x=1 \end{cases}$

$E[X_i] = x_i p_x(x_i) \rightarrow E[X_i] = (-1)(\frac{\theta}{2}) + 0(\cancel{1}-\theta) + (1)(\frac{\theta}{2})$

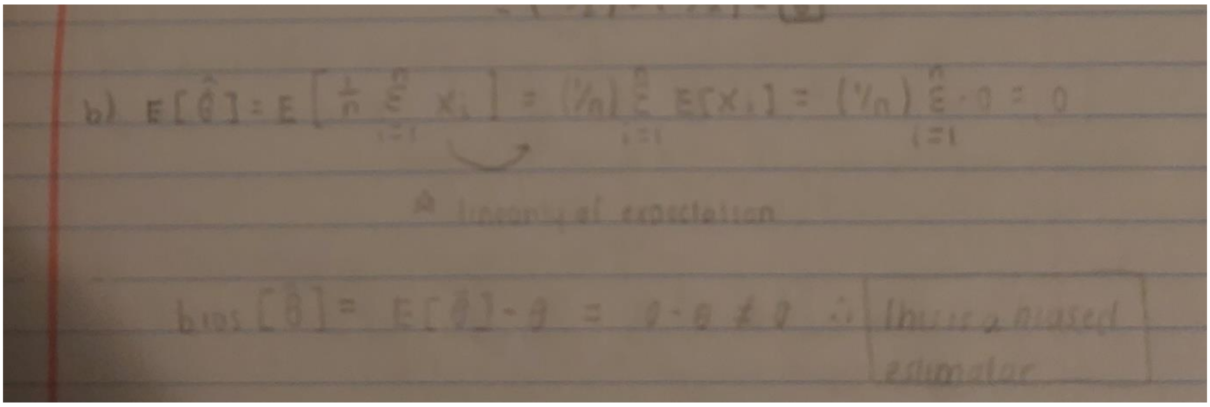
$E[X_i] = -(\frac{\theta}{2}) + (\frac{\theta}{2}) = \boxed{0}$

$Var[X_i] = E[X_i^2] - E[X_i]^2 = E[X_i^2] - 0$

$Var[X_i] = E[X_i^2] = (-1^2)(\frac{\theta}{2}) + 0^2(1-\theta) + (1^2)(\frac{\theta}{2})$

$= (\theta/2) + (\theta/2) = \boxed{\theta}$

(b) Given the estimand θ and its estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}|X_i|$, show if this estimator is biased or not. [5pts]

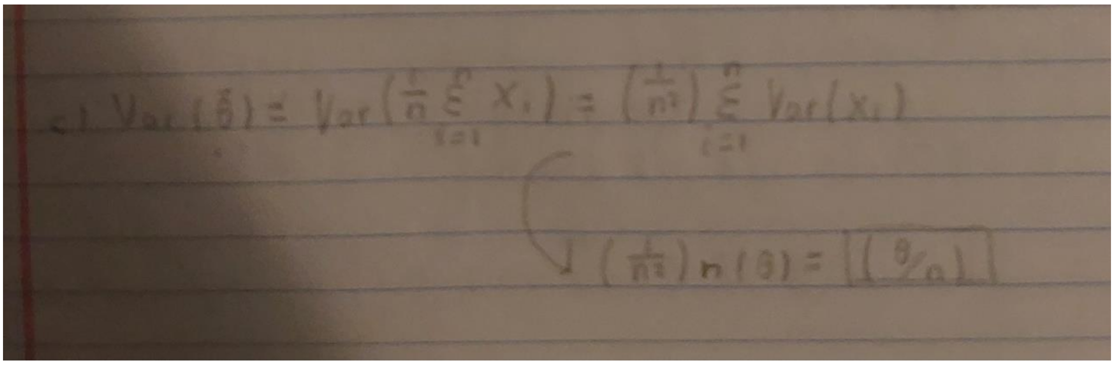


b) $E[\hat{\theta}] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = (1/n)\sum_{i=1}^{n} E[X_i] = (1/n)\sum_{i=1}^{n} 0 = 0$

linearity of expectation

$bias[\hat{\theta}] = E[\hat{\theta}] - \theta = 0 - \theta \neq 0 \therefore$ $\boxed{\text{This is a biased estimator}}$

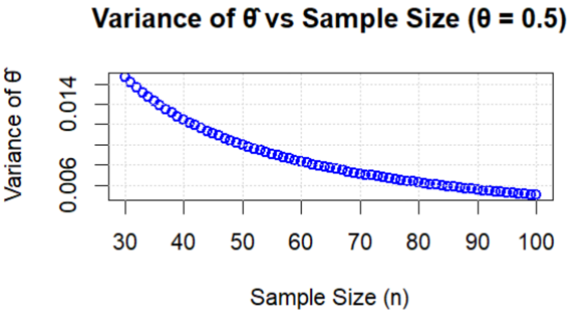
(c) Derive the variance of the estimator $\hat{\theta}$. [5pts]



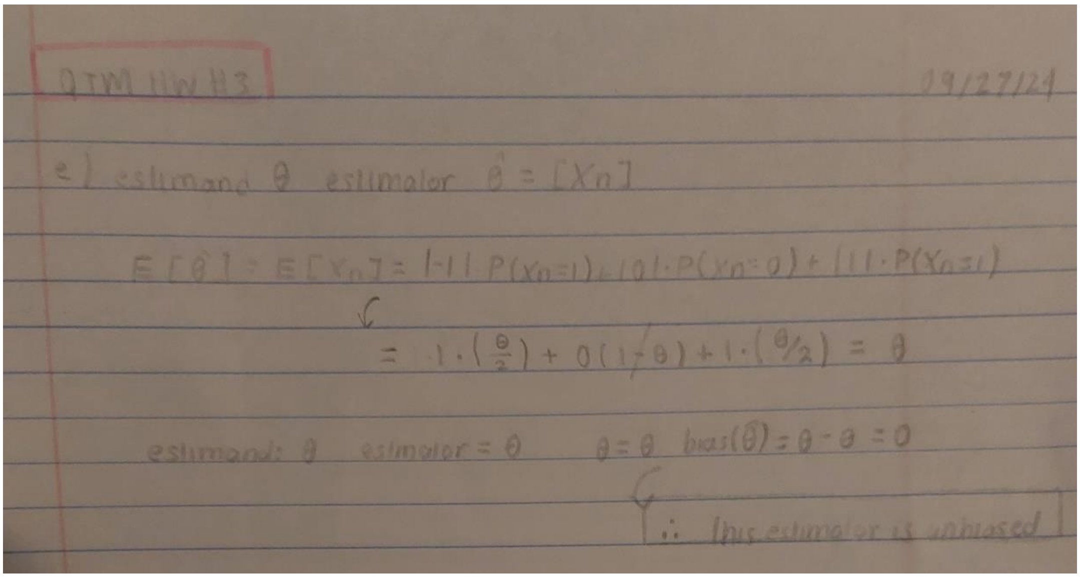
c) $Var(\hat{\theta}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = (\frac{1}{n^2})\sum_{i=1}^{n} Var(X_i)$

$(\frac{1}{n^2}) n(\theta) = \boxed{(\theta/n)}$

(d) Plot using R how $Var[\hat{\theta}]$ changes for $30 \leq n \leq 100$ when $\theta = \frac{1}{2}$. [5pts]



(e) Given the estimand θ and its estimator $\hat{\theta}^* = |X_n|$, show if this estimator is biased or not. [5pts]



QTM HW #3

09/27/24

e) estimand θ estimator $\hat{\theta} = [X_n]$

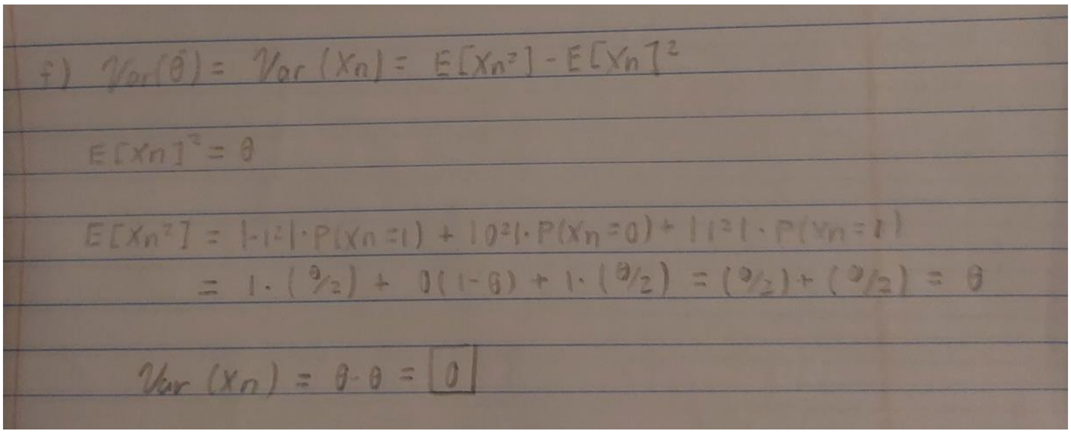
$E[\hat{\theta}] = E[X_n] = |-1| \cdot P(X_n=-1) + |0| \cdot P(X_n=0) + |1| \cdot P(X_n=1)$

$= 1 \cdot (\frac{\theta}{2}) + 0(1-\theta) + 1 \cdot (\theta/2) = \theta$

estimand: θ estimator $= \theta$ $\theta = \theta$ $\text{bias}(\hat{\theta}) = \theta - \theta = 0$

$\boxed{\therefore \text{ This estimator is unbiased}}$

(f) Derive the variance of the estimator $\hat{\theta}^*$. [5pts]



f) $Var(\hat{\theta}) = Var(X_n) = E[X_n^2] - E[X_n]^2$

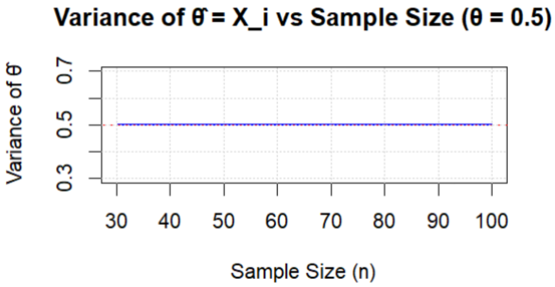
$E[X_n]^2 = \theta$

$E[X_n^2] = |-1^2| \cdot P(X_n=-1) + |0^2| \cdot P(X_n=0) + |1^2| \cdot P(X_n=1)$

$= 1 \cdot (\theta/2) + 0(1-\theta) + 1 \cdot (\theta/2) = (\theta/2) + (\theta/2) = \theta$

$Var(X_n) = \theta - \theta = \boxed{0}$

(g) Plot using R how $Var[\hat{\theta}^*]$ changes for $30 \leq n \leq 100$ when $\theta = \frac{1}{2}$. [5pts]



(h) 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.

3 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.

4 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