EC 320: Introduction to Econometrics

Problem Set 1

Total: 45 points
Due: On canvas, January 17th at 5 pm

Learning Outcomes:

- Understanding random variables and probabilities
- Understanding bias and variance of estimators
- Basic data summary and visualization

Checklist Before Handing In:

- Did you answer all questions?
- Did you answer all parts for each question?
- Were your answers too vague? If so, make them more precise to make sure they really answer the question being asked.

Instructions: You are encouraged to work with other students in the class, but you must provide original responses. To receive full credit, justify your answers and list your collaborators. For full credit on the computational exercises, include your code and output in addition to your answers. You will turn in digital copies of your responses on Canvas. Please note the list of acceptable file types on the submission page.

Name:

Collaborator 1:

Collaborator 2:

Analytical Questions

1. Consider an experiment where you flip three fair coins, record whether each coin landed heads or tails, and assign heads a value of 2 and tails a value of -1. You then multiply the values from each flip and describe this product as the random variable X. For example, if the coins landed `heads`, `heads`, `tails`, then you would multiply $2 \cdot 2 \cdot -1$ to find an outcome of $X = -4$.

   (a) Is X a discrete or continuous random variable?

   X is a discrete random variable.

   (b) Outline the probability distribution of X in a table.

Outcomes	X	**Probability**
HHH	8	1/8
HTT, TTH, THT	2	3/8
TTT	-1	1/8
HHT, THH, HTH	-4	3/8

   (c) What is the population mean of X? In other words, what is $E(X)$?

   $E(X) = 8 \cdot 1/8 + 2 \cdot 3/8 - 1 \cdot 1/8 - 4 \cdot 3/8 = 1/8$

   (d) What is the population variance of X?

   $\text{Var}(X) = E(X^2) - [E(X)]^2 = 8^2 \cdot 1/8 + 2^2 \cdot 3/8 + (-1)^2 \cdot 1/8 + (-4)^2 \cdot 3/8 - (1/8)^2 = 15.625 - 0.015625 = 15.609375.$

2. Consider the distributions of three estimators of the population mean μ:

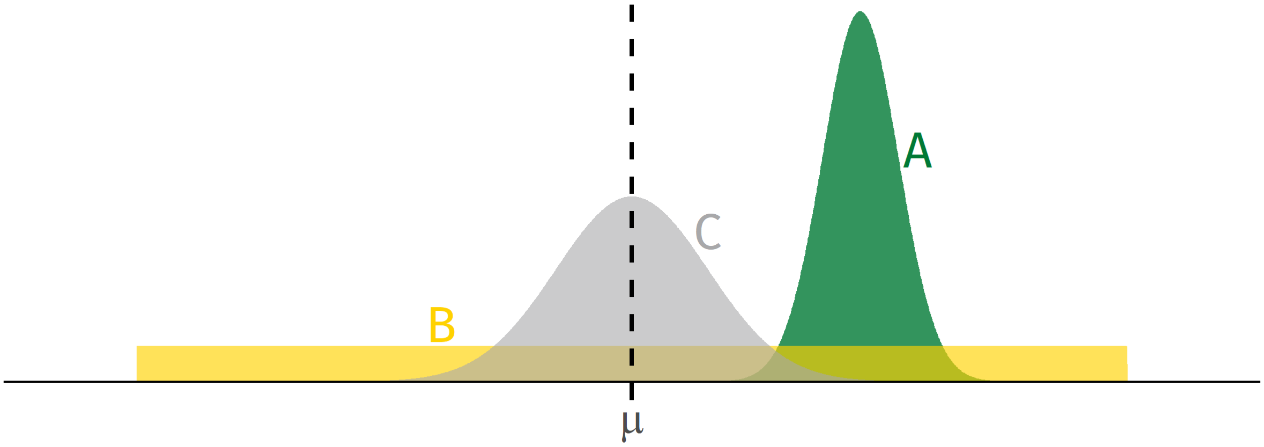


Figure 1: *

Note: Shows the distributions of three estimators (A, B, and C) that provide estimates of the unkown population parameter μ. $E(A) = \mu + 3$, $E(B) = \mu$, and $E(C) = \mu$.

(a) Which of the estimators above is unbiased?

Estimators B and C are unbiased because their expected values are equal to the unknown population parameter μ.

(b) Which of the estimators above has the smallest variance?

Of the three estimators, Estimator A has the smallest variance. Compared to those of the other estimators, the estimates from Estimator A are more tightly distributed around the estimator's expected value.

(c) Which of the estimators above is the "best" unbiased estimator?

The "best" unbiased estimator is the unbiased estimator with the least variance. Both Estimator B and Estimator C are unbiased, but Estimator C has a smaller variance. In this sense, Estimator C is the best unbiased estimator.

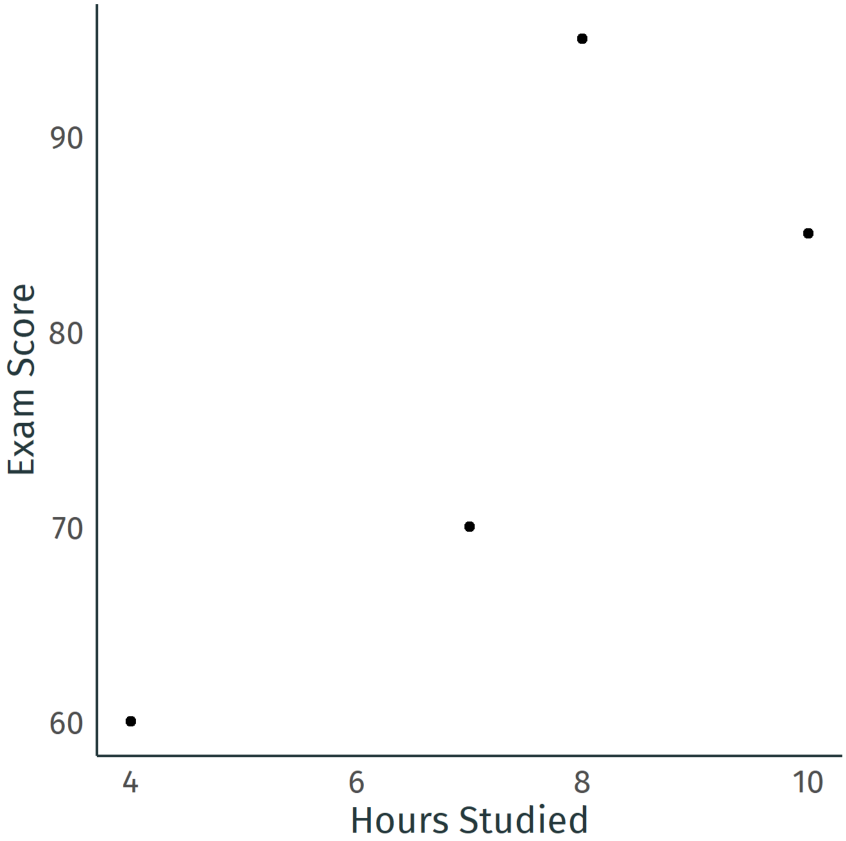
3. Use the rules of the expected values operator to **prove** that the population variance $E\left[(X - E(X))^2\right]$ is equivalent to $E(X^2) - [E(X)]^2$.

$$\begin{aligned} E[(X-E(X))^2] &= E[(X-E(X))(X-E(X))] \\ &= E[X^2 - 2XE(X) + E(X)^2] \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

4. Consider data gathered from students on the number of hours they slept the night before their midterm exam and their subsequent exam scores:

Student	**Hours Slept**	**Exam Score**
Ahmed	8	95%
Chen	10	85%
Klara	7	70%
Peter	4	60%

(a) In the space provided below, draw a scatter plot with **Exam Score** on the Y-axis and **Hours Slept** on the X-axis. Do they appear positively or negatively correlated?



They appear positively correlated.

(b) What is the sample mean of **Exam Score**?

The mean **exam score** is 77.5.

$$\begin{aligned}\overline{\text{Exam Score}} &= \frac{95+85+70+60}{4}\\ &= \frac{310}{4}\\ &= 77.5\end{aligned}$$

(c) What is the sample variance of **Hours Slept**?

The variance of **hours slept** is 6.25. To calculate the variance, you first need to calculate the sample mean of **hours slept**:

$$\begin{aligned}\overline{\text{Hours Slept}} &= \frac{8+10+7+4}{4}\\ &= \frac{29}{4}\\ &= 7.25.\end{aligned}$$

Then you calculate the sample variance:

$$\begin{aligned}\text{Var}(\text{Hours Slept}) &= \frac{(8-7.25)^2+(10-7.25)^2+(7-7.25)^2+(4-7.25)^2}{4-1}\\ &= \frac{0.5625+7.5625+0.0625+10.5625}{4-1}\\ &= \frac{18.75}{3}\\ &= 6.25.\end{aligned}$$

(d) What is the sample correlation coefficient between **Exam Score** and **Hours Slept**?

The correlation between **exam score** and **hours slept** is approximately 0.79. Start by calculating the sample covariance between **exam score** and **hours slept**:

$$\begin{aligned}\text{Cov}(\text{Score}, \text{Sleep}) &= \frac{(8-7.25)(95-77.5)+(10-7.25)(85-77.5)+(7-7.25)(70-77.5)+(4-7.25)(60-77.5)}{4-1}\\ &= \frac{13.125+20.625+1.875+56.875}{4-1}\\ &= \frac{92.5}{3}\\ &\approx 30.83.\end{aligned}$$

The correlation coefficient is given by

$$\text{Cor}(\text{Exam Score}, \text{Hours Slept}) = \frac{\text{Cov}(\text{Exam Score}, \text{Hours Slept})}{\sqrt{\text{Var}(\text{Exam Score})}\cdot\sqrt{\text{Var}(\text{Hours Slept})}}.$$

To find the standard deviations, you need the variances. Calculate the sample variance of **Exam Score**:

$$\begin{aligned}\text{Var}(\text{Exam Score}) &= \frac{(95-77.5)^2+(85-77.5)^2+(70-77.5)^2+(60-77.5)^2}{4-1}\\ &= \frac{725}{4-1}\\ &= \frac{725}{3}\\ &\approx 241.67.\end{aligned}$$

Finally, you can plug in what you know:

$$\begin{aligned}
\text{Cor}(\text{Exam Score}, \text{Hours Slept}) &= \frac{\text{Cov}(\text{Exam Score}, \text{Hours Slept})}{\sqrt{\text{Var}(\text{Exam Score})} \cdot \sqrt{\text{Var}(\text{Hours Slept})}} \\
&= \frac{30.83}{\sqrt{6.25} \cdot \sqrt{241.67}} \\
&= \frac{30.83}{2.5 \cdot 15.54574} \\
&\approx 0.79.
\end{aligned}$$

Computational Questions

For this portion of the problem set, you will use the file `election_2016.csv` in the `Problem_Set_1` folder on Canvas. The file contains county-level data on the 2016 presidential election from the MIT Election Data and Science Lab and the US Census Bureau. You will first need to download the data from Canvas and import it into `R`. Then you will need to load the `tidyverse` package.

Variable Name	Description
`stname`	Name of state
`fips`	County identifier (FIPS code)
`trump`	Number of votes cast for Donald Trump
`clinton`	Number of votes cast for Hillary Clinton
`totalvotes`	Total number of votes cast for any presidential candidate
`white`	Percentage of eligible voters who are non-Hispanic white
`poverty`	Percentage of residents below the poverty line
`population`	Number of residents

Use the data to complete the tasks and questions below.

1. Reduce your data set to include only those counties with a population of 50,000 or more.

2. Generate new variables that give the percentage of votes cast for each candidate using the `mutate` function. With these new variables, generate a third variable called `trump_margin` that gives Trump's "margin of victory" in each county. (For counties that Clinton won, the value of `trump_margin` will be negative.)

3. Produce summary statistics (min, max, mean, median, standard deviation, and the number of observations) for `trump_margin`, `white`, `poverty`, and `population` using the `summarize` function. Use your summary statistics to answer the following:

   (a) What is Trump's average vote margin?

   (b) What is the median poverty rate?

   (c) How many people live in the least populous county? How many live in the most populous county?

4. Create a histogram of Trump's margin of victory. Which candidate won more counties? Is it necessarily the case that this candidate won more votes nationally? *Hint*: Look at a histogram of county population, too.

5. Create a scatter plot with `trump_margin` on the Y-axis and `poverty` on the X-axis. Does support for Trump appear positively or negatively correlated with poverty rates? Use the `cor` function to calculate the correlation coefficient. Interpret your result.

6. Create a scatter plot with `trump_margin` on the Y-axis and `white` on the X-axis. Does support for Trump appear positively or negatively correlated with the percentage of white voters? Use the `cor` function to calculate the correlation coefficient. Interpret your result.

7. Using `group_by` and `summarize`, aggregate the county-level data on `white`, `poverty`, and `population` to the state level. The first two variables are rates. To obtain accurate state-level rates, you need to account for differences in population across counties. You can do this by using the `weighted.mean` function inside `summarize`. For population, you can use the `sum` function. Your new `tibble` should have 50 rows.

8. Using the state-level data, create a scatter plot with `poverty` on the Y-axis and `white` on the X-axis. To illustrate differences in population across states, vary the size of each point by population. You can this by including `size = population` in the `aes` function of `ggplot`. Describe any patterns you see.