EC 320: Introduction to Econometrics
Instructor: T. Ren
Problem Set 1
Summer 2022
Due: 11:59 p.m. on Saturday, June 25

Setup

For the analytical portion, your typed responses/answers to the question (in a Word file or something similar). Ideally, use LaTeX to fill out your answers. Otherwise, feel free print off this sheet and fill in the blanks. Answers must be submitted online through Canvas by the stated deadline (see above). For the computational portion, you may provide a separate document with your answers.

Analytical Questions: 95 total possible points

1) Let $\sum_{i=1}^{n} x_i = 30$. Determine the value of $\sum_{i=1}^{n} x_i^2$. (5 points)

2) If $\mathbb{E}(X) = 28$ and $Y = 32 + \frac{9}{5}X$, what is $\mathbb{E}(Y)$? (5 points)

3) Random variable X takes the value of 1 with probability 0.5 and value 2 with probability 0.5, what is the expectation of $E(X^2)$? (5 points)

4) Let X be the total when two dice are thrown.

Calculate the possible values of Y, where $Y(X)$ is given by

$$Y = 10X - 5$$

(10 points)

a) Calculate $\mathbb{E}(Y)$. Show that this is equal to $10\,\mathbb{E}(X) - 5$.

b) Provide the population variance and standard deviation of X as defined above (recall that the population parameter $\mu_x = E(X)$).

5) Let $\text{Var}(X) = 25$ and $Y = 32 + \frac{9}{5}X$, what is the standard dev of the random variable Y? (5 points)

6) Fill in the gaps:(5 points)

$$\text{Cov}(X, Y) \equiv \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

$$=$$

$$=$$

$$= \mathbb{E}(XY) - \mu_x\mu_y$$

7) Definition: If X and Y are statistically independent then, $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$. Evaluate the following statement : If $\text{Cov}(X, Y) = 0$ then X and Y are necessarily independent. Is this true or false? (5 points)
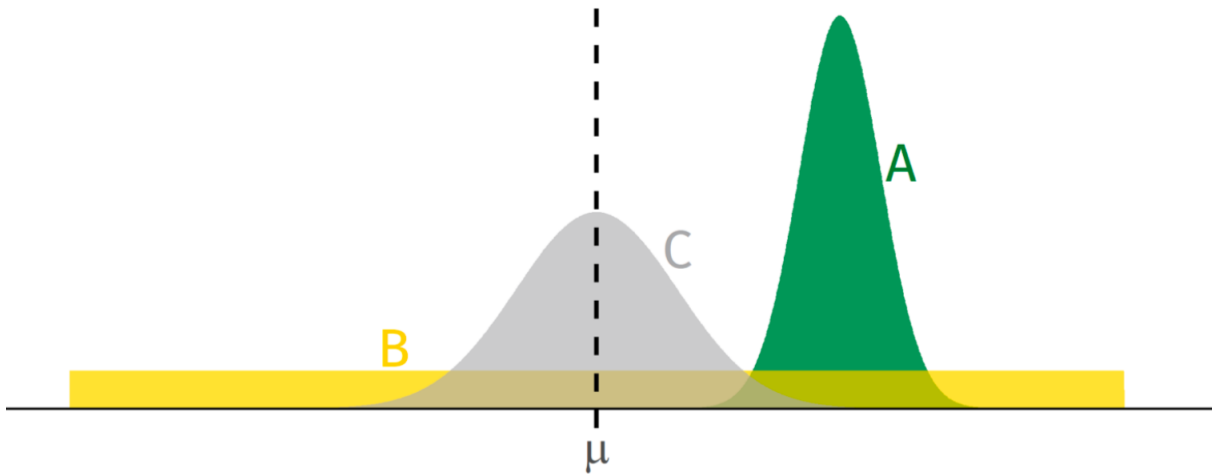
For the distributions of these three estimators, estimates are provided for the unknown parameter μ. $E(A) = \mu + 3$, $E(B) = \mu$ and $E(C) = \mu$. (15 total points, 5 points per question)

i) Which of these three estimators is unbiased?

ii) Which of these three estimators has the lowest variance?

iii) Among the estimators is the "best" unbiased estimator?

8) Suppose a survey of 50 households showed the mean household expenditure increased by 200 dollars after a new tax policy and the standard error of the increase was 75 dollars. (15 points)

a) Is the increase in expenditure significant at the 5% level? the 1% level? (Use a 2-sided test)

b) Calculate the 95% confidence interval and the 99% confidence interval

c) Explain why we reject/accept the null hypothesis.

9) Define the following types of data (10 points, 2.5 points per question)

i) Cross-sectional

ii) Time series

iii) Pooled cross sectional

iv) Panel data

10) Consider hypothetical data on the counterfactuals of 6 individuals (15 points)

i	$Treatment$	$Y_{1,i}$	$Y_{0,i}$
1	1	5	0
2	1	8	3
3	1	5	4
4	0	4	4
5	0	8	4
6	0	6	1

i) Determine the individual treatment effects. Are the treatment effects constant?

ii) Determine the average treatment effect, based on treatment assignment.

iii) Why is it impossible to observe data on counterfactual outcomes in real life?

iv) Estimate the average treatment effect by comparing the mean of the treatment group to the mean of the control group.

v) Do you think the estimator in iv is unbiased? Why or why not?

Computational Questions: 70 Total Possible Points

For this portion of the assignment you will need to use the hw1.csv file in the Problem Set 1 folder on Canvas. The file contains county-level data on the 2020 presidential election from the MIT Data Science Lab. You will first need to download the file from Canvas and import it into R. Then, load the tidyverse package. Use the data to complete the tasks and answer the questions below.

Variable Name	Description
fips	County identifier (FIPS Code)
state	Name of state
county	Name of county
candidate	Name of the presidential candidate
party	The part of the candidate
candidatevotes	Total number of votes cast for each candidate
total votes	The Total number of votes cast for all candidates

1. (10 points) Generate new variables that give the percentage of votes cast for each candidate.

2. (10 points) Generate a new data frame in which you generate a variable for Biden's votes, Trump's votes, and Other votes for each county in each State. Make sure you keep the totalvotes variable in this data frame.

3. (10 points) With these new variables, generate a third variable called **trumpmargin** that gives Trump's "margin of victory" in each county.

4. (10 points) Produce summary statistics (min, max, mean, median, standard deviation, and the number of observations) for trumpmargin. What is his average vote margin?

5. (10 points) 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?

6. (10 points) Now, repeat problem 4-5 but do it by state, not by county.

7. (10 points) Create a graph with trump margin on the Y-axis and state on the X-axis.

Which state did Trump have the greatest margin? Which state did he have the smallest

margin?