Math Camp 2025 – Problem Set 10

Problem 1

Suppose that X and Y are independently distributed random variables with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Note that μ_X represents the expectation of X, σ_X represents the standard deviation, and σ_X^2 the variance.

Calculate the following quantities:

(1) $\mathbb{E}[5X - 2Y + 8]$

(2) $\text{Var}[X + 3Y]$

(3) $\mathbb{E}[4X^2 - 10XY + 25Y^2]$

Problem 2

X and Y are discrete random variables with the following joint distribution:

	X			
Y	1	2	3	4
1	.10	.07	.03	.01
2	.08	.13	.04	.02
3	.03	.04	.11	.09
4	.02	.03	.12	.08

Answer the following questions:

(a) Calculate $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{E}[XY]$.

(b) Calculate $\text{Var}[X]$ and $\text{Var}[Y]$.

(c) Prove that the following two expressions of the covariance between X and Y are equivalent:

$$\text{cov}(X, Y) = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big]$$
$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

(d) Calculate $\text{cov}(X, Y)$.

(e) Calculate the correlation between X and Y: ρ_{XY} with the following formula:

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

(f) Write out the probability mass function (PMF) of the random variable $Z = \mathbb{E}[Y \mid X]$.

Comment. This means: find the possible values that $Z = \mathbb{E}[Y \mid X]$ can take, and calculate the probability that it takes each of those values.

Problem 3

Imagine that every person in the United States has a fixed preference for redistribution, which is defined as a continuous variable, where smaller values mean that the individual favors less redistribution.

Let N equals the size of the entire population of the US. Let Y_i be the preference for redistribution of person $i \in \{1, \ldots, N\}$, and Y be preference for redistribution of a person chosen uniformly at random. Suppose that the average preference for redistribution of the entire population is μ and the variance is σ^2.

As political scientists, we would like to make inferences about the aggregated preference for redistribution of the entire population, but we can't go out and measure every single person's view. So instead, we are going to sample n people from the entire population at random (but not necessarily with the same probability), and measure their preferences for each person in our sample (imagine for now that every person who is sampled responds, and that we measure their views correctly each time).

Let S_i be an indicator variable for person i being sampled, i.e., $S_i = 1$ if i is in the sample and $S_i = 0$ if not. Let S be an indicator variable for a person chosen uniformly at random being sampled.

Let $\hat{\mu}$ be the average preference for redistribution in our sample. The "hat" notation indicates that $\hat{\mu}$ is intended to estimate μ using only our sample.

Now answer the following questions:

(a) Consider Y_i, S_i, μ, and $\hat{\mu}$. Which of these are random variables? Which, if any, are not? Please make sure to explain your answer.

(b) Conceptually, what does $\mathbb{E}[Y]$ refer to?

(c) Conceptually, what do $\mathbb{E}[\hat{\mu}]$ and $\text{Var}[\hat{\mu}]$ refer to?

(d) Calculate $\mathbb{E}[S]$ and $\text{Var}[S]$.

(e) Write $\hat{\mu}$ as an expression of n, N, S_i, and Y_i.

(f) Write $\mathbb{E}[\hat{\mu}]$ as an expression of n, N, S and Y. What is $\mathbb{E}[\hat{\mu}]$ if S and Y are independent?

Hint. If I is selected uniformly at random from $\{1, \ldots, N\}$ and $X_1, \ldots, X_N$ are numbers then $\mathbb{E}[X_I] = \frac{1}{N}\sum_{i=1}^{N} X_i$. Notice that $Y = Y_I$ and $S = S_I$.

(g) If S and Y were not independent, would we necessarily get the same result as in (f)? Practically, what does this mean for surveys? Could you provide an example where S and Y might not be independent?

Problem 4

You are preparing to run a field experiment on the effectiveness of oversight in curbing corruption, as in Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal*

of Political Economy, vol. 115, no.2.

As part of a major infrastructure project, the Indonesian government has allocated funds for roads in an "infinite" (we assume for the purpose of this question) number of villages. Your experimental intervention is whether you inform the village that: "after funds had been awarded but before construction began, the project would subsequently be audited by the central government audit agency."

After the road is constructed, you will conduct an extensive (and expensive) data-collection effort to estimate the amount actually spent on each village's roads, based on the quality of materials found in excavated core samples, estimated wages based on interviews, and so on. The discrepancy between allocated funds and estimated expenditures will be your outcome variable.

Let T_i be an indicator variable for the treatment, so $T_i = 1$ refers to village i has been "treated." Let Y_i be the outcome you measured as described above, then the observed outcome is a function of the treatment variable, i.e. $Y_i = Y_i(T_i)$. Each village i has two potential outcomes:

- $Y_i(0)$: the outcome that would be observed if village i is assigned to the "control," or non-informed, group.
- $Y_i(1)$: the outcome that would be observed if village i is assigned to the "treatment," or informed, group.

Because the field experiment is such a large undertaking, you want to carefully design your experiment to ensure success. At the same time, you can only afford to measure 100 villages. Suppose that you randomly sample N_T villages to assign to treatment and N_C to control (so $N_T + N_C = 100$). Now answer the following questions:

(a) Interpret the meaning of $Y_i(1) - Y_i(0)$ for a single village i.

(b) Interpret the meaning of $\mathbb{E}[Y_i(1) - Y_i(0)]$, where i now is a random village.

(c) Interpret the meaning of $\mathbb{E}[Y_i(0)]$ and $\mathbb{E}[Y_i(1)]$, where i denotes a random village. As an expert on local bureaucracy and corruption, what kind of relationship do you expect between these two quantities?

(d) Interpret the meaning of $\frac{1}{100}\sum_{i\in S}\big(Y_i(1) - Y_i(0)\big)$, where S denotes the set of villages in your sample of 100 villages. Compare it to the quantity in (b).

(e) Interpret the meaning of $\text{Var}[Y_i(0)]$ and $\text{Var}[Y_i(1)]$, where again i denotes a random village. Again, as an expert, what kind of relationship do you expect between these two quantities?

(f) Assume that the treatment assignment is random, that is, $\Pr(T_i = 1) = \frac{N_T}{100}$ for each $i \in S$. For simplicity, assume that villages $i = 1$ through $i = 100$ are in our sample, or equivalently, $S = \{1, 2, \ldots, 100\}$. Calculate the expectation of the following expression and interpret your result:

$$\hat{\tau} = \frac{1}{N_T}\sum_{i=1}^{100} T_i Y_i - \frac{1}{N_C}\sum_{i=1}^{100}(1 - T_i)Y_i.$$