Homework 7

Due November 14 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Short questions) Justify all your answers mathematically.

   (a) For any random variable $\tilde{a}$, can $\mathrm{E}^2(\tilde{a})$ be smaller than $\mathrm{E}\left(\tilde{a}^2\right)$?

   (b) If $\tilde{a}$ and $\tilde{b}$ have the same distribution and are independent, is it true that $\mathrm{E}(\tilde{a}\tilde{b}) = \mathrm{E}^2\left(\tilde{a}\right)$?

   (c) A teacher of a class of n children asks their parents to leave a present under the Christmas tree in the classroom. The day after, each child picks a present at random. What is the expected number of children that end up getting the present bought by their own parents? (Hint: Define a random variable I_i that is equal to one when kid i gets the present bought by their own parents, and to zero otherwise.)

2. (Computer) We model the time (in years) until a computer breaks down as a random variable $\tilde{t}$. The time depends on whether the computer has a defect or not, which is modeled by a random variable $\tilde{d}$. If the computer has a defect ($\tilde{d} = 1$), then $\tilde{t}$ is an exponential with parameter 2. If it does not ($\tilde{d} = 0$), then $\tilde{t}$ is an exponential with parameter 1. The probability that the computer has a defect is 0.1.

   (a) Is the conditional expectation of $\tilde{t}$ given $\tilde{d}$ a discrete or continuous random variable? What is its pmf or pdf?

   (b) What is the variance of $\tilde{t}$?

   (c) A company buys 100 of these computers. If the time until they break down is distributed as explained above, and they are all independent, what is the mean and variance of the number of computers that break down during the first year?

3. (Law of conditional variance) In this problem we define the conditional variance in a similar way to the conditional expectation.

   (a) What is the object $\mathrm{Var}(\tilde{b} \,|\, \tilde{a} = a)$ (i.e. is it a number, a random variable or a function)? What does it represent?

(b) Setting $h(a) = \mathrm{Var}(\tilde{b} \,|\, \tilde{a} = a)$ we define the conditional variance as $\mathrm{Var}(\tilde{b} \,|\, \tilde{a}) = h(\tilde{a})$. What is this object?

(c) Prove the law of conditional variance:

$$\mathrm{Var}(\tilde{b}) = \mathrm{E}\left(\mathrm{Var}(\tilde{b} \,|\, \tilde{a})\right) + \mathrm{Var}\left(\mathrm{E}(\tilde{b} \,|\, \tilde{a})\right) \tag{1}$$

and describe it in words.

4. (Water salinity and temperature) In this question, we use oceanographic data to study the relationship between salinity and temperature in sea water. We perform our analysis on a cleaned and subsampled version of the data, `bottle.csv`. The script is available at `https://github.com/cfgranda/prob_stats_for_data_science/blob/main/hw7/conditional_expectation_EXERCISE.ipynb`

   (a) Plot an estimate of the conditional mean of salinity given the temperature along with the scatter plot of data. Justify any choices you make.

   (b) Annotate your plot to incorporate the conditional standard deviation of salinity given the temperature (apart from plotting the conditional mean, plot the conditional mean $\pm$ the conditional standard deviation).

   (c) Do you expect your estimates to be equally reliable at every point? Please explain your reasoning. (We are not looking for a mathematical answer, you can just explain intuitively.)