Homework 5

Due October 11 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 the median of $\tilde{a}$ equals m, what is the median of $\tilde{a} + b$?

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

   (d) 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. (Pasta and rice) You are hired by the management of a restaurant to model its stock probabilistically. You talk to the cook and she says:

   We cook pasta and rice. We always make sure that we have at least 100 lb of pasta or 100 lb of rice (if there is at least 100 lb of pasta, for example, we could have no rice at all); the logic being that we have daily specials and we want to be able to feed a lot of people with the same dish. However we never have more than 300 lb of rice or of pasta because we have no space to store it (we are able to store 300 lb of rice and 300 lb of pasta at the same time).

   You decide to model the quantity of pasta as a random variable $\tilde{x}$ and the quantity of rice as a random variable $\tilde{r}$. As you have no information beyond what you have heard, you assume that their joint pdf is constant (within the restrictions that you deduce from talking to the cook).

   (a) Draw the joint pdf of $\tilde{x}$ and $\tilde{r}$.

   (b) Are $\tilde{x}$ and $\tilde{r}$ uncorrelated? Justify your answer.

(c) Are $\tilde{x}$ and $\tilde{r}$ independent? Justify your answer.

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.

   (d) We model the time at which a runner gets injured (in hours) during a marathon as an exponential random variable with parameter equal to 1 if the runner is under 30 years old and 2 if she is over 30. What is the mean and the standard deviation of the time at which a runner gets injured if 20% of the runners are over 30?

4. (Water salinity and temperature) A quick Google search will tell you that the salinity of water, which is the salt content in water, increases with temperature. This is because water expands at larger temperature and can fit in more molecules, including salt, increasing the salinity (according to Sciencing). In this question, we will use oceanographic data to understand the relationship between salinity and temperature. We will perform our analysis on a cleaned and subsampled version of the data, `bottle.csv`. Please refer to the Kaggle website for any details about the data.

   (a) Find the best linear MMSE estimator of salinity with temperature . Plot the line and the scatter plot of data on the same graph. According to the relationship you uncovered here, does water salinity increase with temperature?

   (b) Plot an estimate of the conditional mean of salinity given the temperature along with the scatter plot of data. What trend do you see from this plot? (Hint: this question closely follows example 5.2)

   (c) Are your conditional mean estimates equally reliable at every point? If not, which estimates are more reliable? Please explain your reasoning. (We are not looking for a mathematical answer, you can reason in words)

   (d) (Not graded for points) Why do you think the trend you find is different from what Sciencing suggests? It is not because of limited data - the full dataset has $810k$ data points and we still observe the same trend.