Queen Mary
University of London

January Examination Period 2024

Duration: 2 hours (+1 for uploads)

ECS764P Main Applied Statistics

This paper requires **two hours work**. There is an extra hour allowance for downloading the paper and uploading your answers.

You MUST submit your answers before the exam end time.

All instructions and guidelines from the exam page should be followed.

This is an open-book exam and you may refer to lecture material, text books and online resources. The usual referencing and plagiarism rules apply, and you must clearly cite any reference used.

Calculators and/or Jupyter Notebooks are required in this examination.

Answer ALL questions

You MUST adhere to the word limits, where specified in the questions. Failure to do so will lead to those answers not being marked.

YOU MUST COMPLETE THE EXAM ON YOUR OWN, WITHOUT CONSULTING OTHERS.

Examiners:
Dr. Fredrik Dahlqvist

Question 1

Consider the density function given by

$$f(x) = \begin{cases} (ax-2)^2 & \text{if } x \in [0,1] \\ 0 & \text{else} \end{cases}$$

where a is some constant real number.

(a) Write the condition that f must satisfy in order to be a *probability* density function

[2 marks]

(b) Find the value of a for which this condition holds. Show your calculations.

[4 marks]

(c) Let $\mathbb{P}$ be the probability measure defined by this density. What is the probability mass of the interval $\left[-1, \frac{1}{2}\right]$ under $\mathbb{P}$? Show your calculations.

[6 marks]

(d) Compute the CDF of the distribution above. Plot the CDF (you can use a Jupyter notebook or do it by hand). Based on this plot, give an approximate value for the median m of the distribution. The CDF has an almost flat section, why?

[6 marks]

(e) Compute the mean of the distribution (show your calculations!). Sketch a plot of the probability density function. Based on this plot, what is/are the mode/modes of this distribution? Display the mean, the median and the mode of the distribution on your plot. Which measure of centrality would you *not* choose? Explain briefly why. Based on these measures of centrality, what is the sign of the skewness? Explain briefly why.

[7 marks]

Question 2

(a) Consider the measure $\mathbb{P}$ with density

$$\mathbb{P}\left(\{-1\}\right) = \frac{1}{4} \qquad \mathbb{P}\left(\{1\}\right) = \frac{3}{4}$$

(i) What is the support of $\mathbb{P} + \mathbb{P} + \mathbb{P}$?

(ii) Compute the PMF of $\mathbb{P} + \mathbb{P} + \mathbb{P}$ from first principles (i.e. from the definitions, justifying each step).

(iii) Consider the following random walk on the real line: at each step we sample from $\mathbb{P}$ and add the value of this sample to our position. So for example, if we are at 0 and we sample -1, our next position is $0 + (-1) = -1$. What is the mean of the position after three step if we started at 0?

[8 marks]

(b) Using the same distribution $\mathbb{P}$ as in the previous question.

(i) What is the support of the distribution $\frac{\mathbb{P}+\mathbb{P}+\mathbb{P}}{3}$?

(ii) Compute the PMF of $\frac{\mathbb{P}+\mathbb{P}+\mathbb{P}}{3}$ (*Hint: Use your results from the previous section*).

(iii) Using the random walk picture of the previous question, $\frac{\mathbb{P}+\mathbb{P}+\mathbb{P}}{3}$ can be seen as the average step size of a walk with three steps. Using the Central Limit Theorem, approximate the probability that the average step size of a walk with $n = 25$ steps is negative. You will need a scientific calculator or a Jupyter Notebook to compute this number.

[7 marks]

(c) Let $\mathbb{P}$ be a probability measure whose mean $\mu(\mathbb{P}) < \infty$ exists and let $\overline{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}$ be the distribution of the length-n sample mean. Which of the following statement(s) follow(s) from the weak law of large numbers (there may be more than one)?

(i) There exists an $N \in \mathbb{N}$ such that for all $n > N$,

$$\overline{\mathbb{P}}_n\left([\mu(\mathbb{P}) - 0.001, \mu(\mathbb{P}) + 0.001]\right) > 0.99$$

(ii) There exists an $N \in \mathbb{N}$ such that for all $n > N$,

$$\overline{\mathbb{P}}_n\left([\mu(\mathbb{P}) - 0.001, \mu(\mathbb{P}) + 0.001]\right) = 1$$

(iii) There exists an $N \in \mathbb{N}$ such that

$$\lim_{n\to\infty} \overline{\mathbb{P}}_N\left([\mu(\mathbb{P}) - 0.001, \mu(\mathbb{P}) + 0.001)\right) = 0.99$$

(iv) There exists an $N \in \mathbb{N}$ such that for all $n > N$,

$$\overline{\mathbb{P}}_n\left(\mu(\mathbb{P}))\right) > \varepsilon$$

(v) There exists an $N \in \mathbb{N}$ such that ,

$$\overline{\mathbb{P}}_N\left([\mu(\mathbb{P}) - 0.1, \mu(\mathbb{P}) + 0.1]\right) > 0.87$$

(vi) There exists an $N \in \mathbb{N}$ such that ,

$$\overline{\mathbb{P}}_N\left([\mu(\mathbb{P}) - 0.1, \mu(\mathbb{P}) + 0.1]\right) < 0.87$$

[6 marks]

(d) Cite one probability measure for which the weak LLN does not apply and briefly explain why it doesn't.

[4 marks]

Question 3

Consider the family of probability measures $\mathbb{P}_\nu$ indexed by a positive real parameter $\nu > 0$ and defined by the following PDF

$$f_\nu(x) = \begin{cases} 2\nu x e^{-\nu x^2} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

Assume that we have n observations $(x_1, \ldots, x_n)$ that are sampled from one of the probability measures in this family.

(a) Briefly explain how the MLE $\hat{\nu}$ is computed: what needs to be maximised, what is the difficulty, what is the trick used to solve this difficulty, how is $\hat{\nu}$ computed.

[8 marks]

(b) Compute the MLE $\hat{\nu}$.

[8 marks]

(c) The *second raw moment* μ'_2 of a continuous probability measure with density f is defined as

$$\mu'_2 = \int x^2 f(x)\, dx$$

Using the law of the unconscious statistician, show that $\frac{1}{\hat{\nu}}$ is an unbiased estimator of the second raw moment of the probability measure $\mathbb{P}_\nu$.

[9 marks]

Question 4

(a) Consider the following two arrays:

$$x = [1, 3, 3, 4, 7, 7, 9, 9, 10, 12, 15, 16]$$
$$y = [17, 12, 8, 10, -2, -6, -9, -5, -9, -19, -23, -28]$$

You will first perform a linear regression of x against y.

(i) Using the formula from the lecture and a notebook/calculator, compute the coefficient β_1. Show some of the details of your calculations.

(ii) Using the formula from the lecture and a notebook/calculator, compute the coefficient β_0. Show the details your calculations.

(iii) What are the geometric interpretations of β_0. and β_1?

(iv) What does β_1 say about the correlation coefficient ρ_{xy}? Justify your answer.

[8 marks]

(b) Continuing from the previous question, we now make the further assumption that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{where } \varepsilon_i \sim \text{Normal}(0, \sigma), 1 \leq i \leq 12$$

for some $\sigma > 0$.

(i) Test the hypothesis $H_0 : \beta_1 = -3.0$ at confidence level $\alpha = 0.95$.

(ii) Test the hypothesis $H_0 : \beta_0 = 20.0$ at confidence level $\alpha = 0.99$.

(iii) Does the model above define a joint distribution on $\mathbb{R} \times \mathbb{R}$? Briefly justify your answer.

[10 marks]

(c) Consider the following density

$$f(x, y) = \mathbb{1}_{[0,1]}(x)\mathbb{1}_{[x-1/2, x+1/2]}(y).$$

It defines a joint probability distribution $\mathbb{P}$ on $\mathbb{R} \times \mathbb{R}$.

(i) Compute (the density of) the marginal $(\pi_1)_*(\mathbb{P})$ along the x-axis.

(ii) Compute (the density of) the marginal $(\pi_2)_*(\mathbb{P})$ along the y-axis.

(iii) Compute the mean of both marginals and, using these values, write the expression for the covariance of $\mathbb{P}$ (you don't need to evaluate it, just write it).

[7 marks]

End of questions