Imperial College London

MATH40005

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2020

This paper is also taken for the relevant examination for the
Associateship of the Royal College of Science

Probability and Statistics

Date: 14th May 2020

Time: 09.00am – 12.00 noon (BST)

Time Allowed: 3 Hours

Upload Time Allowed: 30 Minutes

This paper has 6 Questions.

Candidates should start their solutions to each question on a new sheet of paper.

Each sheet of paper should have your CID, Question Number and Page Number on the top.

Only use 1 side of the paper.

Allow margins for marking.

Any required additional material(s) will be provided.

Credit will be given for all questions attempted.

Each question carries equal weight.

SUBMIT YOUR ANSWERS AS SEPARATE PDFs TO THE RELEVANT DROPBOXES ON BLACKBOARD (ONE FOR EACH QUESTION) WITH COMPLETED COVERSHEETS WITH YOUR CID NUMBER, QUESTION NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.

Throughout the exam, we assume that $(\Omega, \mathcal{F}, \mathrm{P})$ denotes a probability space.

Please remember to justify all your answers and state carefully which results from the lectures you apply in your proofs.

1. (a) Define a σ-algebra. (3 marks)

   (b) Define a probability measure on $(\Omega, \mathcal{F})$. (3 marks)

   (c) A diagnostic test has a probability 0.9 of giving a positive result when applied to a person suffering from a certain disease, and a probability 0.2 of giving a (false) positive when applied to a non-sufferer. It is estimated that 10 % of the population are sufferers. Suppose that the test is now administered to a person about whom we have no relevant information relating to the disease (apart from the fact that he/she comes from this population). Calculate the following probabilities:

      (i) that the test result will be positive; (3 marks)

      (ii) that, given a positive result, the person is a sufferer; (3 marks)

   (d) How many possibilities are there to write the number 7 as an ordered sum of 3 positive integers? [E.g. 7=1+3+3 would be one possible case and 7=3+1+3 would be another case.] (4 marks)

   (e) Let $k, n \in \mathbb{N} = \{1, 2, \ldots\}$. How many possibilities are there to write the number k as an ordered sum of n positive integers? (4 marks)

(Total: 20 marks)

2. (a) Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space with $\Omega = \{1, 2, \ldots, 10\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (the power σ-algebra of Ω). Let $X : \Omega \to \mathbb{R}$ with $X(\omega) = \omega + 5$. Prove that X is a discrete random variable. (3 marks)

(b) Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space with $\Omega = \{1, 2, \ldots, 100\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (the power σ-algebra of Ω). Consider a discrete random variable X on this probability space with probability mass function given by $\mathrm{P}(X = 3) = \frac{1}{2}, \mathrm{P}(X = 5) = \frac{1}{5}, \mathrm{P}(X = 100) = \frac{3}{10}, \mathrm{P}(X = x) = 0$ for $x \notin \{3, 5, 100\}$. Find the cumulative distribution function F_X of X. (4 marks)

(c) Let X and Y denote independent geometric random variables with parameters p_1 and p_2, respectively, where $p_1, p_2 \in (0, 1)$. (Please refer to the hint below for the probability mass function of a geometric random variable.)

(i) Derive the cumulative distribution functions of X and Y. (3 marks)

(ii) Show that $Z = \min\{X, Y\}$ follows a geometric distribution and find the corresponding parameter. (4 marks)

(d) Imagine you toss a fair coin repeatedly. You denote by H the outcome Heads and by T the outcome Tails. How many times, on average, do you need to toss the coin to see the pattern HT (i.e. Heads followed by Tails) for the first time? (6 marks)

(Total: 20 marks)

Hint: If X is **geometrically** distributed with parameter $p \in (0, 1)$, then its probability mass function p_X is given by

$$p_X(x) = \begin{cases} (1-p)^{x-1}p, & \text{for } x = 1, 2, \ldots, \\ 0, & \text{otherwise.} \end{cases}$$

3. (a) Which properties does a function $f : \mathbb{R} \to \mathbb{R}$ need to satisfy in order to be a valid probability density function? (2 marks)

(b) For each of the functions $f(x)$ given below determine whether $f(x)$ is a valid probability density function (p.d.f.). If $f(x)$ is not a valid p.d.f., determine if there exists a constant c such that $cf(x)$ is a valid p.d.f.. Note that in each case, $f(x) = 0$ for all x not in the interval(s) specified.

(i) $f(x) = 3x$ for $0 < x < 1$, (2 marks)

(ii) $f(x) = -1$ for $0 < x < 1$, (2 marks)

(iii) $f(x) = 1$ for $0 < x < 1$ and $f(x) = -1$ for $1 < x < 2$. (2 marks)

(c) Consider three jointly continuous random variables X, Y, Z with joint probability density function given by

$$f_{X,Y,Z}(x, y, z) = \begin{cases} c, & \text{for } 0 < x < y < z < 1, \\ 0, & \text{otherwise,} \end{cases}$$

for a constant $c \in \mathbb{R}$.

(i) Show that $c = 6$. (3 marks)

(ii) Find $\mathrm{E}(XYZ)$. (3 marks)

(d) (i) Define a partition of the sample space Ω and give an example of $(\Omega, \mathcal{F})$ with a valid partition. (3 marks)

(ii) Prove the law of the total expectation for a discrete random variable X. I.e. consider a partition $\{B_i : i \in \mathcal{I}\}$ of Ω with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$. Let X denote a discrete random variable with finite expectation. Show that

$$\mathrm{E}(X) = \sum_{i \in \mathcal{I}} \mathrm{E}(X|B_i)\mathrm{P}(B_i),$$

whenever the sum converges absolutely. (3 marks)

(Total: 20 marks)

4. (a) Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and each follows a normal distribution with mean μ and variance σ^2. We define the following estimators

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad Z = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

where $\overline{X}$ is an estimator of μ, and S^2 and Z are estimators of σ^2. Carefully justifying all your steps and stating any results used:

(i) Find the distribution of $\overline{X}$. (3 marks)

(ii) Given that $\mathrm{E}(S^2) = \sigma^2$, compute $\mathrm{E}(Z)$. (1 mark)

(iii) Given that $\mathrm{Var}(S^2) = \dfrac{2\sigma^4}{n-1}$, compute $\mathrm{Var}(Z)$. (1 mark)

(iv) Compute the bias of Z. (1 mark)

(v) Compute the mean squared error of Z. (2 marks)

(vi) Choose a constant b so that the quantity bZ has a chi-squared distribution, and state the degrees of freedom for this chi-squared distribution. (2 marks)

(vii) Compute $\mathrm{Cov}(\overline{X}, Z)$. (2 marks)

(b) Markov's inequality states that if a random variable X can only take nonnegative values, then

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}, \qquad \text{for all } a > 0.$$

Prove Markov's inequality. (4 marks)

(c) Suppose that a medical research lab is testing for the association of different genetic variants with a particular disease. The research team decides in advance that a significance threshold of $\alpha = 0.01$ will be used for each test. A total of 100 genetic variants are tested for association based on the data the team has available. The following table lists the five smallest p-values (in decreasing order) and the genetic variants for which these p-values were found:

Genetic variant	A	B	C	D	E
p-value from test	3×10^{-2}	9×10^{-3}	4×10^{-4}	2×10^{-5}	5×10^{-6}

Which of the genetic variants in the table (if any) should the research team declare to be significantly associated with the disease given the data, the statistical test and the significance threshold that were used? Provide justification and state any results used. (4 marks)

(Total: 20 marks)

5. (a) Given a sample of real-valued observations $x_1, x_2, \ldots, x_n$, prove that for any constant $a \in \mathbb{R}$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - a)^2,$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of the observations. (4 marks)

(b) Given n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ for quantities X and Y, define the sample means $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and define

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2, \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Consider the model given by

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i \in \{1, 2, \ldots, n\},$$

where the e_i, $i \in \{1, 2, \ldots, n\}$, are unobservable errors. Find the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the parameters β_0 and β_1, respectively, such that

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \min_{b_0, b_1} \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2.$$

(7 marks)

(c) Suppose we have two samples of data, independent observations $x_1, x_2, \ldots, x_n$ of the random variable X and independent observations $y_1, y_2, \ldots, y_m$ of the random variable Y. We wish to use the two-sample t-test to decide whether or not $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$ are equal.

(i) What is the null hypothesis for the t-test in this case? (1 mark)

(ii) What assumptions are required in order to have theoretical justification for conducting the t-test in this case? (2 marks)

(d) Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed according to a uniform distribution on the closed interval $[0, \theta]$, for some parameter $\theta > 0$, where the exact value of the parameter θ is unknown. Given that $\mathbf{X} = (X_1, X_2, \ldots X_n)$ is observed as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, find the maximum likelihood estimator of θ. Provide justification for all of your steps. (6 marks)

(Total: 20 marks)

Hint: If X is **uniformly** distributed on the interval $[a, b]$, with $a < b$, then its probability density function f_X is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in [a, b], \\ 0, & \text{otherwise.} \end{cases}$$

Note that this question is split over two pages. Please turn the page to see the rest of Question 6.

6. (a) Suppose that the random variables $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed according to a distribution F_Y, which has an unknown mean μ that we wish to estimate. Suppose that $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ is observed as $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, and we are given that

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 5, \qquad \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = 4, \qquad n = 10.$$

Noting that you have access to Tables 1 and 2 below:

(i) If we can assume that the random variables $Y_1, Y_2, \ldots, Y_n$ are normally distributed with variance $\mathrm{Var}(Y) = \sigma^2 = 9$, construct a 90% confidence interval for the unknown mean μ based on the data $\mathbf{y}$. (2 marks)

(ii) If we can assume that the random variables $Y_1, Y_2, \ldots, Y_n$ are normally distributed but the variance $\mathrm{Var}(Y) = \sigma^2$ is unknown, construct a 95% confidence interval for the unknown mean μ based on the data $\mathbf{y}$. (2 marks)

(iii) If we cannot assume that the random variables $Y_1, Y_2, \ldots, Y_n$ are normally distributed but we can assume that the variance is $\mathrm{Var}(Y) = \sigma^2 = 16$, construct a confidence interval for the unknown mean $\mu = \mathrm{E}(Y)$ which has coverage probability at least 0.99, whatever the distribution of F_Y. (2 marks)

Table 1: Partial table showing values of t for $\mathrm{P}(T < t)$, where T has Student's t-distribution with ν degrees of freedom

ν	0.90	0.95	0.975	0.99
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764

Table 2: Partial table showing values of z for $\mathrm{P}(Z < z)$, where Z has a standard normal distribution

z	$\mathrm{P}(Z < z)$
1.281	0.900
1.645	0.950
1.960	0.975
2.326	0.990

[IMPORTANT: Question 6 continues on the next page.]

[**Question 6 continues on this page**]

(b) Suppose one fits a simple linear regression model to the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ as

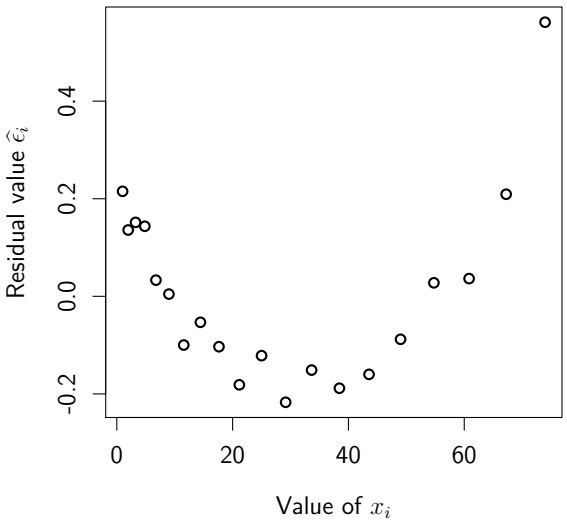
$$Y_i = \beta_0 + \beta_1 g(x_i) + \epsilon_i, \qquad i \in \{1, 2, \ldots, n\},$$

where $g : \mathbb{R} \to \mathbb{R}$ is some univariate transformation and $n = 20$.

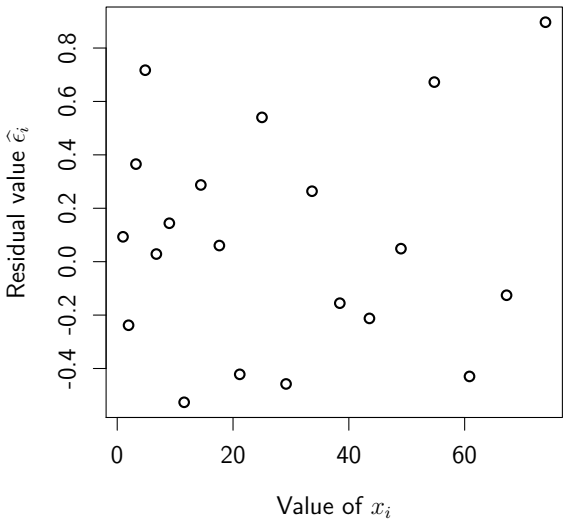
(i) What joint distribution are the errors ϵ_i assumed to follow? (1 mark)

(ii) For two different choices of transformation g, one has two models with the fitted residuals $\widehat{\epsilon}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 g(x_i)$ shown in the figures below. For each model, state whether the model fits the data well or not and justify your answer. (4 marks)

Model 1



Model 2



(c) Suppose that the random variables $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed as an exponential distribution with unknown parameter θ, which has probability density function

$$f(z) = \theta \exp(-\theta z), \qquad \text{with support } z > 0.$$

Following a Bayesian approach and assuming that θ is a random variable with a $\Gamma(\alpha, \beta)$ prior which has probability density function

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta), \qquad \text{with support } \theta > 0,$$

and given that $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ is observed as $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, find the posterior distribution of θ given $\mathbf{z}$ and give the name of this distribution. (3 marks)

(d) Suppose that a random variable X has mean $\mathrm{E}(X) = 2$, another random variable Y has mean $\mathrm{E}(Y) = 3$, and it is known that $\mathrm{E}(XY) = 4$. It is also known that $2 \leq Y \leq 5$. Find a nontrivial lower bound on the standard deviation of X. (6 marks)

(Total: 20 marks)