Queen Mary
University of London

Main Examination period 2023 – January – Semester A

MTH794P: Probability & Statistics for Data Analytics

Duration: 3 hours

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

The exam is intended to be completed within **3 hours**. However, you will have a period of **4 hours** to complete the exam and submit your solutions.

You should attempt ALL questions. Marks available are shown next to the questions.

The New Cambridge Statistical Tables are provided.

You are allowed to bring **three A4 sheets of paper** as notes for the exam.

Only approved non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used. **Statistical functions** provided by the calculator may be used provided that you state clearly where you have used them.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: I. Goldsheid, A. Gnedin

Question 1 [12 marks]. An urn contains 3 black and 5 white balls. Consider the following game:
1. A player picks a ball at random and replaces it by a ball of the opposite colour.
2. Then the player repeats this action one more time.

The game is won if the content of the urn remains unchanged.

(a) What is the probability of picking a black ball at step 2? [**4**]

(b) What is the probability that the payer wins the game? [**5**]

(c) Suppose now that the player pays £1 for the right to play this game. If the player wins the game then he gets back £1 and in addition is given £x. If he loses the game then he also loses his pound. For what value of x is the game fair?

Remark. In a fair game, the mean value of a player's gain is equal to zero. [**3**]

Question 2 [26 marks]. Two random variables X and Y have a joint probability density function $f_{X,Y}(x,y)$.

(a) (i) State the definition of the probability density function $f_{X,Y}(x,y)$ of two jointly continuous random variables (X,Y). [**3**]

(ii) Prove that the probability density function $f_X(x)$ of the random variable X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y.$$

[**6**]

(b) Suppose now that

$$f_{X,Y}(x,y) = \begin{cases} 3\mathrm{e}^{-2x-y} & \text{if } y > x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find $P(Y > 2X)$. [**5**]

(ii) Find the marginal probability density function f_X of X. [**4**]

(iii) Find the conditional density function $f_{Y|X=x}(y)$ of Y given that $X = x$. [**4**]

(iv) Calculate the conditional expectation $E(\mathrm{e}^{0.5Y}|X = x)$. [**4**]

Question 3 [16 marks].

(a) The random variable X has Gamma distribution $\mathrm{Ga}(\alpha, \beta)$.

You are reminded that the probability density function of X is given by $f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}$ if $x > 0$ and $f_X(x) = 0$ otherwise.

(i) Find the moment generating function $M_X(t)$. **[5]**

(ii) Find the expectations $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$.
Hint. Use the moment generating function found in (a)(i). **[4]**

(b) Let X_1, X_2, ..., X_n be independent identically distributed exponential random variables, $X_i \sim \mathrm{Exp}(\lambda)$.

Find the probability density function $f_Y(y)$ of the sum $Y = \sum_{i=1}^n X_i$.

Hint. Compute $M_Y(t)$. **[7]**

Question 4 [6 marks]. Suppose that you roll a fair 4-sided die n times. The sides of the die are marked with numbers 1, 2, 3, and 4. Denote by X_i the number on the side of the die on which it lands on the i^{th} roll. Set $S_n = \sum_{i=1}^n X_i$. Prove that

$$\lim_{n\to\infty} \mathbb{P}(2.48n < S_n < 2.53n) = 1. \qquad \textbf{[6]}$$

Question 5 [30 marks]. Consider two independent random samples $X_1, X_2, ..., X_{n_1} \overset{iid}{\sim}$ Bernoulli(p_1) and $Y_1, Y_2, ..., Y_{n_2} \overset{iid}{\sim}$ Bernoulli(p_2), where p_1 and p_2 are unknown parameters. Set $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$.

(a) This question is about the approximate two-sample binomial Z-test for testing the hypothesis $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$ at significance level α.

(i) Compute the variance of $\overline{X} - \overline{Y}$ and write down the expression for this variance under H_0 in terms of p, where $p = p_1 = p_2$. **[6]**

(ii) State the pooled estimate $\hat{p}$ for p under H_0 and explain why $\hat{p} \to p$ as $n_1 + n_2 \to \infty$. **[3]**

(iii) State the statistic you are going to use in this test. **[3]**

(iv) What is the approximate distribution of the test statistic (given that H_0 is true)? **[3]**

(v) State the rejection rule for this test. **[3]**

(b) A basketball coach has to decide who of the two members of his team, John or Bill, is better prepared for the next game. During the last training session John scored 16 times out of 20 attempts and Bill scored 10 times out of 16 attempts.

Let p_1 be the frequency of successes for John and p_2 be the frequency of successes for Bill. The coach computes the estimates $\bar{x} = \frac{16}{20} = 0.8$ for p_1 and $\bar{y} = \frac{10}{16} = 0.625$ for p_2. Since $\bar{x} > \bar{y}$, he decides that at present John is in a better shape than Bill.

Test the hypothesis $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$ at significance level $\alpha = 0.02$. Does the result of your test support the coach's decision? **[7]**

(c) Using the above data find the 95% Wald confidence interval for $p_1 - p_2$. **[5]**

Question 6 [10 marks]. A six sided die was rolled repeatedly 300 times. The results of the observation were divided in three groups. Group A consists of results with a 1 on the top. Group B consists of results with a 2 or 3 on the top. Group C consist of all other results. The total number of results in each group is presented in the following table:

A	B	C
58	110	132

(a) Test the claim that the die is fair at 5% significance level. **[7]**

(b) Find the P-value of the test. **[3]**

End of Paper.