Main Examination period 2023 – January – Semester A

MTH5129: Probability & Statistics II

Duration: 2 hours

The exam is intended to be completed within **2 hours**. However, you will have a period of **4 hours** to complete the exam and submit your solutions.

For actuarial students only: This module also counts towards IFoA exemptions. For your submission to be eligible, **you must submit within the first 3 hours.**

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

All work should be **handwritten** and should **include your student number**. Only one attempt is allowed – **once you have submitted your work, it is final**.

In completing this assessment:

- You may use books and notes.
- You may use calculators and computers, but you must show your working for any calculations you do.
- You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.
- You must not seek or obtain help from anyone else.

When you have finished:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;

Examiners: C. Beck, D. Kalogiros

Question 1 [27 marks]. Suppose that X and Y have a joint probability density function given by

$$f_{X,Y}(x,y) = \begin{cases} 7e^{-x-7y} & \text{if } x, y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Verify that $f_{X,Y}$ is indeed a probability density function. [**5**]

(b) Find the marginal probability density function f_X and state the name of the distribution of X. [**6**]

(c) Find the conditional probability density function $f_{Y|X=x}$. [**5**]

(d) Are the random variables X and Y independent? Justify your answer. [**3**]

(e) Let another probabilty density $\widetilde{f}_{X,Y}$ be given by

$$\widetilde{f}_{X,Y}(x,y) = \begin{cases} ce^{-(\sqrt{x}+\sqrt{y})^2+2\sqrt{xy}} & \text{if } x \geq y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Determine the normalization constant c in this case. Are X and Y statistically independent? Justify your answer. [**8**]

Solution:

(a) [**seen similar**]
We observe that $f_{X,Y}(x,y) \geq 0$ for all x and y.

Then, we prove that

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx &= \int_0^{\infty}\int_0^{\infty} 7e^{-x-7y}\,dy\,dx \\ &= \int_0^{\infty} [-e^{-x-7y}]_0^{\infty}\,dx \\ &= \int_0^{\infty} e^{-x}\,dx = [-e^{-x}]_0^{\infty} = 1.\end{aligned}$$

(b) [**Seen similar**]
The marginal density f_X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt.$$

Now if $x < 0$ then, regardless of t, $f_{X,Y}(x,t) = 0$ and so, in this case, the integral is zero. I.e., if $x < 0$ then $f_X(x) = 0$.

If $x > 0$ then,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt \\ &= \int_{-\infty}^{0} f_{X,Y}(x,t)\,dt + \int_{0}^{\infty} f_{X,Y}(x,t)\,dt \\ &= 0 + \int_{0}^{\infty} 7e^{-x-7t}\,dt \\ &= \Big[-e^{-x-7t}\Big]_{t=0}^{\infty} = e^{-x}. \end{aligned}$$

Thus the marginal density f_X is given by

$$f_X(x) = \begin{cases} e^{-x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Therefore, X is an Exp(1) random variable.

(c) **[Seen similar]**
Using $f_X(x)$ from part (a), we have by definition

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

only for $x \geq 0$. It is not defined for $x < 0$ (as $f_X(x)$ would be zero).

Then, if $y < 0$ then $f_{X,Y}(x,y) = 0$ so $f_{Y|X=x}(y) = 0$.

Finally if $y \geq 0$ then

$$f_{Y|X=x}(y) = \frac{7e^{-x-7y}}{e^{-x}} = 7e^{-7y}$$

Hence, for any given $x \geq 0$, we have

$$f_{Y|X=x}(y) = \begin{cases} 7e^{-7y} & \text{if } y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(d) **[Seen similar]**
The two random variables are independent, as the probability density function can be written as the product of two functions where the first one only depends on x and the other one only on y (Theorem in the lectures). Other arguments are allowed as well.

(e) **[unseen]**
Because of $(\sqrt{x}+\sqrt{y})^2 - 2\sqrt{xy} = x + 2\sqrt{xy} + y - 2\sqrt{xy} = x + y$ we have $\tilde{f}_{X,Y}(x,y) = ce^{-x-y}$ for $x \geq y \geq 0$ and 0 else. Hence

$$\begin{aligned} 1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{f}_{X,Y}(x,y)\,dxdy &= \int_{0}^{\infty}\int_{y}^{\infty} ce^{-(x+y)}\,dxdy \\ &= \int_{0}^{\infty} ce^{-y}[-e^{-x}]_{x=y}^{x=\infty}\,dy \\ &= c\int_{0}^{\infty} e^{-2y}\,dy = \frac{c}{-2}[e^{-2y}]_0^{\infty} = \frac{c}{2}. \end{aligned}$$

Thus $c = 2$. X and Y are not independent since the condition $X \geq Y$ influences the possible value of the random variable X via the value taken by Y (other arguments allowed as well, can for example explicitly show that $f_X(x)f_Y(y) \neq f_{X,Y}(x, y)$).

Question 2 [11 marks]. Consider a standard Normal random variable $Z \sim N(0, 1)$. Use the method of cumulative distribution functions, or any other method, to find the probability density function $f_U(u)$ of

$$U = Z^5 \,. \qquad [\mathbf{11}]$$

Solution: [seen other example] For $Z \sim N(0, 1)$, we have

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad -\infty < z < \infty.$$

Then if $U = Z^5$, we have

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(Z^5 \leq u) \\ &= P\left(Z \leq u^{\frac{1}{5}}\right) \\ &= \int_{-\infty}^{u^{\frac{1}{5}}} f_Z(z)dz \\ &= \Phi\left(u^{\frac{1}{5}}\right) \,. \end{aligned}$$

So if we differentiate both sides with respect to u we find

$$\begin{aligned} f_U(u) &= \frac{1}{5} u^{-\frac{4}{5}} f_Z\left(u^{\frac{1}{5}}\right) \\ &= \frac{1}{5\sqrt{2\pi}} u^{-\frac{4}{5}} \exp\left(-\frac{1}{2} u^{\frac{2}{5}}\right) \end{aligned}$$

for all $u \in \mathbb{R}$.

Question 3 [12 marks]. Suppose that $Z_1, Z_2, \ldots, Z_n$ are statistically independent random variables. Define Y as the sum of squares of these random variables:

$$Y = \sum_{i=1}^{n} Z_i^2 \quad (n \geq 2)$$

(a) Express the moment generating function $M_Y(t)$ of the random variable Y in terms of moment generating functions involving the random variables Z_i^2, $i = 1, \ldots, n$. [**4**]

(b) Determine $M_Y(t)$ for the special case that $Z_i \sim N(0, 1)$. [**4**]

(c) For the above special case, calculate $E[Y]$ by using the moment generating function. [4]

Solution:

(a) [**seen slightly different example**] We have

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t\sum_{i=1}^n Z_i^2}] \\ &= E[e^{tZ_1^2}e^{tZ_2^2}\cdots e^{tZ_n^2}] \\ &= E[e^{tZ_1^2}]E[e^{tZ_2^2}]\cdots E[e^{tZ_n^2}] \qquad \text{(by independence)} \\ &= M_{Z_1^2}(t)M_{Z_2^2}(t)\cdots M_{Z_n^2}(t) \end{aligned}$$

(b) [**new example**] From the lecture notes we have that if $Z \sim N(0,1)$ then Z^2 has moment generating function $M_{Z^2}(t) = \left(\frac{1/2}{1/2-t}\right)^{\frac{1}{2}}$. Thus $M_Y(t) = \left(\frac{1/2}{1/2-t}\right)^{\frac{n}{2}} = (1-2t)^{-\frac{n}{2}}$

(c) [**new example**] $M_Y'(t) = -\frac{n}{2}(-2)(1-2t)^{\frac{n}{2}-1} = n(1-2t)^{-\frac{n}{2}-1}$. Thus $E[Y] = M_Y'(t)|_{t=0} = n$.

Question 4 [25 marks]. A family software company operating in 4 cities in the UK would like to automate customer service and reduce repeat calls in their call centres. Hence, they would like to fully implement AI-powered virtual agents in order to reduce costs as well as enhance the customer service performance. In order to fully understand consumer acceptance of this new technology, they collected data from a pilot survey during the last 90 days on the number of complaints received after contacting customer service and being served either by a virtual agent or not (e.g. served by a human agent or directed to the FAQs (Frequently Asked Questions) online page, among others). The relevant data are presented in the following tables below:

Number of complaints received when being served by a Virtual Agent	Observed Frequency
0	35
1	12
2	13
3	11
4	5
5	4
6	3
7	2
8	2
9	2
10	1
11 or more than 11	0

Number of complaints received when not being served by a Virtual Agent	Observed Frequency
0	19
1	14
2	15
3	9
4	14
5	6
6	4
7	2
8	2
9	3
10	2
11 or more than 11	0

(a) Estimate the average number of complaints received per day in the two different types of customer service where a customer is served by a Virtual Agent or not served by a Virtual Agent. [**6**]

(b) Test the hypothesis that the average number of complaints received per day when served by a Virtual agent is the same as the number of complaints received per day after not being served by a Virtual Agent at the 5% significance level. [**15**]

(c) Find an approximation to the 95% confidence interval for the difference of the average number of complaints received in the two types of customer service (with or without Virtual Agent). [**4**]

NB. In your answers to all the questions above, report the numerical computations to three decimal places.

Solution:

(a) [**Seen similar**]
We assume that number of complaints received follow a Poisson distribution with means μ_1 and μ_2 for the two different types of customer service with and without a Virtual Agent, respectively. We can estimate μ_1 and μ_2 by evaluating the sample means of the observed data, i.e.

$$\hat{\mu}_1 = \frac{0 \times 35 + 1 \times 12 + 2 \times 13 + 3 \times 11 + 4 \times 5 + 5 \times 4 + 6 \times 3 + 7 \times 2 + 8 \times 2 + 9 \times 2 + 10 \times 1}{35 + 12 + 13 + 11 + 5 + 4 + 3 + 2 + 2 + 2 + 1}$$

$$= 2.078$$

$$\hat{\mu}_2 = \frac{0 \times 19 + 1 \times 14 + 2 \times 15 + 3 \times 9 + 4 \times 19 + 5 \times 6 + 6 \times 4 + 7 \times 2 + 8 \times 2 + 9 \times 3 + 10 \times 2}{19 + 14 + 15 + 9 + 14 + 6 + 4 + 2 + 2 + 3 + 2}$$

$$= 2.867$$

(b) **[Seen similar]**
In order to test $H_0 : \mu_1 = \mu_2$, we use the following test statistic

$$T = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\frac{\hat{\mu}_2}{n_2} + \frac{\hat{\mu}_1}{n_1}}}$$

$$= \frac{2.867 - 2.078}{\sqrt{\frac{2.867}{90} + \frac{2.078}{90}}}$$

$$= 3.365$$

where n_1 and n_2 denote the sample sizes which are both sufficiently large, as they are equal to 90 (days). The distribution of the statistic T can, therefore, be approximated by a standard normal distribution and based on the value of the statistic that we calculated above, we reject the null hypothesis H_0 as the value of T (3.365) is greater than the cut-off value 1.96 at the 5% level of significance.

(c) **[Seen similar]**
Based on the lecture notes, an approximate 95% CI for the difference of the average number of complaints received in the two types of customer service is

$$\hat{\mu}_2 - \hat{\mu}_1 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}}$$

$$\begin{aligned} &= (2.867 - 2.078) \pm 1.96\sqrt{\frac{2.078}{90} + \frac{2.867}{90}} \\ &= (0.328, 1.250) \end{aligned}$$

Question 5 [25 marks]. Lecturers at a university intend to see the effect of stress relief strategies such as deep breathing and relaxation techniques to the student performance in the final exam. For this reason, they assigned students of the same year in two different cohorts. More precisely, students in Cohort 1 did not do any relaxation techniques before the exam, while students in Cohort 2 performed a full range of the relaxation techniques approved by the British Psychological Society. The average mark of a random sample of 17 students from Cohort 1 was 61, with a standard deviation of 17, while the average mark of a random sample of 11 students from Cohort 2 was 67 with a standard deviation of 19.

(a) Test whether the variances of marks for the two cohorts are the same at the 5% significance level. **[7]**

(b) Test the hypothesis that the average mark in the exam for Cohort 1 students is the same as that for students of Cohort 2 at the 5% significance level. For this test, what is the p-value? **[9]**

(c) What does the p-value that you calculated in (b) indicate about the null hypothesis? **[2]**

(d) Calculate a 95% confidence interval for the difference in average marks in the exam for the two cohorts of students. Explain the meaning of the 95% confidence interval that you have calculated. **[7]**

NB. In your answers to all the questions above, report the numerical computations to three decimal places.

Solution:

(a) **[Seen similar]**
We will use the F test in order to test whether the variances of marks for the two cohorts are the same. Let the variance of marks at Cohort 1 be denoted as σ_1^2 and the variance of marks for Cohort 2 be given by σ_2^2. From the given sample standard deviations we have $S_1^2 = 289$ and $S_2^2 = 361$ for the corresponding sample variances. We test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$. In order to do this, we use the F-test statistic, i.e.

$$F = \frac{S_1^2}{S_2^2}.$$

Under the null hypothesis, F has an F_{10}^{16} distribution. The lower and upper cutoff values are given by 0.335 and 3.496. The observed value of F is $289/361 = 0.801$. We, therefore, do not reject the null hypothesis of equality of variances.

(b) **[Seen similar]**
Given that we have not rejected the null hypothesis of equality of variances, we use the standard 2-sample t-test. We start by computing the pooled estimate of variance

$$S_0^2 = \frac{16 \times 289 + 10 \times 361}{17 + 11 - 2} \qquad = 316.692.$$

The test statistic is

$$\begin{aligned} T &= \frac{61 - 67}{\sqrt{316.692}\sqrt{\frac{1}{17} + \frac{1}{11}}} \\ &= -0.871. \end{aligned}$$

Under the null hypothesis, the distribution of the test statistic is t_{26}. The cutoffs are -2.056 and 2.056 hence we do not reject the null hypothesis that the average mark in Cohort 1 is the same as in Cohort 2.

(c) [**Seen similar**]
The p-value of the test is given by $2 \times P(t_{26} < -0.8735) = 0.392$. This p-value indicates that we have weak evidence against the null hypothesis and we cannot reject it.

(d) [**Seen similar and the last part is new**]
Let $\bar{X}_1$ and $\bar{X}_2$ denote the sample means for the marks from Cohort 1 and Cohort 2, respectively. A confidence interval for the difference is given by

$$\bar{X}_1 - \bar{X}_2 \pm t_{26}(0.975)\sqrt{316.692}\sqrt{\frac{1}{17} + \frac{1}{11}}$$

$$= (-20.160, 8.160).$$

Regarding the interpretation of the calculated confidence interval, there are different formulations of the correct answer based on the definition of the confidence interval as it is presented and highlighted during the lectures.

End of Paper.