Queen Mary
University of London

Main Examination period 2022 – January – Semester A

MTH5129: Probability & Statistics II

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

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.**

All work should be **handwritten** and should **include your student number**.

The exam is available for a period of **24 hours**. Upon accessing the exam, you will have **3 hours** in which to complete and submit this assessment.

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;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

Please try to upload your work well before the end of the submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

IFoA exemptions. For actuarial students, this module counts towards IFoA actuarial exemptions. To be eligible for IFoA exemption, **your must submit your exam within the first 3 hours of the assessment period**.

Examiners: C. Beck, P. Bhuyan

 Continue to next page

In this exam, $P(\cdot)$ denotes a probability measure defined on a space $(\Omega, \mathcal{F})$ and $E(\cdot)$ denotes the expectation with respect to P.

Question 1 [30 marks]. Suppose that X and Y have a joint probability density function given by

$$f_{X,Y}(x,y) = \begin{cases} ce^{-3x-5y} & \text{if } x, y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Determine the value of the normalization constant c. [**5**]

(b) Find the marginal probability density function f_X and state the name of the distribution of X. [**8**]

(c) Find the conditional probability density function $f_{Y|X=x}$. [**8**]

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

Consider now a different joint probability density function for X and Y, namely

$$\widetilde{f}_{X,Y}(x,y) = \begin{cases} 12ye^{-3x-2y^2} & \text{if } x, y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(e) What is the probability $P(Y^2 > 2X > 0)$? [**6**]

Solution:

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

Then, we need to consider normalization

$$\begin{aligned} 1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx &= \int_0^{\infty}\int_0^{\infty} ce^{-3x-5y}\,dy\,dx \quad \text{[2 marks]} \\ &= \int_0^{\infty} [-\frac{c}{5}e^{-3x-5y}]_0^{\infty}\,dx \\ &= \int_0^{\infty} \frac{c}{5}e^{-3x}\,dx = \frac{c}{15}[-e^{-3x}]_0^{\infty} = 1. \end{aligned}$$

Hence $c = 15$. [3 marks]

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

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt. \quad \text{[2 marks]}$$

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 $f_X(x) = 0$.

 Continue to next page

If $x > 0$ then,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt \\ &= \int_{-\infty}^{0} f_{X,Y}(x,t)\,dt + \int_{0}^{\infty} f_{X,Y}(x,t)\,dt \\ &= 0 + \int_{0}^{\infty} 15e^{-3x-5t}\,dt \\ &= \Big[-3e^{-3x-5t}\Big]_{t=0}^{\infty} = 3e^{-3x}. \end{aligned} \quad \text{[3 marks]}$$

Thus the marginal density f_X is given by

$$f_X(x) = \begin{cases} 3e^{-3x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{[2 marks]}$$

Therefore, X is an Exp(3) random variable [1 mark].

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

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{[2 marks]}$$

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

$$f_{Y|X=x}(y) = \frac{15e^{-3x-5y}}{3e^{-3x}} = 5e^{-5y} \quad \text{[4 marks]}$$

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

$$f_{Y|X=x}(y) = \begin{cases} 5e^{-5y} & \text{if } y \geq 0 \\ 0 & \text{otherwise.} \end{cases} \quad \text{[2 marks]}$$

(d) **[Seen similar]**
Yes, they are independent [2 marks], because e.g. $f_{Y|X=x}(y)$ does not depend on the value of x [1 mark].

(e) **[More difficult and partially new]**
We have $P(Y^2 > 2X) = P(\frac{Y^2}{2} > X)$

$$P(\frac{Y^2}{2} > X > 0) = \int_0^{\infty} \int_0^{\frac{y^2}{2}} \widetilde{f}_{X,Y}(x,y)\,dx\,dy = \int_0^{\infty} \int_{\sqrt{2x}}^{\infty} \widetilde{f}_{X,Y}(x,y)\,dy\,dx \quad \text{[2 marks]}$$

where we can use either of the two ways of parameterising the region $y^2 > 2x > 0$.

 Continue to next page

Using the former (either would work):

$$\begin{aligned} P(\frac{Y^2}{2} > X > 0) &= \int_0^\infty \int_0^{\frac{y^2}{2}} \widetilde{f}_{X,Y}(x,y)\,dx\,dy \\ &= \int_0^\infty \int_0^{\frac{y^2}{2}} 12ye^{-3x-2y^2}\,dx\,dy \\ &= \int_0^\infty \Big[-4ye^{-3x-2y^2} \Big]_0^{\frac{y^2}{2}}\,dy \\ &= \int_0^\infty (4ye^{-2y^2} - 4ye^{-\frac{7}{2}y^2})\,dy \\ &= \Big[-e^{-2y^2} + \frac{4}{7}e^{-\frac{7}{2}y^2} \Big]_{y=0}^\infty = \frac{3}{7} \quad \text{[4 marks]} \end{aligned}$$

Question 2 [10 marks]. Suppose that X_1 and X_2 have joint probability density function

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} \frac{1}{6}x_1x_2, & \text{if } 1 < x_1 < 2 \text{ and } 1 < x_2 < 3 \\ 0, & \text{otherwise.} \end{cases}$$

What is the joint probability density function f_{Y_1,Y_2} of $Y_1 = X_1/X_2$ and $Y_2 = X_2$? **[10]**

Solution: [Method seen but example is new]

The inverse is

$$x_1 = y_1\,y_2 \quad \text{and} \quad x_2 = y_2. \quad \text{[2 marks]}$$

The determinant of the Jacobian is

$$J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2. \quad \text{[2 marks]}$$

The maximum value of $Y_1 = X_1/X_2$ is given by $X_1^{max}/X_2^{min} = 2/1 = 2$. The minimum value of Y_1, likewise, is given by $X_1^{min}/X_2^{max} = 1/3$.
So, using the direct transformation of two random variables method

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \frac{1}{6}[(y_1\,y_2)y_2]\,y_2 = \frac{1}{6}y_1\,y_2^3, & \text{for } \frac{1}{y_2} < y_1 < \frac{2}{y_2} \text{ and } 1 < y_2 < 3 \\ 0, & \text{otherwise.} \end{cases}$$

[6 marks]

Question 3 [10 marks]. Suppose that X, Y and Z are statistically independent random variables, each of them with a $\chi^2(2)$ distribution.

(a) Find the moment generating function of $U = X + 3Y + Z$. State clearly and justify all steps taken. **[7]**

(b) Calculate the expectation $E(U)$ using the moment generating function. **[3]**

Hint: You may use without proof the fact that the moment generating function of a $\chi^2(\nu)$ random variable W is

$$M_W(t) = \left(\frac{1}{1-2t}\right)^{\nu/2}.$$

Solution: [Seen similar, but for only 2 variables.]
The mgf of X as $\chi^2(2)$ random variable is

$$M_X(t) = \frac{1}{1-2t}. \quad [\ 2 \text{ marks}]$$

For $3Y$, we have for $Y \sim \chi^2(2)$ that

$$M_{(3Y)}(t) = E(e^{t\,(3Y)}) = E(e^{(3t)\,Y}) = \frac{1}{1-6t}. \quad [2 \text{ marks}]$$

Since X, Y and Z are independent, then equivalently X and $3Y$ and Z are also independent. Hence, we have

$$M_U(t) = M_{X+3Y+Z}(t) = M_X(t)\, M_{3Y}(t)\, M_Z(t).$$

Thus

$$M_U(t) = \frac{1}{1-2t}\frac{1}{1-6t}\frac{1}{1-2t} = \frac{1}{1-4t+4t^2}\cdot\frac{1}{1-6t} = \frac{1}{1-10t+28t^2-24t^3}. \quad [3 \text{ marks}]$$

We then calculate by differentiating the mgf

$$E(U) = M_U'(0) = \left.\frac{10-56t+72t^2}{(1-10t+28t^2-24t^3)^2}\right|_{t=0} = 10. \quad [3 \text{ marks}]$$

Question 4 [25 marks]. Transport engineers are interested to study the pattern and the risk of accidents at a junction. Suppose we observe the following frequency distribution of the number of accidents in 150 weeks.

Number of Accidents	0	1	2	3	4	5	6	7+
Observed frequency	50	30	24	30	10	5	1	0

(a) Find an estimate (to four decimal places) of the average number of accidents in a week. **[6]**

(b) Perform a goodness of fit test at the 5% significance level of the null hypothesis that the observed number of accidents follow a Poisson distribution.
Hint: Use the estimate (to four decimal places) you computed in (a) and report the expected frequency and observed value of the statistic to four decimal places. **[15]**

 Continue to next page

(c) What is the p-value of your test in (b)? Does the p-value indicate that there is evidence against the null hypothesis? [**4**]

Solution:

(a) **[Seen similar]**
To find the Poisson probabilities we need the mean μ. A reasonable estimate of μ, is given by sample mean of the observed data

$$\frac{0\times 50+1\times 30+2\times 24+3\times 30+4\times 10+5\times 5+6\times 1}{50+30+24+30+10+5+1} \qquad \text{[2 marks]}$$

$$= 1.5933. \qquad \text{[4 marks]}$$

(b) **[Seen similar]**
Now using the Poisson formula

$$P[Y=y]=\frac{e^{-\mu}\mu^y}{y!}$$

or R functions we can compute the probabilities in the following table. [5 marks]

Number of Accidents	0	1	2	3	4	5	6	7+
Observed frequency (O_i)	50	30	24	30	10	5	1	0
Expected frequency (E_i)	30.4881	48.5766	38.6986	20.5528	8.1867	2.6088	0.6928	0.1957

Now the last three expected frequencies are all less than 5. If we group them together into a class 5+ the expected frequency will be 3.4973, still less than 5. So we group the last four classes into a class 4+ with expected frequency 11.6840 and observed frequency 16. Now we compute

$$\begin{aligned} X^2 &= \frac{(30.4881-50)^2}{30.4881}+\frac{(48.5766-30)^2}{48.5766}+\frac{(38.6986-24)^2}{38.6986} \\ &\quad +\frac{(20.5528-30)^2}{20.5528}+\frac{(11.6840-16)^2}{11.6840} \\ &= 31.1110. \qquad \text{[5 marks]} \end{aligned}$$

Now after our grouping there are four classes, $k=5$ and we estimated one parameter, the mean, from the data so $d=1$. Thus $\nu=5-1-1=3$.

Computing the observed value of the test statistic X^2 we get 31.1110. This is greater than the cut-off value of 7.81 for χ^2 distribution with 3 degrees of freedom, hence we reject the null hypothesis that the observed data follows Poisson distribution. [5 marks]

(c) **[Seen similar]**
The p-value for this test is given by $P(X^2>31.1110)$, assuming that $X^2\sim\chi^2_3$. [2 marks]
We have from R that the P value is given by

$> 1 - pchisq(31.1110, 3)$
$8.055093e - 07$
Such a small p-value shows evidence against the hypothesis that the data have a Poisson distribution. [2 marks]

Question 5 [25 marks]. Twenty patients sampled at random were matched by age and BMI. One of each pair was assigned at random to a new treatment and the other to an existing treatment. Ultrasound examination of a certain tumour weight (in grams) produced the following results.

Existing Treatment	15.5	5	8.5	10.6	12.5	6.5	8.4	10.5	8	10
New Treatment	14	5.7	9	11.5	10	7	7.5	8	9.9	10.5

To answer the following questions, report the numerical computations to four decimal places.

(a) Test whether there is a difference in the two treatments at 5% level of significance. [**10**]

(b) Find a 90% confidence interval for the mean difference in the treatments. [**5**]

(c) What would be the conclusion in (a) if we had wrongly ignored the pairing? [**10**]

Solution:

(a) **[Seen similar]**
The differences are $+1.5, -0.7, -0.5, -0.9, -2.5, -0.5, -0.9, -1.5, -1.9, -0.5$. The mean difference is $\bar{d} = -0.84$, and the standard deviation of the differences is $s_d = 1.0627$. [3 marks]
The null hypothesis is $\mu_d = 0$ versus an alternative that $\mu_d \neq 0$. The test statistic is

$$t = \frac{\bar{d}\sqrt{n}}{s_d}$$

which follows a t distribution with 9 degrees of freedom if H_0 is true. The observed value of the statistic is $t = -2.4996$. [4 marks]
The rejection region is $\{t^* : |t^*| > 2.2622\}$. Therefore we reject H_0 at the 5% significance level. [3 marks]

(b) **[Seen similar]**
A 90% confidence interval is of the form

$$\bar{d} \pm t_9(0.95)\frac{s_d}{\sqrt{n}} \qquad \text{[2 marks]}$$

$$= -0.84 \pm 1.8331 \times \frac{1.0627}{\sqrt{10}} \qquad \text{[1 marks]}$$

$$= (-1.4560, -0.2240) \qquad \text{[2 marks]}$$

.

(c) **[Seen similar]**
We would use a 2 sample t-test. The mean and variance of the new treatment is 9.31 and 5.8321, respectively. The mean and variance of the existing treatment is 9.55 and 9.0161, respectively. The pooled estimate of variance is 7.0485. [3 marks]
The observed value of the test statistic is

$$\frac{9.31 - 9.55}{\sqrt{7.0485}(\frac{1}{10} + \frac{1}{10})^{1/2}} = -0.2021. \qquad \text{[4 marks]}$$

We are comparing with a t_{18} and rejection region is $\{t^* : |t^*| > 2.1009\}$. Hence, the conclusion would be that there is no difference in means if we had wrongly ignored the pairing. [3 marks]

End of Paper.