

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

21 April 2023 (am)

Subject CS1 – Actuarial Statistics Core Principles

Paper A

Time allowed: Three hours and twenty minutes

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator.

If you encounter any issues during the examination please contact the Assessment Team on T. 0044 (0) 1865 268 873.

1 Two independent random variables, X and Y , have densities $f_X(x) = \lambda^2 x e^{-\lambda x}$ and $f_Y(y) = \lambda e^{-\lambda y}$ respectively, where $\lambda > 0$, $x \geq 0$ and $y \geq 0$. The random variable Z is defined as the sum of X and Y ($Z = X + Y$).

(i) Identify the distribution of Y with its parameter. [1]

(ii) Identify which **one** of the following expressions gives the probability density function of Z :

A $f_Z(z) = \frac{\lambda^3}{2} z e^{-\lambda z}$

B $f_Z(z) = \frac{\lambda^3}{2} z^2 e^{-\lambda z}$

C $f_Z(z) = \frac{\lambda^2}{2} z^3 e^{-\lambda z}$

D $f_Z(z) = \frac{\lambda}{2} z^2 e^{-\lambda z}$.

[3]

[Total 4]

2 Two friends are chatting on a social media platform: the first friend has a typing speed equal to three messages per minute while the second friend has a typing speed of two messages per minute. The process of writing the messages for each of the two friends is modelled with a Poisson process and these processes are assumed independent.

Calculate the probability that only two messages in total are exchanged between the two friends during the first minute. [4]

3 An Actuary determines that the claim size for a certain class of accident is a random variable, X , with moment-generating function:

$$M_X(t) = \frac{1}{(1 - 2500t)^4}, \text{ where } t < \frac{1}{2500}$$

Determine, using $M_X(t)$, the standard deviation of the claim size for this class of accident. [4]

4 Statisticians A and B obtain independent samples X_1, \dots, X_{10} and Y_1, \dots, Y_{17} respectively, both from a Normal distribution with expectation μ and variance σ^2 , with both μ and σ unknown. The variance, σ^2 , can be estimated by the sample variance of each sample, denoted as S_X^2 and S_Y^2 .

(i) Identify which one of the following options gives the correct probability that S_X^2 exceeds $1.5\sigma^2$:

A $P(S_X^2 > 1.5\sigma^2) = 0.14$

B $P(S_X^2 > 1.5\sigma^2) = 0.18$

C $P(S_X^2 > 1.5\sigma^2) = 0.10$

D $P(S_X^2 > 1.5\sigma^2) = 0.21$.

[3]

(ii) Calculate the probability that S_Y^2 exceeds $1.5\sigma^2$.

[3]

[Total 6]

5 Two people are playing a game together, involving the toss of a single coin. The coin used is biased so that the probability of throwing a head is an unknown constant, h . It is known that h must be either 0.35 or 0.85. Prior beliefs about h are given by the following distribution:

$$P(h = 0.35) = 0.7 \quad P(h = 0.85) = 0.3$$

The coin is tossed 15 times, and nine heads are observed.

Determine the posterior probabilities for the two possible values of h .

[7]

- 6 An Analyst concludes from an empirical study that the number of children, X , per family in a certain region has the following distribution:

Number of children, x	0	1	2	3	More than 3
Probability $p_x = P[X=x]$	0.05	0.42	0.4	0.1	0.03

When asked for the expected number of children per family, the Analyst claims that an exact value cannot be calculated for the expectation, but that a lower limit can be provided.

- (i) Explain whether the Analyst is right. [3]
- (ii) Calculate a lower limit for $E[X]$, the expected number of children per family in this region. [3]

From a further study of families with more than three children, it is concluded that the conditional expectation $E[X|X > 3] = 4.5$.

- (iii) Identify which **one** of the following options gives the expectation of X :

- A $E[X] = 1.915$
B $E[X] = 1.175$
C $E[X] = 1.655$
D $E[X] = 1.325$.

[3]
[Total 9]

7 Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be pairs of random variables with each pair (X_i, Y_i) being independent of all other pairs. The distribution of X_i is $N(0, 1)$, for $i = 1, \dots, n$. The conditional distribution of Y_i , given that X_i takes a particular value x_i , is $N(x_i\theta, 1)$, for $i = 1, \dots, n$ where $\theta \in (-\infty, +\infty)$ is an unknown parameter.

(i) Identify which **one** of the following options gives the correct expression of the likelihood function:

A
$$L(\theta) = \prod_{i=1}^n \exp \left[\frac{(y_i + x_i\theta)^2 - x_i^2}{2} \right]$$

B
$$L(\theta) = \prod_{i=1}^n \frac{1}{2\pi} \exp \left[-\frac{(y_i - x_i\theta)^2 + x_i^2}{2} \right]$$

C
$$L(\theta) = \prod_{i=1}^n \frac{\pi}{2} \exp \left[-\frac{(y_i - x_i\theta)^2 - x_i^2}{2} \right]$$

D
$$L(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(y_i + x_i\theta)^2 + x_i^2}{2} \right].$$

[3]

(ii) Determine the maximum likelihood estimator $\hat{\theta}$ of θ . You do not need to check that your solution is a maximum of the likelihood function. [4]

(iii) Determine the Cramer–Rao lower bound for $\hat{\theta}$. [4]

(iv) Write down the asymptotic distribution of $\hat{\theta}$. [2]

[Total 13]

- 8 A space rocket contains six identical mechanical components that work independently of each other and need to be in operation for a successful launch. Data from simulated launches are available to establish the relationship between the number of damaged components (Y) on the rocket and air temperature (X , in degrees Fahrenheit). It is suggested to analyse the simulated data using a binomial Generalised Linear Model (GLM) with the canonical link function, where $Y \sim \text{Binomial}(6, p)$, p is the probability of a component being damaged, and the linear predictor has the form:

$$\beta_0 + \beta_1 X.$$

The analysis of the simulated data gave the following estimates for the model:

	<i>Estimate</i>	<i>Standard error</i>
β_0	11.6630	3.2963
β_1	-0.2162	0.0532

- (i) Determine whether air temperature significantly affects the number of damaged components on the rocket by computing a suitable p -value. [5]
- (ii) Estimate (to two decimal places) the probability that a component will be damaged when the air temperature is 31 degrees Fahrenheit. [2]
- (iii) Estimate the expected value of the number of components that will be damaged when the air temperature is 31 degrees Fahrenheit. [1]

It is believed that the launch is safe when at least five of these six components are not damaged.

- (iv) (a) Calculate the probability that the launch is safe when the air temperature is 31 degrees Fahrenheit. [3]
- (b) Comment on the safety of the launch when the air temperature is 31 degrees Fahrenheit. [1]

A second approach for analysing the simulated data was suggested, where a logarithmic link function would be used with the same GLM as used before.

- (v) Comment on the suitability of the second approach. [2]
- [Total 14]

- 9 An insurer models the number of cars owned by a single policyholder as a discrete random variable with the following distribution, where p is an unknown parameter:

Number of cars	0	1	2	3	More than 3
Probability	$\frac{1}{2}p$	p	$\frac{1}{4}p$	$\frac{1}{4}p$	$1 - 2p$

In an empirical study, the number n_k of policyholders owning k cars is recorded, for $k = 0, \dots, 3$, and n_4 is the number of policyholders with more than three cars.

- (i) Identify which **one** of the following functions is the log likelihood function for p using the recorded numbers n_0, \dots, n_4 , where C is a constant independent of p :

A $(n_0 + n_1 + n_2 + n_3)\frac{1}{4}\log p + n_4 \log(1 - 2p) + C$

B $(n_0 + n_1 + n_2 + n_3) \log \frac{1}{4} \log p + n_4 \log(1 - 2p) + C$

C $\log(n_0 + n_1 + n_2 + n_3) \log p + \log n_4 \log(1 - 2p) + C$

D $(n_0 + n_1 + n_2 + n_3) \log p + n_4 \log(1 - 2p) + C.$

[3]

- (ii) Derive the maximum likelihood estimator for the parameter p . You do not need to check that your solution is a maximum of the likelihood function. [4]

In a larger study, the insurer has to restrict the information held for each policyholder due to data protection. The insurer records only the number of policyholders with no car, m_0 , and the number of policyholders with at least one car, m_1 .

- (iii) Show that the log likelihood function for the parameter p based on the observations m_0 and m_1 , and the above distribution for the number of cars is given by:

$$l(p) = m_0 \log\left(\frac{1}{2}p\right) + m_1 \log\left(\frac{2-p}{2}\right)$$

[4]

- (iv) Identify which **one** of the following estimators is the maximum likelihood estimator for the parameter p :

A $\hat{p} = \frac{m_0}{2m_0 + m_1}$

B $\hat{p} = \frac{2m_0}{2m_0 + m_1}$

C $\hat{p} = \frac{2m_0}{m_0 + m_1}$

D $\hat{p} = \frac{m_0}{m_0 + m_1}.$

[3]

The following data have been observed:

$$n_0 = 50, \quad n_1 = 37, \quad n_2 = 17, \quad n_3 = 16, \quad n_4 = 10$$

(v) Estimate the value of p using the estimator derived in part (ii). [1]

(vi) Estimate the value of p using the estimator found in part (iv). [1]

[Total 16]

- 10** A Banking Analyst is assessing the performance of a newly developed credit risk model against experts' knowledge. The credit scores produced on a sample of twelve customers by the experts (x) and the model (y) are the following:

x	65.8	63.7	67.6	64.4	68.2	62.9	70.5	66.4	68.0	67.1	69.5	71.8
y	68.2	66.2	68.1	66.0	69.1	66.1	68.7	65.9	69.3	67.2	67.9	70.4

Summary statistics of the data are given below:

$$\begin{aligned} \sum x_i &= 805.9 & \sum y_i &= 813.1 & \sum x_i^2 &= 54,203.21 \\ \sum y_i^2 &= 55,118.71 & \sum x_i y_i &= 54,643.17 \end{aligned}$$

- (i) Fit a linear regression line of y on x . [4]
- (ii) Calculate Pearson's correlation coefficient between the experts' and the model's scores. [1]
- (iii) Perform a statistical test, using Fisher's transformation, to determine whether the population Pearson's correlation coefficient is significantly different from 0.8. Your answer should include the p -value of the test. [5]
- (iv) Construct a 99% confidence interval for the slope parameter of the linear regression line fitted in part (i). [3]
- (v) Comment on your answers to parts (iii) and (iv). [2]

The Analyst is informed that the scores on their own are not the most important aspect of the model. Instead, the performance of the model is assessed by how well it is able to predict the rank order of the twelve customers provided by the experts. A higher score corresponds to a better customer. The rankings of the customers based on their above scores are provided in the table below:

Rank (x_i)	4	2	7	3	9	1	11	5	8	6	10	12
Rank (y_i)	8	4	7	2	10	3	9	1	11	5	6	12

- (vi) Calculate Spearman's rank correlation for the data between the model's and experts' scores. [6]
- (vii) Comment on the model's alignment with the experts' opinion, based on your results from parts (ii) and (vi). [2]

[Total 23]

END OF PAPER