INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

21 April 2022 (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 The number of emails, X, to be replied to in a day by an employee of the customer service centre of an insurance company is modelled as a Poisson random variable with mean 25. The time (in minutes), Y, that the employee takes to reply to x emails is modelled as a random variable with conditional mean and variance given by:

$$E(Y|X=x) = 3x + 11, \quad \text{Var}(Y|X=x) = x + 9.$$

Calculate the unconditional variance of the time, Y, that the employee takes to reply to emails in a day. [3]

2 The number of claims arriving at an insurance company is assumed to follow a Poisson process $\{N(t)\}_{t \geq 0}$ with rate $m = 2$ per year.

(i) State the distribution of the random variable $N(1)$. [1]

(ii) Calculate the probability of more than two claims arriving in year 2 given that five claims arrived in year 1. [2]

(iii) Calculate the probability of more than two claims arriving in year 2 given that no claims arrived in year 1. [1]

(iv) Compare the results in parts (ii) and (iii). [1]

(v) Identify the distribution of the time of the nth claim, justifying your answer. [2]

(vi) Calculate a random value from the exponential distribution with parameter $m = 2$ using a realised value of 0.201 from a $U(0,1)$ distribution and the inverse transform method. [2]

[Total 9]

3 Let X and Y be two continuous random variables jointly distributed with probability density function:

$$f_{XY}(x, y) = \begin{cases} 6e^{-(2x+3y)}, & x, y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

(i) Identify which **one** of the following options gives the correct expression for the marginal density function $f_X(x)$:

A $\quad f_X(x) = \begin{cases} 2e^{2x}, & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

B $\quad f_X(x) = \begin{cases} e^{-2x}, & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

C $\quad f_X(x) = \begin{cases} 2e^{-x}, & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

D $\quad f_X(x) = \begin{cases} 2e^{-2x}, & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

[1]

(ii) Identify which **one** of the following options gives the correct expression for the marginal density function $f_Y(y)$:

A $\quad f_Y(y) = \begin{cases} 3e^{3y}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

B $\quad f_Y(y) = \begin{cases} e^{3y}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

C $\quad f_Y(y) = \begin{cases} 3e^{-3y}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

D $\quad f_Y(y) = \begin{cases} e^{-3y}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

[1]

(iii) Comment on whether X and Y are independent, by using your results in parts (i) and (ii). [1]

(iv) Calculate the conditional expectation $E[Y|X > 2]$. [2]

(v) Identify which **one** of the following options gives the correct expression for $P(X > Y)$:

A $\quad \frac{1}{5}$

B $\quad \frac{3}{5}$

C $\quad \frac{1}{3}$

D $\quad \frac{1}{2}$

[2]
[Total 7]

4 (i) Describe what is meant by each of the following:

(a) A random sample

(b) A statistic.

[3]

A new political party is interested in the level of support it would have among the voters in a particular country. The random variable X is defined as:

$$X = \begin{cases} 1, & \text{if the voter would support the party,} \\ 0, & \text{otherwise.} \end{cases}$$

A random sample of 50 voters are presented with a simple summary of the party's policies and asked if they would support this new party. The random sample is represented by $X_1, X_2, \ldots, X_{50}$.

(ii) (a) Identify a suitable population together with a possible parameter of interest.

(b) Determine, using your answer to part (ii)(a), the sampling distribution of the statistic:

$$Y = \sum_{i=1}^{50} X_i$$

[4]
[Total 7]

5 Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables following a Poisson(m) distribution. Suppose that, rather than observing the random variables precisely, only the events $X_i = 0$ or $X_i > 0$ are observed for $i = 1, 2, \ldots, n$.

Let Y be a random variable with:

$$Y_i = \begin{cases} 0, & X_i = 0 \\ 1, & X_i > 0 \end{cases}$$

for $i = 1, 2, \ldots, n$.

(i) Explain why the distribution of Y_i is a Bernoulli (p) distribution with parameter $p = 1 - e^{-m}$. [1]

(ii) Identify which **one** of the following expressions gives the correct likelihood function based on observations $y_1, \ldots, y_n$ in terms of $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and the unknown parameter m.

A $L(m) = (1 + e^{-m})^{n\bar{y}}(e^{m})^{n-n\bar{y}}$

B $L(m) = (1 - e^{m})^{n\bar{y}}(e^{-m})^{n-n\bar{y}}$

C $L(m) = (1 - e^{-m})^{n\bar{y}}(e^{-m})^{n-n\bar{y}}$

D $L(m) = (1 - e^{-m})^{n\bar{y}}(e^{-m})^{n+n\bar{y}}$

[2]

(iii) Derive an expression for the Maximum Likelihood Estimate (MLE) $\hat{m}$ of m in terms of $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. [4]

(iv) State the condition that $\hat{m}$ and $L(m)$ must satisfy for $\hat{m}$ to maximise the likelihood function. [1]

[Total 8]

6 The size of claims on a certain type of motor insurance policy are modelled as a random variable X with Probability Density Function (PDF)

$$f(x; \alpha, \beta) = \alpha \frac{\beta^{\alpha}}{x^{\alpha+1}}, \qquad x \geq \beta, \quad \alpha, \beta > 0.$$

(i) Identify which **one** of the following expressions gives the correct log likelihood function in terms of a random sample $(x_1, x_2, \ldots, x_n)$ and the unknown parameters α and β:

A $\quad l(\alpha, \beta) = n \log \alpha + n\alpha \log \beta + (\alpha + 1) \sum_{i=1}^{n} \log x_i$

B $\quad l(\alpha, \beta) = \log \alpha + n\alpha \log \beta - (\alpha + 1) \sum_{i=1}^{n} \log x_i$

C $\quad l(\alpha, \beta) = n \log \alpha + n \log \beta - (\alpha + 1) \sum_{i=1}^{n} \log x_i$

D $\quad l(\alpha, \beta) = n \log \alpha + n\alpha \log \beta - (\alpha + 1) \sum_{i=1}^{n} \log x_i$

[2]

(ii) Derive the MLE $\hat{\alpha}$ of parameter α as a function of parameter β, for a random sample $(x_1, x_2, \ldots, x_n)$. [2]

(iii) Comment on the behaviour of the PDF of X when β increases. [1]

(iv) Determine the MLE $\hat{\beta}$ of parameter β based on your comment in part (iii). [2]

The values (in \$) of a sample of 10 claims are given in the table below:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
10,000	12,000	8,000	16,000	20,000	19,000	17,000	22,000	18,000	5,000

(v) Calculate the mean and standard deviation of the natural logarithm of the sample. [2]

(vi) Calculate the MLEs $\hat{\alpha}$ and $\hat{\beta}$ based on the sample. [2]

[Total 11]

7 The probability density function of a gamma distribution is parameterised as follows:

$$f(x) = \frac{\left(\frac{\mu}{\sigma^2}\right)^{(\mu^2/\sigma^2)}}{\Gamma\left(\frac{\mu^2}{\sigma^2}\right)} x^{\left(\frac{\mu^2}{\sigma^2}\right)-1} e^{-x\mu/\sigma^2}, \quad x \geq 0, \quad \mu, \sigma > 0.$$

This density can be expressed in the form of the exponential family, as follows:

$$\theta = -\frac{1}{\mu}, \quad b(\theta) = -\log(-\theta), \quad \phi = \frac{\mu^2}{\sigma^2}, \quad \alpha(\phi) = \frac{1}{\phi},$$

$$c(x, \phi) = (\phi - 1)\log x - \log\Gamma(\phi) + \phi \log \phi,$$

where the exponential family notation is the same as that in the Actuarial Formulae and Tables book.

(i) Justify that μ and σ^2 are the mean and the variance of the distribution, respectively, using the properties of the exponential family. [3]

An actuary is modelling the relationship between claim size and the time spent processing the claim, called operational time (opt). A statistician suggests using a model with the claim size being the response variable following the gamma distribution given above.

(ii) Comment on why a gamma distribution may be more suitable than the Normal distribution for the claim sizes. [2]

The actuary decided to fit a generalised linear model (GLM) with a gamma family and obtained the following estimates:

Parameters:

	Estimate	*Standard error*
Intercept	7.51621	0.03310
opt	0.06084	0.00296

(iii) Explain, using the model output shown above, whether the variable 'opt' is significant or not. [2]

Another statistician has suggested that an alternative model needs to take into account a legal representation variable, which shows whether or not an insured person has legal representation.

(iv) Explain the difference between the variables 'opt' and 'legal representation' in a statistical sense in the context of a GLM. [2]

The actuary now has to choose between the following two models for the claim size:

Model 1: Only opt is used as a covariate.
Model 2: Both opt and legal representation are used as covariates.

An analysis of variance (ANOVA) was carried out to assess the significance of the two covariates: opt and legal representation (denoted by lr). The results obtained are given below, where claim size is denoted by cs:

Model 1: $cs = 7.52 + 0.06 \times opt$
Model 2: $cs = 3.6 + 0.04 \times opt + 2.32 \times lr$

	Resid. df	*Resid. dev*	*Df*	*Deviance*	*Pr(>Chi)*
Model 1	45	39.987			
Model 2	44	15.869	1	24.118	0.000286

(v) Determine which model provides the better fit to the data. [2]

[Total 11]

8 The time, T, until the next lorry arrives at a customs checkpoint at the border of a country is modelled with an exponential distribution, that is, $T \sim \text{Exp}(\lambda)$, where λ is an unknown parameter. Time is measured in minutes.

(i) Identify which **one** of the following expressions gives the correct likelihood function $L(\lambda)$ for the parameter λ, based on a sample of observed times until the next lorry arrives, t_i, $i = 1, \ldots, n$:

A $L(\lambda|T) = \lambda^n \exp(-\lambda \sum t_i)$

B $L(\lambda|T) = \lambda^{n-1} \exp(-\lambda \sum t_i)$

C $L(\lambda|T) = \lambda^{n+1} \exp(-\lambda \sum t_i)$

D $L(\lambda|T) = \lambda \exp(-\lambda \sum t_i)$

[1]

An analyst uses Bayesian inference to obtain an estimate for λ. They choose a gamma distribution with parameters α and β as the prior distribution for λ.

(ii) Verify that the posterior distribution of the parameter λ is a gamma distribution with parameters $\alpha + n$ and $\beta + \sum t_i$. [4]

Assume that a total of 20 lorries have arrived at the checkpoint.

(iii) Determine the Bayesian estimator for λ, in terms of the parameters α and β, under quadratic loss based on this sample. [2]

(iv) Explain how to determine the Bayesian estimator for λ under all-or-nothing loss based on this sample. [3]

(v) Identify which **one** of the following options gives the correct Bayesian estimator for λ under all-or-nothing loss based on the sample given:

A $\lambda = \dfrac{\alpha}{\beta + 60}$

B $\lambda = \dfrac{\alpha + 19}{\beta + 60}$

C $\lambda = \dfrac{\alpha + 18}{\beta + 60}$

D $\lambda = \dfrac{\alpha + 20}{\beta + 60}$

[2]

(vi) Comment on the difference between the two estimators in parts (iii) and (v).

[1]

[Total 13]

9 Consider the linear regression model in which the response variable Y_i is linked to the explanatory variable X_i by the following equation:

$$Y_i = \alpha + \beta X_i + e_i, \quad i = 1, \ldots, n,$$

where e_i are the error terms and data (x_i, y_i), $i = 1, \ldots, n$, are available.

(i) Comment on whether or not the linear regression model as presented above can be used to make inferences on parameters α and β. [3]

The coefficient of determination for this model is given by $R^2 = \frac{S_{xy}^2}{S_{xx}\, S_{yy}}$.

(ii) Verify that R^2 gives the proportion of the total variability of Y 'explained' by the linear regression model. [3]

Consider the multiple linear regression model where the response variable Y_i is related to explanatory variables $X_1, X_2, \ldots, X_k$ by:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + e_i, \quad i = 1, \ldots, n,$$

where e_i are the error terms and relevant data are available.

(iii) Suggest three ways for assessing the fit of the multiple linear regression model to a set of data. [3]

A forward selection process is used for selecting explanatory variables in the multiple linear regression model.

(iv) Explain whether the coefficient of determination, R^2, can be used as a criterion for selecting variables when applying this process. [3]

A multiple linear regression model with four explanatory variables (X_1, X_2, X_3, X_4) is fitted to a set of data, and a forward selection process is used for selecting the optimal set of explanatory variables.

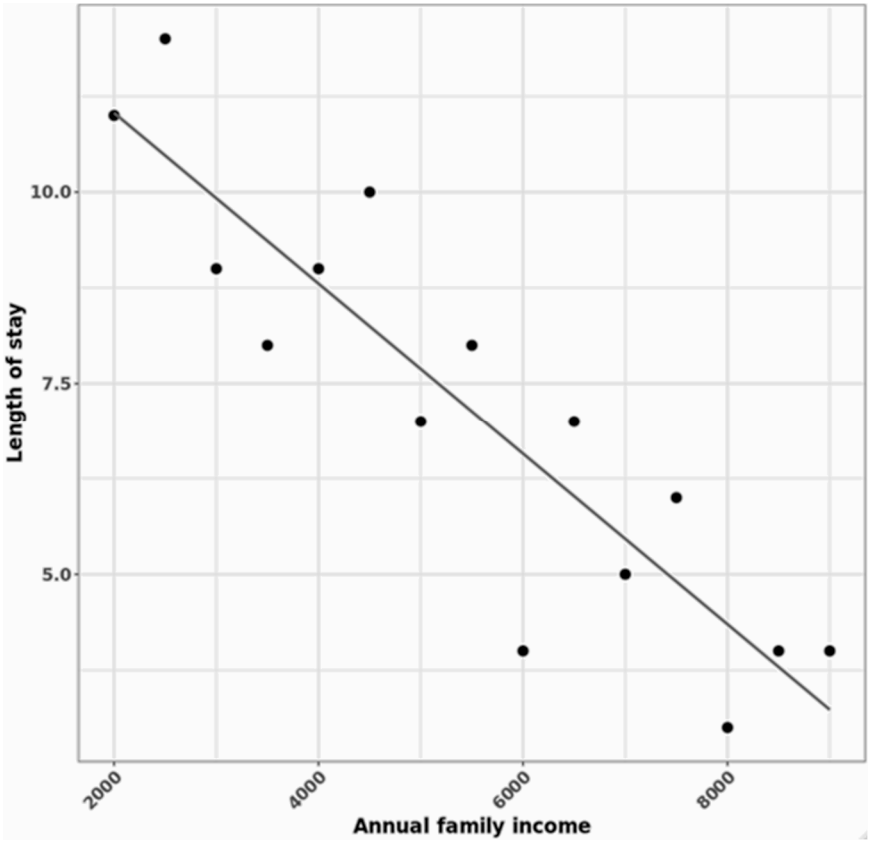
Some output of this process is shown in the following table:

Model	R^2	*Adjusted* R^2
X_1	0.7322	0.7167
$X_1 + X_4$	0.8018	0.7712
$X_1 + X_4 + X_3$	0.8253	0.7805
$X_1 + X_4 + X_3 + X_2$	0.8259	0.7684

(v) Determine the optimal set of explanatory variables using this output. [2]

[Total 14]

10 A random sample of the records of a certain hospital yielded the following information on the length of hospital stay in days (l_i) and the annual family income (a_i, rounded to the nearest £500) of 15 discharged patients. An analyst believes that the relationship between these two variables is linear. The graph below depicts the scatter plot of the annual family income against the length of stay and the simple linear regression line fitted by the analyst.



Summary statistics for these data are given below:

$$\sum a_i = 82{,}500, \quad \sum a_i^2 = 523{,}750{,}000, \quad \sum a_i l_i = 510{,}500, \quad \sum l_i = 107, \quad \sum l_i^2 = 871.$$

(i) Comment on the relationship between the two variables. [2]

(ii) Determine the equation of the simple regression line. [3]

(iii) Perform an ANOVA test to determine whether the slope of the regression line is significantly different from zero. [4]

(iv) Calculate Pearson's correlation coefficient between the annual family income and the length of hospital stay. [1]

(v) Perform a statistical test to determine whether Pearson's correlation coefficient for the corresponding population is significantly different from -0.8. [5]

(vi) Identify which **one** of the following options gives an approximate 95% confidence interval for Pearson's correlation coefficient for the corresponding population:

A (–2.027, –0.896)
B (–0.966, –0.714)
C (–0.989, –0.683)
D (–0.908, –0.794)

[2]
[Total 17]

END OF PAPER