INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

24 September 2020 (am)

Subject CS1A – Actuarial Statistics
Core Principles

Time allowed: Three hours and fifteen minutes

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

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

1 Let $X_1, X_2, \ldots, X_{81}$ be independent and identically distributed continuous random variables, each with expected value $\mu = E(X_i) = 5$, and variance $\sigma^2 = V(X_i) = 4$.

(i) Determine the sampling distribution of the statistic $T = \sum_{i=1}^{81} X_i$. [2]

(ii) Calculate the probability $P(T > 369)$ using your answer to part (i). [2]

[Total 4]

2 A pair of fair six-sided dice is rolled once.

(i) Identify which **one** of the following options gives the probability that the sum of the two dice is seven:

A1 $\frac{1}{36}$

A2 $\frac{1}{6}$

A3 $\frac{1}{12}$

A4 $\frac{1}{3}$

[2]

(ii) Identify which **one** of the following options gives the probability that at least one dice shows three:

A1 $\frac{25}{36}$

A2 $\frac{1}{36}$

A3 $\frac{11}{36}$

A4 $\frac{5}{36}$

[2]

(iii) Identify which **one** of the following options gives the probability that at least one dice shows an odd number:

A1 $\frac{1}{4}$

A2 $\frac{3}{4}$

A3 $\frac{1}{2}$

A4 $\frac{1}{12}$

[2]

The random variables representing the numbers on the first and second dice are denoted by X and Y respectively.

(iv) (a) Identify which **one** of the following options gives the correct expression of $E(X+Y \mid X=4)$, that is the conditional expectation of the sum of the two dice given that $X=4$:

A1 $E(Y)$
A2 $E(X)+E(Y)$
A3 $4E(X)+E(Y)$
A4 $4+E(Y)$

[1]

(b) State a necessary assumption for deriving the answer in part (iv)(a).

[1]

(c) Determine the value of $E(X+Y \mid X=4)$, using your answer to part (iv)(a). [2]

[Total 10]

3 The following data are available on three television factories that produce all the televisions used in a country.

Factory	*% of total production*	*Probability of defect (Def)*
A	0.35	0.020
B	0.40	0.015
C	0.25	0.010

A television is selected at random and found to have a defect (Def).

(i) Identify which **one** of the following expressions gives the required expression to correctly calculate the probability that the selected television was made in factory B.

A1 $$\frac{\text{P(made in B}\,|\,\text{Def)} \times \text{P(Def)}}{\text{P(made in A}\,|\,\text{Def)P(Def)} + \text{P(made in B}\,|\,\text{Def)P(Def)} + \text{P(made in C}\,|\,\text{Def)P(Def)}}$$

A2 $$\frac{\text{P(Def}\,|\,\text{made in B)} \times \text{P(made in B)}}{\text{P(Def}\,|\,\text{made in A)P(made in A)} + \text{P(Def}\,|\,\text{made in B)P(made in B)} + \text{P(Def}\,|\,\text{made in C)P(made in C)}}$$

A3 $$\frac{\text{P(Def}\,|\,\text{made in B)} + \text{P(made in B)}}{[\text{P(Def}\,|\,\text{made in A)} + \text{P(made in A)}] \times [\text{P(Def}\,|\,\text{made in B)} + \text{P(made in B)}] \times [\text{P(Def}\,|\,\text{made in C)} + \text{P(made in C)}]}$$

A4 $$\frac{\text{P(Def}\,|\,\text{made in B)}}{\text{P(Def}\,|\,\text{made in A)} + \text{P(Def}\,|\,\text{made in B)} + \text{P(Def}\,|\,\text{made in C)}}$$

[2]

(ii) Calculate, by using your answer to part (i), the probability that the selected television was produced by Manufacturer B. [2]

[Total 4]

4 A random variable Y has probability density function

$$f(y) = ae^{-5y}, \quad y > b,$$

where a, b are positive constants.

The moment generating function of Y is denoted by $M_Y(t)$.

(i) Write down the bounds of the integration required to calculate $M_Y(t)$. [1]

(ii) Identify which **one** of the following options gives the correct expression for $M_Y(t)$. [2]

A1 $a\dfrac{e^{-(1-5t)b}}{1-5t}$

A2 $\dfrac{a}{b}\dfrac{e^{-(1-5t)b}}{1-5t}$

A3 $\dfrac{a}{b}\dfrac{e^{-(5-t)b}}{5-t}$

A4 $a\dfrac{e^{-(5-t)b}}{5-t}$

(iii) Write down the condition on t for $M_Y(t)$ to be finite. [1]

(iv) Determine an expression giving the constant a in terms of b, using your answer for $M_Y(t)$ from part (ii). [3]

[Total 7]

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

$$Y_i = a + bX_i + e_i,\ i = 1, ..., n$$

assuming that the error terms e_i are independent and Normally distributed with expectation 0 and variance σ^2. In a sample of size $n = 10$, the following statistics have been observed:

$$\sum_{i=1}^{n} x_i = 141, \quad \sum_{i=1}^{n} y_i = 127,$$

$$\sum_{i=1}^{n} x_i^2 = 2{,}014, \quad \sum_{i=1}^{n} y_i^2 = 1{,}629, \quad \sum_{i=1}^{n} x_i y_i = 1{,}810.$$

(i) Calculate values for S_{xx}, S_{yy}, and S_{xy}. [3]

(ii) Write down, using your answers to part (i), the value of Pearson's correlation coefficient between the variables X_i and Y_i. [1]

(iii) Calculate estimates of the parameters a and b in the regression model. [2]

[Total 6]

6 (i) State the three components of a Generalised Linear Model (GLM). [3]

In a mortality model, the number of deaths D_x at age x is modelled with a GLM. D_x is assumed to have a Poisson distribution with expectation $m_x = \exp(a + bx)$ for each age x, such that $D_x \sim$ Poisson(exp($a + bx$)).

(ii) State the specific form of each of the three components of the GLM for the above mortality model. [3]

(iii) Identify which **one** of the following expressions gives the correct likelihood function as a function of the unknown parameters a and b based on the observed number of deaths for all ages 20 to 80 given by $d_{20},\ldots, d_{80}$, assuming that the numbers of deaths at different ages are independent.

A1 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} \frac{1}{d_x!} e^{-e^{(a+bx)}} e^{(a+bx)d_x}$$

A2 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} e^{e^{(a+bx)}} e^{(a+bx)d_x}$$

A3 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} \frac{1}{d_x!} e^{-e^{(a-bx)}} e^{(a-bx)d_x}$$

A4 $$L(a,b) = \prod_{x=20}^{80} P[D_x = d_x] = \prod_{x=20}^{80} \frac{1}{d_x!} e^{e^{(a+bx)d_x}} e^{-(a+bx)}$$

[2]

An analyst is reviewing the mortality model and is considering deaths only for ages between 40 to 43 inclusive.

The analyst collects data for deaths and estimates the parameters for a and b as follows:

$$d_{40} = 2 \qquad d_{41} = 3 \qquad d_{42} = 1 \qquad d_{43} = 0$$

$$a = 0.01512 \qquad b = -0.00686$$

(iv) Identify, using your answer to part (iii), which **one** of the following options gives the correct value of the likelihood function, based on the analyst's data and parameter estimates.

A1 0.00222
A2 4.05473
A3 0.0008
A4 4.32729

[2]
[Total 10]

7 The probability density function of a Normal distribution is given as follows:

$$f(x; m, s^2) = \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{1}{2s^2}(x-m)^2\right)$$

with $-\infty < x < \infty$, $-\infty < m < \infty$, $s > 0$.

(i) Identify which **one** of the following options gives the correct expression for the exponential family of the density f. [2]

A1 $\frac{1}{\sqrt{2\pi}} \exp\left(\frac{xm - m^2/2}{s^2} - \frac{x^2}{2s^2} - \ln s\right)$

A2 $\exp\left(\frac{xm - m^2/2}{s^2} - \frac{x^2}{2s^2} - \frac{\ln(2\pi s^2)}{2}\right)$

A3 $\exp\left(\frac{x(2m-x)}{2s^2} - \frac{m^2/2}{s^2} - \frac{\ln(2\pi s^2)}{2}\right)$

A4 $\exp\left(\frac{1}{s^2}\left(xm - m^2/2 - \frac{x^2}{2}\right) - \frac{\ln(2\pi s^2)}{2}\right)$

(ii) Identify which **one** of the following options gives the natural parameter θ, the scale parameter ϕ, and the relevant functions $b(\theta)$, $a(\phi)$ and $c(x, \phi)$ of the exponential family for this distribution, using your answer to part (i).

A1 $\theta = m,\ \phi = s^2,\ b(\theta) = m^2,\ a(\phi) = \frac{s^2}{2},\ c(x, \phi) = -\frac{1}{2}\left(x^2 + \ln(2\pi s^2)\right)$

A2 $\theta = m,\ \phi = \frac{s^2}{2},\ b(\theta) = m^2,\ a(\phi) = \frac{s^2}{2},\ c(x, \phi) = -\frac{1}{2}\left(\frac{x^2}{s^s} + \ln(2\pi s^2)\right)$

A3 $\theta = s^2,\ \phi = m,\ b(\theta) = m^2,\ a(\phi) = \frac{s^2}{2},\ c(x, \phi) = -\frac{1}{2}\left(x^2 + \frac{\ln(2\pi s^2)}{2}\right)$

A4 $\theta = m,\ \phi = s^2,\ b(\theta) = \frac{m^2}{2},\ a(\phi) = s^2,\ c(x, \phi) = -\frac{1}{2}\left(\frac{x^2}{s^2} + \ln(2\pi s^2)\right)$

[3]

An analyst found that the mean and standard deviation of this distribution are $E(X) = m$ and $\text{SD}(X) = s^2$. In your answer you may denote θ by theta and ϕ by phi.

(iii) Justify, using the properties of the exponential family, whether or not the analyst is right about the mean and standard deviation of this distribution. [3]

(iv) Contrast a numerical variable and a factor covariate in the context of a generalised linear model. [2]

[Total 10]

8 A statistician has recorded the number of advertising telephone calls that their office received over 2 years. The statistician has recorded data $X_{ij,}$ which is the number of calls received in the ith quarter of the jth year (where $i = 1, 2, 3, 4$ and $j = 1, 2$):

	X_{i1}	X_{i2}	$\bar{X}_i$	$\sum_j (X_{ij} - \bar{X}_i)^2$
$i = 1$	43	29	36	98
$i = 2$	38	42	40	8
$i = 3$	22	18	20	8
$i = 4$	68	56	62	72

(i) Calculate values for:

(a) $E[m(\theta)]$

(b) $E[s^2(\theta)]$

(c) $\text{Var}[m(\theta)]$.

[4]

(ii) Calculate an estimate for X_{13}, the number of advertising telephone calls that the statistician's office expects to receive in the first quarter of year 3, using your answers to part (i) and the assumptions of the Empirical Bayes Credibility Theory Model 1 (EBCT Model 1). [2]

(iii) (a) State two key assumptions underlying the EBCT Model 1.

(b) Explain what these assumptions mean for the data X_{ij} above.

[4]

[Total 10]

9 For an empirical investigation into the amount of rent paid by tenants in a town, data on income X and rent Y have been collected. Data for a total of 300 tenants of one-bedroom flats have been recorded. Assume that X and Y are both Normally distributed with expectations μ_X and μ_Y, and variances σ_X^2 and σ_Y^2. S_X and S_Y are the sample standard deviation for random samples of X and Y, respectively.

The random variable Z_X is defined as

$$Z_X = 299 \frac{S_X^2}{\sigma_X^2}.$$

(i) State the distribution of Z_X and all of its parameters. [2]

(ii) Write down the expectation and variance of Z_X. [2]

(iii) Explain why the distribution of Z_X is approximately Normal. [2]

(iv) Calculate values of an approximate 2.5% quantile and 97.5% quantile of the distribution of Z_X using your answers to parts (ii) and (iii). [3]

In the collected sample, the mean income is \$1,838 with a realised sample standard deviation of \$211, the mean rent is \$608 with a realised sample standard deviation of \$275 and $\Sigma x_i y_i = 348 \times 10^6$.

(v) Calculate a 95% confidence interval for the mean income. [2]

(vi) Calculate a 95% confidence interval for the mean rent. [2]

(vii) Calculate an approximate 95% confidence interval for the variance of income using your answer to part (iv). [2]

(viii) Identify which **one** of the following options gives the correct form of the equation for the simple linear regression model of rent on income, including any assumptions required for statistical inference. [2]

A1 $y_i = a + bx_i$

A2 $y_i = a + bx_i + z_i$ with $E[z_i] = 0$

A3 $y_i = a + bx_i + z_i$ with $z_i \sim \chi^2$, 299 df

A4 $y_i = a + bx_i + z_i$ with $z_i \sim N(0, \sigma^2)$

(ix) Calculate estimates of the slope and the intercept of the model in part (viii) based on the above data for the 300 tenants. [4]

[Total 21]

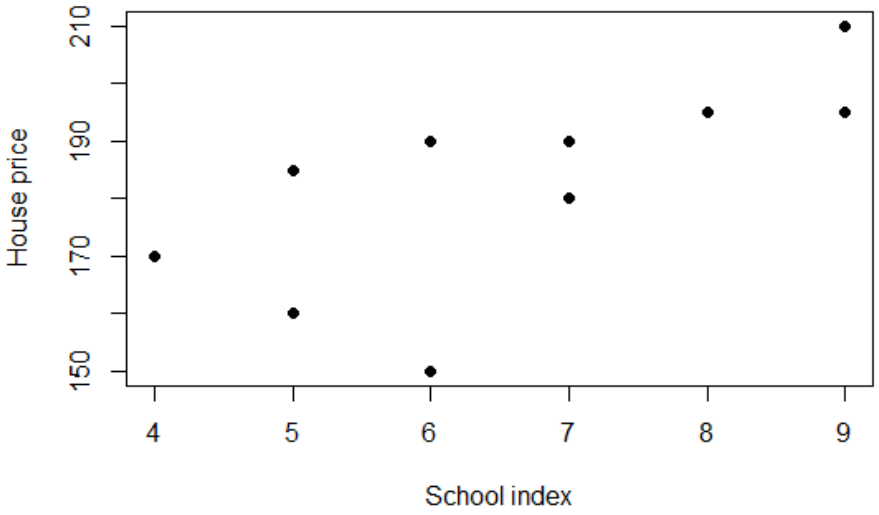
10 It is thought that house prices in certain areas are correlated with the quality of schools in the same areas. A study has been carried out in ten regions where average house prices and school quality indices ranging from 1 (very poor) to 10 (excellent) have been recorded:

Region i	1	2	3	4	5	6	7	8	9	10
School index x_i	9	5	7	6	4	9	7	8	5	6
House prices y_i (£1,000s)	210	185	190	190	170	195	180	195	160	150

$$\sum x_i y_i = 12{,}240; \quad \sum x_i^2 = 462; \quad \sum y_i^2 = 335{,}975.$$

(i) State what is meant by response and explanatory variables in a linear regression. [1]

A plot of the data is given below.



(ii) Comment on the relationship between school quality index and house price, using the plot. [2]

Pearson's correlation coefficient between the data is given as $r = 0.7$.

(iii) A statistical test is performed, using Fisher's transformation, to determine whether Pearson's population correlation coefficient is significantly different from zero, i.e. for

$$H_0: \rho = 0 \text{ vs } H_1: \rho \neq 0.$$

(a) Identify which **one** of the following options gives the correct value of the test statistic for this test:

A1 2.295
A2 6.071
A3 2.743
A4 4.009

[2]

(b) Write down the conclusion of the test at the 5% level of significance, including the relevant critical value(s) from the Actuarial Formulae and Tables. [3]

The linear regression line, of house prices (y) on school index (x), is given as

$$\hat{y} = 133.8 + 7.386x.$$

(iv) A t test is performed to determine if the slope parameter is significantly different from 0.

(a) Identify which **one** of the following options gives the correct values of the sums S_{xx}, S_{yy}, S_{xy} for the house prices (y) and school index (x) data:

A1 $S_{xx} = 32.8$; $S_{yy} = 2{,}415.4$; $S_{xy} = 235$
A2 $S_{xx} = 20.5$; $S_{yy} = 3{,}131.2$; $S_{xy} = 182$
A3 $S_{xx} = 26.4$; $S_{yy} = 2{,}912.5$; $S_{xy} = 195$
A4 $S_{xx} = 35.2$; $S_{yy} = 2{,}817.4$; $S_{xy} = 247$

[2]

(b) Calculate the value of the test statistic. [2]

(c) Write down the distribution of the test statistic, if the null hypothesis of the test is correct. [1]

(d) Write down the conclusion of the test at the 5% level of significance, including the relevant critical value(s) from the Actuarial Formulae and Tables. [3]

(v) Comment on the results in parts (iii)(b) and (iv)(d). [2]

[Total 18]

END OF PAPER