INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2022

Subject CS1A – Actuarial Statistics
Core Principles

Introduction

The Examiners' Report is written by the Chief Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the Specialist Advanced (SA) and Specialist Principles (SP) subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

Sarah Hutchinson
Chair of the Board of Examiners
July 2022

A. General comments on the *aims of this subject and how it is marked*

The aim of the Actuarial Statistics subject is to provide a grounding in mathematical and statistical techniques that are of particular relevance to actuarial work.

Some of the questions in the examination paper accept alternative solutions from those presented in this report, or different ways in which the provided answer can be determined. All mathematically correct and valid alternative solutions or answers received credit as appropriate.

Rounding errors were not penalised. However, candidates may have lost marks where excessive rounding led to significantly different answers.

In cases where the same error was carried forward to later parts of the answer, candidates were given appropriate credit for the later parts.

In questions where comments were required, valid comments that were different from those provided in the solutions also received full credit where appropriate.

The paper included a number of multiple choice questions, where showing working was not required as part of the answer. In all multiple choice questions, the details provided in the answers below (e.g. calculations) are for information.

In all numerical questions that were not multiple-choice, full credit was given for correct answers that also included appropriate workings.

Standard keyboard typing was accepted for mathematical notation.

B. Comments on *candidate performance in this diet of the examination.*

Performance was generally satisfactory, with many candidates showing good understanding of the topics in this subject. Well prepared candidates were able to achieve high scores.

A smaller number of candidates appeared to be inadequately prepared, in terms of not having covered sufficiently the entire breadth of the subject.

Candidates scored lower in questions with unusual style (e.g. Question 4) despite the tested topics being standard basic statistical concepts from the CS1 Core Reading. This highlights the need for candidates to cover the whole syllabus when they revise for the exam and not rely heavily on questions appearing in recent papers.

Candidates are encouraged to practise more on the fundamentals of mathematical calculus and probability. For example, mixed answers in Question 3 suggest that a number of candidates would benefit from additional work on joint and conditional probability, as well as standard integration.

There was an error in multiple choice Question 8 of the paper, where the value of $\sum t_i = 60$ was not specified in the question. As a result, the term "60" appeared as part of the given answers, instead of $\sum t_i$. As this appeared identically in all four possible answers, it did not distinguish any of the given answers. The error was taken into account when marking the question, with the Examiners applying flexibility in awarding full credit where appropriate.

C. Pass Mark

The Pass Mark for this exam was 59.
1311 presented themselves and 579 passed.

Solutions for Subject CS1A – April 2022

Q1

Firstly, to calculate $Var(Y)$ the following result is required:
$Var(Y) = Var[E(Y \mid X)] + E[Var(Y \mid X)]$ [½]

which gives:
$Var(Y) = Var(3X + 11) + E(X + 9)$
$= 3^2 Var(X) + E(X) + 9.$ [1]

Using the information in the question, $X \sim Poi(25)$, $E(X) = 25$ and $Var(X) = 25$. [½]

Therefore:
$Var(Y) = 3^2 \times 25 + 25 + 9$
$= 225 + 34$
$= 259$ [1]

[Total 3]

The question was very well answered.

Q2

(i)
$N(1) \sim$ Poisson(m) , i.e. Poisson (2). [1]

(ii)
$P(N(2) > 7 | N(1) = 5) = P(N(2) - N(1) > 2) = P(N(1) > 2)$ [1]
$= 1 - P(N(1) \leq 2) = 1 - 0.67668 = 0.323$ [1]

(iii)
$P(N(2) > 2 | N(1) = 0) = P(N(1) > 2) = 0.323$ [1]

(iv)

Both probabilities are the same as a direct consequence of the Poisson process i.e. the number of events in the time interval (*s*, *t*] is independent of the number of events up to time *s*. [1]

(v)
The time to the nth event in a Poisson process with rate λ is the sum of *n* individual inter-event times. [½]
Since the inter-event time is an Exp(m) random variable, [½]
the distribution of interest is Gamma(n, m). [1]

(vi)
The CDF of an exponential distribution is given by $F(x) = 1 - e^{-mx}$.
The random number can be generated by solving for x the equation $F(x) = u$. [1]
This gives $x = -\log(1-u)/m$ i.e. x =0.1121972. [1]
[Total 9]

Candidates did not perform well in this question.

In part (iv), alternative answers mentioning the memory-less property of Poisson process were given credit as appropriate

In part (v), only a small number of candidates identified correctly the Gamma(n, m) distribution.

Part (vi) was answered well by the majority of candidates..

Q3
(i)
Answer: D [1]

$$f_X(x) = \int_0^\infty f_{XY}(x,y)\,dy = \int_0^\infty 6e^{-(2x+3y)}\,dy = 6e^{-2x}\int_0^\infty e^{-3y}\,dy$$
$$= 6e^{-2x}\left[-\frac{e^{-3y}}{3}\right]_0^\infty = 2e^{-2x}.$$

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

(ii)
Answer: C [1]

$$f_Y(y) = \int_0^\infty f_{XY}(x,y)\,dx$$
$$= \int_0^\infty 6e^{-(2x+3y)}\,dx = 6e^{-3y}\int_0^\infty e^{-2x}\,dx = 6e^{-3y}\left[-\frac{e^{-2x}}{2}\right]_0^\infty = 3e^{-3y}.$$

Therefore,

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

(iii)

$f_{XY}(x, y) = f_X(x) \times f_Y(y)$ for all x and y. [½]

Therefore, X and Y are independent. [½]

(iv)

Since X and Y are independent, $E[Y|X > 2] = E[Y]$. [1]

$Y \sim \text{Exp}(3)$. Therefore $E[Y|X > 2] = \frac{1}{3}$. [1]

(v)

Answer: B [2]

$$P(X > Y) = \int_0^\infty \int_y^\infty 6e^{-(2x+3y)}\, dx\, dy = \int_0^\infty 6e^{-3y} \left(\int_y^\infty e^{-2x} dx \right) dy$$

$$= \int_0^\infty 6e^{-3y} \left[-\frac{e^{-2x}}{2} \right]_y^\infty dy = \int_0^\infty 3e^{-5y}\, dy = \frac{3}{5}$$

[Total 7]

Performance in this question was mixed. Candidates are recommended to practice the mechanics of integration.

Q4

(i)(a)

A random sample is made up of independent [½]

and identically distributed random variables, [½]

typically denoted as $X_1, \ldots, X_n$.

(i)(b)

A statistic is a function of random variables. [1]

It will be a random variable itself and will have a distribution, its sampling distribution. [½]

A statistic does not involve any unknown parameters. [½]

(ii)(a)

A suitable population in this case is the set of all voters. [1]

In terms of the random variable X it will consist of a series of 1s and 0s depending on whether an individual voter would or would not support the new party.

The parameter of interest is p, representing the proportion of 1s in the population, [1]

i.e. the proportion of voters in the population that support the party.

(ii)(b)

Y is the number of voters who would support the party.

Since the sample is random, therefore each observation is independent, p is constant and the responses are either success (1) or failure (0), then [1]
Y will have a binomial distribution with $n = 50$ and parameter p , i.e. $Y \sim Bin(50\,, p)$. [1]
[Total 7]

A number of candidates found the question challenging.

The question examines basic statistical concepts from the CS1 Core Reading.
In part (ii)(b), alternative answers referring to the sum of n independent Bernoulli trials received credit as appropriate.

Q5

(i)
Y_i are independent random variables with only two outcomes. [½]
The two outcomes are 0 and 1 with 1 having a "success" probability of
$p = 1 - e^{-m}$. [½]

This is the definition of the Bernoulli(p) distribution.

(ii)
Answer: C [2]

The likelihood function, in terms of y_i, i =1, 2, …, n, is given as

$$L(m) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} (1 - e^{-m})^{y_i} (e^{-m})^{1-y_i}$$
$$= (1 - e^{-m})^{\sum y_i} (e^{-m})^{n - \sum y_i}$$
$$= (1 - e^{-m})^{n\bar{y}} (e^{-m})^{n - n\bar{y}}$$

(iii)
Taking logarithms we have:

$$l(m) = \sum y_i \log(1 - e^{-m}) - m(n - \sum y_i)$$ [1]

and
$$\frac{dl}{dm} = \frac{e^{-m} \sum y_i}{1 - e^{-m}} - n + \sum y_i$$ [1]

The MLE, in terms of y_i, will be given by

$$\frac{dl}{dm} = 0 \Rightarrow e^{-\hat{m}} \sum y_i - n\left(1 - e^{-\hat{m}}\right) + \sum y_i \left(1 - e^{-\hat{m}}\right) = 0$$ [1]

$$\Rightarrow e^{-\hat{m}} = 1 - \bar{y} \Rightarrow \hat{m} = -\log(1 - \bar{y})$$ [1]
where $\bar{y} = \frac{\sum y_i}{n}$.

(iv)

The second derivative of L evaluated at $\hat{m}$ must be strictly negative, that is

$\frac{\partial^2}{\partial m^2} L(\hat{m}, y_1, \ldots, y_n) < 0$ [1]

[Total 8]

Candidates generally answered this question well.

In part (iv) alternative answers in terms of the log likelihood were given credit as appropriate.

Q6

(i)

Answer: D [2]

The likelihood of the model is given by

$$L(\alpha, \beta) = \prod_{i=1}^{n} f(x_i, \alpha, \beta) = \prod_{i=1}^{n} \alpha \frac{\beta^\alpha}{x_i^{\alpha+1}} = \alpha^n \beta^{\alpha n} \prod_{i=1}^{n} \frac{1}{x_i^{\alpha+1}}.$$

Taking the log leads to

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

(ii)

Differentiating the log-likelihood with respect to α gives

$\frac{\partial l(\alpha,\beta)}{\partial \alpha} = \frac{n}{\alpha} + n \log \beta - \sum_{i=1}^{n} \log x_i.$ [1]

Hence the MLE of α is

$\hat{\alpha} = \frac{n}{-n \log \beta + \sum_{i=1}^{n} \log x_i}.$ [1]

(iii)

The PDF increases as β increases. [½]

Also, the support of the PDF (i.e. $\{x: x \geq \beta\}$) moves to the right. [½]

(iv)

The increase in the PDF when β increases implies that the likelihood is higher for higher values of β. This means that the likelihood is maximised for the highest value of β. Since $\beta \leq x_i$ for all i, [1]

the MLE of β is the smallest value of x_i. [1]

(v)

meanlog $= \sum_{i}^{10} \frac{\log x_i}{n} = 9.508.$ [1]

$$\text{sdlog} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\log x_i - \text{meanlog})^2} = 0.476$$ [1]

(vi)
$\hat{\beta}$ is the smallest value of x_i: $\hat{\beta} = 5000$ [1]

$$\hat{\alpha} = \frac{10}{-10\times\log 5000 + \sum_{i=1}^{10}\log x_i} = 1.009$$ [1]

[Total 11]

Candidates generally answered this question well.

In part (iv) many candidates failed to determine the MLE of beta. Credit was given for partial answers referring to standard MLE methods not working here.

Q7
(i)
Using the properties of the exponential family, the derivative w.r.t. θ of the $b(\theta)$ function gives the mean:

Mean $= b'(\theta)$ [½]
$= -1/\theta$ i.e. $b'(\theta) = \mu$. [½]

The second derivative w.r.t. θ of the $b(\theta)$ function multiplied by the $a(\phi)$ function gives the variance:

Variance $= b''(\theta)a(\phi)$ [½]
$= (1/\theta^2)(\sigma^2/\mu^2)$ [1]
$= \mu^2 \times \sigma^2/\mu^2 = \sigma^2$. [½]

(ii)
The gamma distribution is more suitable because claim sizes are always positive [1]
and their distribution is usually non-symmetrical. [1]

(iii)
The model output suggests that:

parameter estimate > 2*(standard error) since 0.06084 > 2*0.00296 = 0.00592. [1]

We conclude that the covariate operational time is significant. [1]

(iv)
The variable "legal representation" is a factor that takes a categorical value (yes/no), [1]
while the operational time is a continuous covariate (or a variable taking a numerical value). [1]

(v)
The deviance improved significantly when using the legal representation as a second covariate. [1]

Therefore, legal representation is a significant covariate of the claim sizes and Model 2 is preferred. [1]

[Total 11]

Most candidates answered well this question.

In part (ii) many candidates failed to make the point of the suitability of the Gamma distribution due to claim sizes being positive.

In part (v) stating a significant change in the deviance is important for receiving full marks.

Also in part (v), correct answers included reference to the p-value or the appropriate quantile involving the χ^2 statistic and χ^2 distribution with 1 degree of freedom.

Q8

(i)
Answer: A [1]

$L(\lambda|T) = \lambda^n \exp(-\lambda\sum t_i)$ where t_i are the observed times between the arrival of two lorries.

(ii)
Density of gamma distribution with parameters α and β: $f(\lambda) = C\lambda^{\alpha-1}e^{-\beta\lambda}$ [1]

Posterior distribution:
$f(\lambda|T) \propto f(\lambda)L(\lambda|T)$
$\propto \lambda^{\alpha-1}e^{-\beta\lambda}\lambda^n \exp(-\lambda\sum t_i)$ [1]
$= \lambda^{\alpha+n-1}\exp[-\lambda(\beta + \sum t_i)]$. [1]

We recognise this as the density function of a gamma distribution
with parameters $\alpha + n$ and $\beta + \sum t_i$. [1]

(iii)
Under quadratic loss, the Bayesian estimator is the mean of the posterior distribution. [1]

In this case the mean is $\frac{\alpha+20}{\beta+\sum t_i}$. [1]

(iv)
Under all-or-nothing loss, the Bayesian estimator is the mode of the posterior [1]
To find the mode we need to maximise the density. [1]
To maximise the density, we need to differentiate the density or log-density w.r.t λ, set this expression to zero and solve for λ. [1]

(v)
Answer: B [2]

$$f'(\lambda|T) = (\alpha + n - 1)\lambda^{\alpha+n-2} \exp[-\lambda(\beta + \sum t_i)] + \lambda^{\alpha+n-1} \exp[-\lambda(\beta + \sum t_i)]\big(-(\beta + \sum t_i)\big) = 0$$

$$(\alpha + n - 1) - \lambda(\beta + \sum t_i) = 0$$

Solve for λ: $\lambda = \frac{\alpha+n-1}{\beta+\sum t_i}$.
And for the given sample we obtain: $\lambda = \frac{\alpha+19}{\beta+60}$

(vi)
For fixed *n*=20, the estimator in part (iii) will give a higher value than that in part (v) as the numerator is bigger. The difference will be affected by the denominator. [1]

[Total 13]

Candidates answered well this question.

In part (iv) candidates who quoted the mode of the gamma distribution received credit as appropriate.

For part (v) see also the comment in Section B, regarding $\sum t_i = 60$ not being specified in the question. In part (vi) alternative answers were given credit, including answers mentioning that the two estimators will give almost identical estimated value for large sample size n.

Q9
(i)
As it stands the model cannot be used for inference. [1]
We need further assumptions: [½]
the errors e_i are independent [½]
and $e_i \sim N(0, \sigma^2)$. [1]

(ii)
We have $S_{yy} = SS_{TOT}$ [1]

and $\frac{S_{xy}^2}{S_{xx}} = SS_{REG}$ [1]

So, $R^2 = \frac{S_{xy}^2}{S_{xx}\, S_{yy}} = \frac{SS_{REG}}{SS_{TOT}}$ [1]

This verifies that R^2 gives the proportion of the total variability of *Y* "explained" by the linear regression.

(iii)
Possible approaches:
Use R^2 [1]
Use adjusted R^2 [1]
Plot residuals against fitted values (or explanatory variables) [1]

(iv)
R^2cannot be used [1]
Although increased values show a better fit of the model, [1]
it cannot decrease as more explanatory variables are added to the model. [1]

So, we do not have a "stopping rule" for the process.

(v)
The adjusted R^2 should be used as a criterion. [1]

So, the model maximising the adjusted R^2 has explanatory variables $X_1 + X_4 + X_3$. [1]

[Total 14]

Candidates answered well this question.

A common error in part (i) was the omission of the assumptions required for the error terms.

A number of alternative answers received credit, including the following:
Part (i):answers worded in terms of the model being suitable for inference, provided that the assumptions are satisfied.
Part (iii): answers mentioning QQ-plot for residuals.
Part (iv):answers mentioning that R^2 does not take into account the number of covariates/complexity of the model.

Q10
(i)
There appears to exist a strong, linear relationship between annual family income and length of stay. [1]
The length of stay decreases with the Annual family income (negative association). [½]

At some point this relationship cannot hold since length of stay is positive and annual family income is not bounded above. [½]

(ii)
Start by calculating the sum of squares
$S_{aa} = 523{,}750{,}000 - \frac{82{,}500^2}{15} = 70{,}000{,}000$ [½]

$S_{al} = 510{,}500 - \frac{82{,}500 * 107}{15} = -78{,}000$ [½]
$\hat{\beta} = \frac{S_{al}}{S_{aa}} = -0.001$ [½]

$\hat{\alpha} = \bar{l} - \hat{\beta}\bar{a} = \frac{107}{15} + 0.001 * \frac{82500}{15} = 12.63$ [½]

Hence, the fitted regression equation of l on a is:
$\hat{l} = 12.63 - 0.001a$ [1]

(iii)
$H_0 : \beta = 0 \quad vs \quad H_1 : \beta \neq 0$ [½]
$SS_{TOT} = 871 - \frac{107^2}{15} = 107.733$ [½]
$SS_{REG} = \frac{78{,}000^2}{70{,}000{,}000} = 86.914$ [½]
$SS_{RES} = 107.733 - \frac{78{,}000^2}{70{,}000{,}000} = 20.819$ [½]

Under $H_0 : \beta = 0$, we have:

$F = \frac{\frac{SS_{REG}}{1}}{\frac{SS_{RES}}{13}} = 54.27$ on (1,13) degrees of freedom [1]

This is a significant result which exceeds the 0.01 critical value of $F_{1,13} = 9.074$. So, there is sufficient evidence at the 0.01 level to reject H_0 in favour of $\beta \neq 0$. [1]

(iv)
Correlation coefficient ρ
$$\rho(A,L) = \frac{S_{al}}{(S_{aa}S_{ll})^{1/2}} = \frac{-78000}{\sqrt{(70{,}000{,}000 * 107.733)}} = -0.898$$ [1]

(v)
Hypotheses:
$H_0 : \rho = -0.8 \quad vs \quad H_1 : \rho \neq -0.8$ [½]

Under H_0, the test statistic Z_r has a $N(Z_\rho, \frac{1}{\sqrt{12}})$ distribution, where:
$Z_\rho = \frac{1}{2} log \frac{1-0.8}{1+0.8} = -1.098612 \approx -1.099.$ [½]

The observed value of this statistic is:
$Z_r = \frac{1}{2} log \frac{1-0.898}{1+0.898} = -1.461792 \approx -1.462.$ [½]

These correspond to a value of the test statistic: $\frac{-1.462+1.099}{\sqrt{\frac{1}{12}}} = -1.257$ [1]

which under the null hypothesis should be a value from the $N(0,1)$ distribution. [½]

The absolute value is less than 1.96, the upper 2.5% point of the standard normal distribution. [1]

So, there is insufficient evidence to reject H_0 at the 5% level, i.e. the data do not provide enough evidence to conclude that the correlation parameter is different from -0.8. [1]

(vi)
Answer: B [2]

95% confidence interval

$$Z_r \pm Z_{\frac{\alpha}{2}} * \left(\frac{1}{\sqrt{n-3}}\right)$$

Then:

$$-1.462 \pm 1.96 * \left(\frac{1}{\sqrt{15-3}}\right)$$

$$-1.462 \pm 0.566$$

$$Z_r \in [-2.0278, -0.8962]$$

Given that:

$$\frac{1}{2} ln \frac{1+r}{1-r} = Z_r$$

$$r = \frac{e^{2Z_r} - 1}{1 + e^{2Z_r}}$$

Converting these limits which are values of Z_r into values of r gives:

$Z_r = -2.0278$ then $r = -0.966$

$Z_r = -0.8962$ then $r = -0.714$

Therefore, the 95% confidence interval for ρ is $[-0.966, -0.714]$.

[Total 17]

Candidates overall answered well this question.

In Part (iii) an ANOVA test is required; a small number of candidates attempted different tests and received partial credit where appropriate. Part (vi) was answered correctly by candidates who seemed to be well prepared.

A number of alternative answers received credit, including the following:
Part (ii) and throughout the question: work using a higher number of decimal places resulting in answers of varying accuracy – these received full credit where appropriate.
Part (v): answers using the inverse hyperbolic tangent function and/or referring to the p-value of the test.

[Paper Total 100]

END OF EXAMINERS' REPORT

