INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS’ REPORT

September 2021

CS1 – Actuarial Statistics
Core Principles
Paper A

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 later 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
December 2021

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 satisfactory in general, with many candidates showing good understanding of the topics in this subject. Well prepared candidates were able to score highly.

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

Questions corresponding to parts of the syllabus that are not frequently examined were generally poorly answered (e.g. Question 2, parts of Question 8). This highlights the need for candidates to cover the whole syllabus when they revise for the exam and not only rely on themes appearing in past papers.

C. Pass Mark

The Pass Mark for this exam was 58
1372 presented themselves and 578 passed.

Solutions for Subject CS1A – September 2021

Q1

(i)

$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, [½]

$n = 15, \sigma^2 = 2$ so $7S^2 \sim \chi^2_{14}$. [1½]

(ii)

The underlying sample is from the Normal distribution, hence the chi-squared distributional assumption for the sample variance holds true [1]

[Total 3]

Generally well answered.
(i) Common errors included candidates using the wrong expression.
(ii) A number of candidates did not answer this part. Those who answered, did so well.

Q2

(i)

$t_1 = \frac{Z}{\sqrt{Y}}$, where $Z \sim N(0,1)$ and $Y \sim \chi_1^2$ are independent [1]

Simulate Z_1, Z_2 from N(0,1) independently, [½]
then $Z_2^2 \sim \chi_1^2$. [½]
So $\frac{Z_1}{\sqrt{Z_2^2}} = \frac{Z_1}{Z_2} \sim t_1$ [1]

(ii)

Simulate iid $Z_1, Z_2, Z_3 \sim N(0,1)$, so that $Z_1^2 + Z_2^2 + Z_3^2 \sim \chi_3^2$. [2]
This is the same as a Gamma (3/2, 1/2) distribution [1]

(iii)

Simulate iid $Z_1, Z_2 \sim N(0,1)$, so that $Z_1^2,\ Z_2^2 \sim \chi_1^2$ independently [1]
Then $\frac{Z_1^2}{Z_2^2} \sim F_{1,1}$ [1]

[Total 8]

This question was not well answered, with many candidates not attempting it.
In many cases candidates attempted to provide answers using incorrect (or not sufficiently explained) references to the inverse CDF method. Notice that the inverse CDF method is not directly applicable here.

Q3

The mean and variance of the distribution are given by

$E\,[X] = \frac{b}{a-1} = \frac{6}{4-1} = 2$ [½]

$Var\,[X] = \frac{ab^2}{(a-1)^2(a-2)} = \frac{4(6)^2}{(4-1)^2(4-2)} = 8$ [½]

$Var\,(Y) = Var\,[E\,(Y \mid X)] + E\,[Var\,(Y \mid X)]$, so

$Var\,(Y) = Var\,[3X + 6\,] + E\,[X^2 + 4]$ [1]

$\quad = 9\,Var\,[X] + E\,[X^2] + 4$ [1]

Also $E\,[X^2] = Var[X] + (E\,[X])^2 = 8 + 2^2 = 12$ [1]

So, $Var\,(Y) = 9\,(8) + 12 + 4 = 88$ [1]

The standard deviations is $\sqrt{88} = 9.381$ [1]

Generally answered very well.
Common issues involved not providing the standard deviation and calculation errors.
Also, some candidates did not provide sufficient intermediate steps and this may have impacted partial credit given.

Q4

(i)

A gamma distribution with mean 35 and standard deviation 5 has the following parameters:
$\frac{\alpha}{\lambda} = 35$ and $\frac{\alpha}{\lambda^2} = 25$

So: $\alpha = 25\lambda^2$ and $\alpha = 35\lambda$
Solving these equations gives: $\alpha = 49$ and $\lambda = 1.4$ [1]

So the prior distribution of m is $\text{Gamma}(49, 1.4)$ [1]
The prior PDF of m is therefore: $f_{prior}(m) \propto m^{48}e^{-1.4m}$ [1]

(ii)

Answer: D [3]
Likelihood function L $\propto e^{-7m}m^{135}$
The posterior PDF of m is given by:
$f_{posterior}(m) \propto f_{prior}(m) \times Likelihood\ function$
So $f_{posterior}(m) \propto m^{48}e^{-1.4m} \times e^{-7m}m^{135} = m^{183}e^{-8.4m}$

(iii)

Under all or nothing loss, the Bayesian estimate is given by the mode of this $\text{Gamma}(184, 8.4)$ distribution, which can be obtained by finding the value of m that maximises the PDF [1]
Finding the maximum:
$\frac{d}{dm}\left(\log(f_{posterior}(m))\right) = \frac{d}{dm}(183\log m - 8.4m) = \frac{183}{m} - 8.4$ [1]
Setting equal to zero gives $m = \frac{183}{8.4} = 21.786$ [2]

(iv)

Correctly identify mean of gamma posterior distribution as: [1]
$\frac{\alpha}{\lambda} = \frac{184}{8.4} = 21.905$ [1]

[Total 12]

The question was well answered.
Please note that in all multiple choice questions, the details provided in the answers (e.g. calculations) are for information. Candidates were not required to show working.

Q5

(i)

X = number of policies where a claim is made, so
$X \sim Bin(500, 0.06)$ [1]

Use the Normal approximation: $X \;\dot{\sim}\; \mathrm{N}(30, 28.2)$, [2]
as n is sufficiently large

$P(X \leq 40) = P(X \leq 40.5)$ using continuity correction [1]
$= \Phi\left(\frac{40.5-30}{\sqrt{28.2}}\right) = \Phi\left(\frac{10.5}{\sqrt{28.2}}\right) = \Phi(1.97726) = 0.97599$ [1]

(ii)

The 95% confidence interval for the mean claim amount is: $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ [1]
$1.96\frac{\sigma}{\sqrt{n}} = 5$ for a total confidence width of £10 [1]
Solve for n, using $\sigma = £75$, gives n = 864.36, i.e. sample size of 865 [2]
[Total 9]

Generally answered well.
A common issue in part (i) was not using or applying a continuity correction correctly.

Q6

(i)

Answer: C [2]

First derive the cdf of Y as

$$F(y) = \int_0^y 2ct \exp(-ct^2)\, dt = [-\exp(-ct^2)]_0^y$$
$$= 1 - \exp(-cy^2),\; y > 0.$$

So, $y = F^{-1}(u) = \left\{-\frac{1}{c}\log(1-u)\right\}^{1/2}$

(ii)

To generate values of Y:

1. Generate a random variate u from U(0, 1) [½]
2. Return $y = \left\{-\frac{1}{c}\log(1-u)\right\}^{1/2}$ [1½]

(iii)

Answer: B [2]

$$p(c \mid y) \propto \pi(c)L(c; y) \propto c^{a-1}e^{-cb}(2c)^n \prod_i y_i e^{-cy_i^2}$$
$$\propto c^{n+a-1} \exp\{-(b + \textstyle\sum_{i=1}^n y_i^2)c\}$$

(iv)
This is the density of a gamma distribution [1]
with parameters $n + a$ and $b + \sum_{i=1}^{n} y_i^2$ [1]
[Total 8]

> *There were mixed answers here, with a number of candidates not attempting parts of the question.*
> *In part (iv) some candidates failed to identify the parameters of the gamma distribution correctly.*

Q7
(i)
Since X_i are independent, we have that $Y = \sum_i^n X_i$ follows a gamma distribution with parameters n and b [1]
So MGF is given by $M_Y(t) = \left(1 - {}^{t}\!/_{b}\right)^{-n}$ [1]

(ii)
$M_z(t) = \sqrt{M_Y(t)} = \left(1 - {}^{t}\!/_{b}\right)^{-n/2}$ [½]
The MGF of a chi-square distribution with n degrees of freedom
is $(1 - 2t)^{-n/2}$ [½]

So $M_z(t)$ is the MGF of a chi-square distribution with n degrees of freedom [1]
and $b = 0.5$ [1]
[Total 5]

> *There were mixed answers in this question, often with unclear justification.*
> *In part (i) reference to independence of the variables is required to fully justify the answer and obtain full marks.*

Q8
(i)
Y follows a Poisson distribution [1]
$\log(\mu) = \alpha_i + \beta_i X_1$; where $i = 1,2,3$ for low, medium and high pollutant respectively [1]
$\mu = E(Y)$ [1]

Alternative forms for the linear predictor:
The linear predictor above can also be written as:
$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_{2,i} + \beta_{3,i} X_1$; where $i = 2,3$ for medium and high pollutant
Or, a model without the interaction term can be given
$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_{2,i}$; where $i = 2,3$ for medium and high pollutant

(ii)
$\alpha_i,\ i = 1,2,3$ are the coefficients of the main effect for pollutant concentration [1]

We may also need the interaction term $\beta_i X_1$ if the effect of temperature on number of hospitalisations is different for each level of pollutant concentration [1]

Under alternative forms for the linear predictor in (i):
β_0 is the intercept
β_1 is the coefficient for the main effect for temperature
$\beta_{2,i}$ the coefficients of the main effect for pollutant concentration where $i = 2{,}3$ for medium and high pollutant
$\beta_{3,i}$ the coefficients of the effect of temperature on number of hospitalizations where $i = 2{,}3$ for medium and high pollutant

(iii)
$\log(\mu) = -0.372 + 0.09 \times 19 + 0.298 - 0.076 \times 19$ [1]

(iv)
These are not listed as X_2Low is used as the reference category [1]
or, equivalently, their effect is included in the intercept estimate

(v)
Medium concentration has no significant effect, as compared to low concentration, [1]
while high concentration has a significant increasing effect for the number of hospital admissions [1]

Alternative comments include:
The sign of $X_1{:}X_2High$ suggests that temperature becomes less important when pollutant concentration is High (but 0.09-0.076 is still positive)

[Total 9]

Again, the quality of answers given here was mixed.
In part (i) there was no mention of the distribution in many cases.
Parts (iii), (iv) were well answered for candidates that attempted them.

Q9

(i)(a)

Answer: A [2]

$n = 110$
$S_{xx} = Var(x)\,(n-1) = 261.881^2 \times 109 = 7475401$
$S_{yy} = Var(y)\,(n-1) = 0.824^2 \times 109 = 74.008$

$$\hat{b} = r\sqrt{\frac{S_{yy}}{S_{xx}}} = -0.0175\,\frac{0.824}{261.881} = -5.506 \times 10^{-5}$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 0.106 + 5.506 \times 10^{-5} \times 134.487 = 0.113$$

(b)
The fitted return for a firm with $x = 95.55$ is
$y^* = 0.113 - 5.506 \times 10^{-5} \times 95.55 = 0.108$ [1]

(ii)(a)
Using the logarithmic regression,
$y^* = 0.438 - 0.090 \times \ log(95.55) = 0.028$ [1]

(b)
The return estimated with the log revenue is different from the return in part (i)(b) as expected [1]

(c)
$S_{zz} = Var(z)\ (n-1) = 1.698^2 \times 109 = 314.269$ [½]
$S_{zy} = \beta S_{zz} = -0.09 \times 314.269 = -28.284$ [½]

(iii)
$H_0 : \beta = 0\ vs\ H_1 : \beta \neq 0$ [½]

$\hat{\sigma}^2 = \frac{1}{108}\left(74.008\ -\ \frac{28.284^2}{314.269}\right) = 0.662$ [1]
s.e.$\left(\hat{\beta}\right) = (\hat{\sigma}^2/S_{zz})^{1/2} =\ (0.662/314.269)^{1/2} = 0.046$ [½]

Test statistic $=\ -0.09/0.046 = -1.956$ [1]

The test statistic follows a t-distribution with 108 df under the null hypothesis [½]
This is a two-sided test with the 5% critical value being -1.658 for 120 df
(-1.661 using linear interpolation and -1.659 using R) [½]

We have evidence at 10% significance level to reject the null hypothesis that
$\beta = 0$ and we conclude that the logarithmic revenues affect returns [1]

(iv)
$H_0 : \beta = 0\ vs\ H_1 : \beta > 0$ [1]

From (iii), the test statistic is $-\ 1.956$
The test statistic follows a t-distribution with 108 df under the null hypothesis
This is a one-sided test with the 10% critical value approximating 1.289 for
120 df (1.290 using linear interpolation and 1.289 using R) [1]

We do not have evidence to reject H_0 at 10% significance level. Firms with greater revenues do not necessary enjoy a larger return [1]

(v)
$r = \frac{S_{zy}}{(S_{zz}S_{yy})^{1/2}} = \frac{-28.284}{(314.269 \times 74.008)^{1/2}} = -0.185$ [1]

(vi)(a)
$z = 2$, the estimated percentage return is
$y^* = 0.438 - 0.09 \times 2 = 0.258$ [1]

(b)

$Se(y^*) = \left[\left\{1 + \frac{1}{110} + \frac{(2-3.686)^2}{314.269}\right\}\ 0.662\ \right]^{\frac{1}{2}} = 0.821$ [1½]

Confidence interval: $0.258 \pm 1.98 \times 0.821$ i.e. $(-1.367\,, 1.883)$ [1½]
if approximating the percentage points for $t_{108}\ to\ t_{120}$.

(vii)(a)
The expected return is $y^{*} = 0.438 - 0.09\ \times\ 1.76 = 0.28$. [1]
The residual is $\tilde{e} = 4.333 - 0.28 = 4.053$. [1]

(b)
The residual is way above 0 and from the table the percentage return is 3 times the median [1]

Alternative:
This observation seems to be an outlier. Or, the residual appears large given the size of the sample SD of the *y* data

[Total 22]

Generally well answered.
Some common issues included:
(ii)(b) Attention to detail was required here, often candidates made inconsistent comments.
(ii)(c), (iii): Calculation errors.
(iv) Using a two-tailed test was a common error here.
(v) A number of candidates showed lack of understanding where incorrect values for Szy, Szz led to an obviously incorrect Pearson's correlation coefficient, i.e. r < -1 or r > 1.
In parts (iii), (iv), (vi) full credit was given for using alternative critical point values, including values resulting from linear interpolation, extracted form R, or the appropriate critical points of the standard normal distribution with justification – i.e. high df.

Q10
(i)
Answer: C [2]

$$l(\mu) = \log\left(\prod_{i=1}^{n} f(x_i|\mu)\right) = \sum_{i=1}^{n} \log f(x_i|\mu)$$
$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Therefore
$$\frac{dl(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

(ii)
From part (i):

$$\frac{dl(\mu)}{d\mu} = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

Therefore,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$\hat{\mu} = 140{,}000$ [1]

(iii)
Given that $\hat{\mu}$ is the sample mean,
$\hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ [1½]

Confidence interval:
$\hat{\mu} \pm Z_{0.025}\sqrt{\sigma^2/n}$ [1]

$140{,}000 - 1.96\frac{12{,}000}{\sqrt{5}} \leq \mu \leq 140{,}000 + 1.96\frac{12{,}000}{\sqrt{5}}$ [1]

95% CI: $(129{,}481.54\,,\ 150{,}518.46)$ [½]

(iv)
The posterior distribution is a normal distribution with mean: [1]

$\frac{n\tau\bar{x}+\tau_0\mu_0}{n\tau+\tau_0} = 142166.7$ [½]

and variance:
$1/(n\tau + \tau_0) = 4749.77^2$ = 22,560,315 [½]

Hence,
$\hat{\mu} \sim \mathcal{N}(142166.7\,, 4749.77^2)$

(v)
The prior and the posterior distribution are of the same type [½]
The normal distribution is the conjugate prior for the mean of a normal distribution [½]

(vi)
Bayesian credible estimate for μ under quadratic loss is the expectation of the posterior distribution: [1]
$\tilde{\mu} = 142166.67$ [1]

(vii)
$\tilde{\mu} \sim N(142166.67, 4749.77^2)$, therefore the Bayesian interval is

$\left(142{,}166.67 - 1.96\sqrt{4749.77^2}\,, 142{,}166.67 + 1.96\sqrt{4749.77^2}\right)$ [1½]
i.e. $(132857.1, 151476.2)$. [½]

(viii)
The Bayesian interval is different (narrower) than the CI of the MLE [½]
The prior belief has impacted on the estimation of the posterior [½]

(ix)
Answer: B [3]
Given that the prior density of the uniform distribution $f(\mu)$ does not depend on μ, we have:

$$p(\mu|\underline{x}\,) \propto L(\mu)f(\mu)$$
$$\propto \exp\left[-\tfrac{1}{2\sigma^2}\left(\textstyle\sum_{i=1}^{n}(x_i-\mu)^2\right)\right]$$
$$\propto \exp\left[-\tfrac{1}{2\sigma^2}\left(\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2 + 2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) + \mathrm{n}(\bar{x}-\mu)^2\,\right)\right]$$
$$\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right]$$

Since
$\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) = (\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) = 0$
and
$\sum_{i=1}^{n}(x_i-\bar{x})^2$ does not depends on μ.

[Total 18]

Very well answered.
Comments in part (viii) varied, with many candidates failing to mention the impact of the prior distribution.

[Paper Total 100]

END OF EXAMINERS' REPORT