INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2021

Subject 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.

Paul Nicholas
Chair of the Board of Examiners
July 2021

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

1. 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.

2. Some of the questions in the examination paper admit alternative solutions from these 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.

3. Rounding errors were not penalised, but candidates lost marks where excessive rounding led to significantly different answers.

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

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

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

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

9. Standard keyboard typing was accepted for mathematical notation.

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

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

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

3. Questions that required higher order skills and comments were generally not well answered (e.g. Q1(iii)(b), Q8(v),(vi)).

4. Questions corresponding to parts of the syllabus that are not frequently examined were generally poorly answered (e.g. Q5). 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.

5. There was a typing error in Q5(v) of the paper, where the correct answer should be shown as $\int_0^\infty 2te^{-2t}dt + \int_0^\infty 6e^{-2t}dt$. This is also related to the answers in parts (vii) and (viii) of the question. The error was taken into account when marking the

question, with the Examiners applying flexibility in awarding credit where appropriate. The pass mark for this exam was adjusted accordingly, to reflect the marks that affected candidates might not have had the opportunity to score. The Examiners did not find any evidence that the error had any further impact on performance on the remainder of the paper.

C. Pass Mark

The Pass Mark for this exam was 56.
1,482 candidates presented themselves and 779 passed.

Solutions for Subject CS1 Paper A April 2021

Q1

(i)

X ~ Gamma(50, 0.25), and using X ~ Gamma($\alpha,\ \lambda$) $\Rightarrow 2\lambda X \sim \chi^2_{2\alpha}$: [1]

P(X > 270) = P($2\lambda X > 2\lambda \times 270$) = P($0.5X > 135$) = P($\chi^2_{100} > 135$) ≈ 0.01 [1]
(using the Actuarial Tables for chi-square probabilities)

(ii)

The mean and variance of the given gamma distribution are

$E(X) = \frac{\alpha}{\lambda} = \frac{50}{0.25} = 200,\ var(X) = \frac{\alpha}{\lambda^2} = \frac{50}{0.25^2} = 800$ [1]

Using the normal approximation for large α, X ~ Gamma(50, 0.25) can be approximated as
X ~ N(200, 800): [1]

P(X > 270) = $P = \left(Z > \frac{270-200}{\sqrt{800}}\right)$ = P(Z > 2.4749) = 1 – P(Z < 2.4749) = 0.0066642
Alternatively, use tables to interpolate, giving 0.00667
[2]

(iii)

The gamma distribution converges to the normal distribution as $\alpha \to \infty$. [1]
But for $\alpha = 50$, the gamma distribution exhibits positive skew, [½]
and gives a higher tail probability than the symmetric normal distribution [½]
[Total 8]

Generally well answered. In part (i) some candidates did not calculate the probability using the chi-square distribution, as the question asked. In (iii) a number of candidates did not provide any comments.

Q2

(i) $E[U] = E\left[E[Y|X]\right] = E[Y] = 5$ [1]

(ii) $Var(U) = Var(E[Y|X]) = Var(Y) - E[Var(Y|X)] = 4 - 2 = 2$ [2]
[Total 3]

Generally well answered. There were a few slips in the derivation, resulting in incorrect answers.

Q3

(i)

Answer B

$M_X(t) = E[e^{tX}] = \int_0^1 e^{tx}dx$

$M_X(t) = \frac{1}{t}(e^t - 1)$ for $t \neq 0$ [2]

(ii)

$M_x(0) =$ expectation of exp(0*X) $= 1$ [1]

(iii)

For a $U(0,2)$ distributed RV Z, we have:

$M_Z(t) = E[e^{tZ}]$ [½]

$= \frac{1}{2}\int_0^2 e^{tz}dz = \frac{1}{2t}(e^{2t} - 1)$ [1½]

(iv)

Since X and Y are independent, [1]

the MGF of $X + Y$ is given by the product of the MGFs:

$M_{X+Y}(t) = E[e^{tX}]E[e^{tY}] = \frac{1}{t^2}(e^t - 1)^2$. [1]

So, $M_{X+Y}(t) \neq M_{U(0,2)}(t)$, and therefore $X + Y$ does not have a $U(0,2)$ distribution [1]

[Total 8]

Part (i) was well answered. In part (ii), a common error was to state that the MGF is undefined. Common errors in part (iv) involved not stating that X and Y are independent, incorrectly deriving MGF(X+Y) and not summarising a response to the assertion.

Q4

(i)

The sampling distribution of S^2 is: $\frac{(n-1)S^2}{d^2} \sim \chi^2_{n-1}$ with $n = 25$ and $d^2 = 4$

Therefore the sampling distribution of S^2 is: $\frac{25-1}{4} S^2 = 6\,S^2 \sim \chi^2_{24}$ [2]

(ii)

So $E[6S^2] = 24$

And: $E[S^2] = 4$ [1]

(iii)

$\text{var}[6S^2] = 48$

So $\text{var}[S^2] = \frac{48}{36} = \frac{4}{3}$ [1]

[Total 4]

The question was well answered. In part (i) a number of candidates failed to specify the sampling distribution, as the question asked.

Q5

(i)

$f(x,y) = ke^{-(x+2y)} = ke^{-x}e^{-2y}, x > 0, y > 0$

[Or, f(x,y) = k g_X(x) g_Y(y).] [½]

The density function is expressed as a product of a function of x and y. Therefore, the joint probability function is a product of the two marginal probability functions for all (x, y) in the range of the variables hence X and Y are independent [½]

(ii)

The integral over the domain

$$\iint_0^\infty f(x,y)dxdy = k\iint_0^\infty e^{-(x+2y)}dxdy = k\int_0^\infty e^{-x}dx\int_0^\infty e^{-2y}\,dy$$

Or, the integral of f(x,y) is k times the integral of g_X times the integral of g_Y

$\int_0^\infty e^{-x}dx = -e^{-x}|_0^\infty = 1$, that is, the integral of g_X is one [1]

$\int_0^\infty e^{-2y}dy = -\frac{1}{2}e^{-2y}|_0^\infty = \frac{1}{2}$, that is, the integral of g_Y is 0.5 [1]

The integral of f(x,y) is 1 only for k=2 since, [1]

$\int_0^\infty f(x,y)dxdy = k \times 1 \times \frac{1}{2} = 1$, hence $k = 2$

(iii)

The marginal density is

$$f_Y(y) = 2\int_0^\infty e^{-x}e^{-2y}dx = 2e^{-2y}\int_0^\infty e^{-x}dx = 2e^{-2y}$$ [1]

(iv)

The conditional probability $P(Y \le y|Y > 3)$ is

$$P(Y \le y|Y > 3) = \frac{P(Y \le y, Y > 3)}{P(Y > 3)}$$

$$= \frac{P(3 < Y \le y)}{P(Y > 3)}$$

$$= \frac{F_Y(y) - F_Y(3)}{P(Y > 3)}, y > 3.$$

[1]

Therefore,

$f(y|Y > 3) = \frac{f_Y(y)}{P(Y>3)} = \frac{2e^{-2y}}{e^{-6}} = 2e^{6-2y}, \ \ y > 3,$ [1]

since

$$P(Y > 3) = \int_3^\infty f_Y(y)dy = -e^{-2y}|_3^\infty = e^{-6}\ .$$ [1]

(v)

Answer D [1]

The conditional expectation is given as

$$E[Y|Y > 3] = \int_3^\infty yf(y|Y > 3)dy = \int_3^\infty 2ye^{6-2y}dy$$

By taking $t = y - 3$,

$$E[Y|Y > 3] = \int_0^\infty 2(t+3)e^{-2t}dt = \int_0^\infty 2te^{-2t}dt + \int_0^\infty 6e^{-2t}dt$$

(vi)

$$\int_0^\infty 2te^{-2t}dt = \frac{e^{-2t}(-2t-1)}{2}|_0^\infty = (0) - \left(-\frac{1}{2}\right) = \frac{1}{2}$$ [1]

$$\int_0^\infty 6e^{-2t}dt = 3$$ [½]

$$E[Y|Y > 3] = 3.5$$ [½]

(vii)

Answer D [2]

$$E[Y^2|Y > 3] = \int_3^\infty y^2 f(y|Y > 3)dy = \int_3^\infty 2y^2 e^{6-2y}dy$$

Similar to (v),

$$\int_0^\infty 2(t+3)^2 e^{-2t}dt = \int_0^\infty 2t^2 e^{-2t}dt + \int_0^\infty 12te^{-2t}dt + \int_0^\infty 18e^{-2t}dt$$

$$E[Y^2|Y > 3] = 0.5 + 6 \times 0.5 + 9 = 12.5$$

The first integral is the moment of order 2 for the exponential distribution with parameter 2

(viii)

The variance of Y given $Y > 3$

$$\text{Var}[\text{Y}|\text{Y} > 3] = E[Y^2|Y > 3] - (\,E[Y|Y > 3])^2 = 12.5 - 3.5^2 = 0.25$$ [1]

[Total 14]

There were mixed answers in this question. This type of question has not appeared frequently in the presented form and many candidates found it challenging. Parts (i) - (iii) were well answered, while in part (iv) the justification for the conditional probability required was often missed. Parts (v), (vii), v(iii) were not well answered. These parts were potentially affected by the typing error in part (v). Part (vi) was poorly answered.

Q6

(i)(a)

Answer A

The likelihood function is:

$$L = [(1-p)^3]^{40} \times [3p(1-p)^2]^{60} \times [3p^2(1-p)]^{15} \times [p^3]^5$$
$$\propto (1-p)^{120}\, p^{60}\, (1-p)^{120}\, p^{30}\, (1-p)^{15}\, p^{15}$$
$$= (1-p)^{255}\, p^{105}$$

Taking logs:

$$\log L \propto 255 \log (1-p) + 105 \log p \qquad [2]$$

(b)

Using the answer from (i)(a):

Then differentiate with respect to p:

$$\frac{d \log L}{dp} = -\frac{255}{1-p} + \frac{105}{p} \qquad [1]$$

Setting this equal to zero gives:

$$255\hat{p} = 105(1-\hat{p})$$
$$360\hat{p} = 105 \qquad [1]$$
$$\hat{p} = \frac{105}{360} = 0.2917 \qquad [1]$$

(ii)

Specify the hypotheses using a χ^2 goodness of fit test:

H_0 – the probabilities follow a binomial bin(3, p) distribution

H_1 – the probabilities do not follow a binomial bin(3, p) distribution [1]

Using the MLE estimate for p above (0.29166):

$P(X = 0) = (1-p)^3 = 0.35540$

$P(X = 1) = 3p(1-p)^2 = 0.43902$

$P(X = 2) = 3p^2(1-p) = 0.18077$

$P(X = 3) = p^3 = 0.024812$ [1]

Therefore we get the following:

Number of exam passes	0	1	2	3
Observed no. of passes	40	60	15	5
Expected no. of passes	0.35540 x 120 = 42.648	0.43902 x 120 =52.682	0.18077 x 120 = 21.693	0.024812 x 120 = 2.9774

[2]

Combining last two columns, as expected no. of students with 3 exam passes < 5:

Number of exam passes	0	1	2 and 3
Observed no. of passes	40	60	20
Expected no. of passes	42.648	52.682	24.670

[1]

So: degrees of freedom = 3 – 1 – 1 = 1 [1]

The test statistic is:

$$\sum \frac{(O-E)^2}{E} = \frac{(40-42.648)^2}{42.648} + \frac{(60-52.682)^2}{52.682} + \frac{(20-24.670)^2}{24.670} = 2.0649$$

[1]

The test statistic is less than the 5% χ_1^2 critival value of 3.841 – therefore there is insufficient evidence at the 5% level to reject H_0. Therefore there is no evidence to conclude that the model is not a good fit [1]

[Total 13]

Part (i) was well answered. Part (ii) was reasonably well answered, but with a number of common errors, including: incorrect hypotheses stated, incorrect expected numbers calculated, no attempt at combining final 2 cells, incorrect degrees of freedom and a number of candidates not clearly showing their working.

Q7

(i)

t distribution would be suitable, with 33 df. [1]

(ii)

Assumed that the variances (rural and urban) are equal [1]

Equal variances seem to be justified given the S_y values for rural and urban areas are similar given the small sample sizes [1]

Assumption of Normality [½]

[Marks available 2½, maximum 2]

(iii)

Answer A

Test statistic $t = (\bar{Y}_{rural} - \bar{Y}_{urban}) / \left(S_P \sqrt{\frac{1}{n_{rural}} + \frac{1}{n_{urban}}} \right) \sim t_{n_{rural}+n_{urban}-2}$

under the null hypothesis that phone usage is equal.

$$S_P^2 = \frac{14S_{rural}^2 + 19S_{urban}^2}{33} = \frac{1}{33}(14 \times 2.1^2 + 19 \times 1.9^2) = 3.949$$

$$t = \frac{3.7 - 4.4}{\sqrt{3.95 \times (\frac{1}{15} + \frac{1}{20})}} = -1.031$$ [2]

(iv)

We are applying a two-sided test [1]

Critical values for t_{33} are not in the tables, but at the 2.5% level they are between 2.032 (t_{34}) and 2.037 (t_{32}) [1]

Since the test statistic lies in-between the table values,
i.e. $-2.032 < -1.031 < 2.037$,
[Or, as t33;2.5% is between 2.032 and 2.037, we have
t33;97.5% < -1.031 < t33;2.5%)]
we conclude that the null hypothesis of equal phone usage being equal cannot be rejected [1]

(v)

Assume $Y_{rural} \sim N(\mu, \sigma^2)$. [1]

Critical values for t_{14} at the 2.5% level are: -2.145 and +2.145 [1]

Confidence interval = $3.7 \pm \frac{t_{14,0.025}\, 2.1}{\sqrt{15}}$

$= [3.7 - 2.145 \times 0.542\ ,\ 3.7 + 2.145 \times 0.542]$ [1]

$= [2.537, 4.863]$ [1]

[Total 12]

Parts (i)-(iii) were generally well answered – common errors here included the justification of equal variances often being omitted. In part (iv), a number of candidates did not clearly refer to the critical values required, while in (v) the assumption of Normality was often missed.

Q8

(i)

There appears to be a number of possible outliers, [½]
(i.e. c0 or c365 days, these should be rechecked as they may be an error in the data or analysis.)

The plot exhibits a strong positive linear relationship between days and year [1]

R^2 percentage looks too high when compared to the scatterplot and the several outliers [½]

α value looks too high, we would expect it lower than 100 days, looking at the scatterplot [½]

β value sign looks to be the wrong way around, i.e. should be a positive [½]

The number of days is bounded in the interval $[0,366]$. If the intention is to project into future years, it may have been better to fit a model that respects this restriction, e.g. do a logistic transformation on the number of days first (although the relationship may no longer be linear) [1]

[Marks available 4, maximum 2]

(ii)

The required values are:

$s_{tt} = \sum t^2 - n\bar{t}^2$

= 42,925 – 50 * (1,275 / 50)^2 = 10,412.50 [½]

$s_{td} = \sum td - n\bar{t}\bar{d}$

= 282,724 – 50 * (1,275 / 50) * (8,502 / 50) = 65,923.00 [½]

Therefore:

$\hat{\beta} = \frac{s_{td}}{s_{tt}}$

= 65,923.00 / 10,412.50 = 6.331 [1]

$\hat{\alpha} = \bar{d} - \hat{\beta}\bar{t}$

= 8,502 / 50 – 6.331 * (1,275 / 50) = 8.596 [1]

Hence the regression line equation as given in the question

(iii)(a)

$s_{dd} = \sum d^2 - n\bar{d}^2$

= 1,911,378 – 50 * (8,502 / 50)^2 = 465,697.92 [1]

$\hat{\sigma}^2 = \frac{1}{n-2}\left(s_{dd} - \frac{s_{td}^2}{s_{tt}}\right)$

= (1 / 48) * (465,697.92 – 65,923^2 / 10,412.50) = 1,006.878 [1]

So the standard error of $\hat{\beta}$ is:

$\sqrt{\frac{\hat{\sigma}^2}{s_{tt}}}$

= sqrt(1,006.878 / 10,412.50) = 0.311 [1]

(iii)(b)

The test is as follows:

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

Under the null hypothesis, the corresponding test statistic is:

$\frac{\hat{\beta}-0}{\sqrt{\frac{\hat{\sigma}^2}{s_{tt}}}}$ = 6.331 / 0.311 = 20.360 [½]

The 1% critical values from the t_{48} distribution are *circa* ± 2.678 (using t_{50} for simplicity) [½]

Or, interpolate to find critical values ± 2.6832

Since $20.35998 > 2.678$ there is strong evidence to reject H_0 at the 1% level,
i.e. there is sufficient evidence to suggest that there is a strong linear relationship [½]

(iii)(c)

Using the same standard error and percentage point in (iii)(b), the confidence interval is found by:

$$\hat{\beta} \pm 2.678\sqrt{\frac{\hat{\sigma}^2}{s_{tt}}}$$

$= 6.331 \pm 2.678 \times 0.311$ [1]

$= (5.498\ ,7.164)$ [1]

(iv)(a)

The test is as follows:

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

Under the null hypothesis, the corresponding test statistic is:

= 5.215 / 1.983 = 2.630 [½]

The 1% critical values from the t_{48} distribution are circa ± 2.678 (using t_{50} for simplicity)
Since $2.62995 < 2.678$ we have no evidence to reject H_0 at the 1% level [½]
We conclude that there is insufficient evidence of a linear relationship [½]

(iv)(b)

Using the same standard error and percentage point in (iv)(a), the confidence interval is:

$5.215 \pm 2.678 \times 1.983$ [1]

$= (-0.095\ ,10.524)$ [1]

(v)

The two confidence intervals overlap, with one being a subset of the other [1]
This suggests that we cannot confidently conclude that the underlying slope coefficients are different [1]
However, the large standard error leads to a wide confidence interval, meaning we lack evidence in the conclusion to the above bullet points [1]

Alternative comments:
For deciding if the two underlying slope parameters are equal, a formal test would be required for the difference between the two parameters [1]
where the variance of the difference should also be taken into account properly [1]
[Marks available 5, maximum 3]

(vi)

The test conclusions in (iii)(b) and (iv)(a) appear to disagree [1]
The test statistic in (iii)(b) lies well over the critical value whereas the test statistic in (iv)(a) lies just under the critical value [1]
So this suggests that the slope coefficients may be different for the two sets of climate change data [1]
Recording of past data, method of collection, errors in collection / the data etc from the alternative sources, treatment of outliers, differences in definition (e.g. location used) of extreme weather, may lead to the apparent differences observed [1]
Alternative comments:
We reject this hypothesis of the slope parameter being significantly different from 0 in part (iii)(b) but not in part (iv)(a) [1]

From these results alone it appears that the two parameters are therefore different, which seems to contradict the conclusion in part (v) [1]
However, for deciding if the two underlying slope parameters are equal, a formal test would be required for the difference between the two parameters, where the variance of the difference should also be taken into account properly [1]

[Marks available 7, maximum 4]

[Total 23]

Part (i) required more analysis and judgement from candidates, compared to the usual comments required for this question type. Many candidates made generic comments regarding the plot, with very little challenge or comment regarding the statistics given in the question. Parts (ii)-(iv) were generally answered well, with the only issue being numerical errors. Parts (v)-(vi) were poorly answered.

Q9

(i)

Answer C [3]

The likelihood is

$f(x|\mathrm{u}) = \prod_{i=1}^{n} \frac{\mathrm{u}^{x_i} e^{-\mathrm{u}}}{x_i!}$ and prior for u is $f(\mathrm{u}) \propto \mathrm{u}^{\mathrm{a}-1} e^{-\mathrm{bu}}$

So the posterior density is given by

$$f(\mathrm{u}|x) \propto f(x|\mathrm{u}) \times f(\mathrm{u}) \propto \prod_{i=1}^{n} \frac{\mathrm{u}^{x_i} e^{-\mathrm{u}}}{x_i!} \times \mathrm{u}^{\mathrm{a}-1} e^{-\mathrm{bu}} \propto \mathrm{u}^{\mathrm{a}+\sum x_i - 1} e^{-(\mathrm{b}+\mathrm{n})\mathrm{u}}$$

(ii)

$u|x$ follows a gamma(a+ $\sum x_i$, b+n) distribution [2]

(iii)(a)

The Bayesian estimate of μ under quadratic loss is the posterior mean and so:

$\hat{\mathrm{u}} = E(\mathrm{u}|x) = \frac{\mathrm{a}+\sum x_i}{\mathrm{b}+\mathrm{n}}$. [2]

(b)

This can be written as:

$$\hat{\mathrm{u}} = \frac{a}{\mathrm{b}} \frac{\mathrm{b}}{\mathrm{b}+\mathrm{n}} + \frac{\sum x_i}{\mathrm{n}} \frac{\mathrm{n}}{\mathrm{b}+\mathrm{n}} \quad [1]$$

$$= (1 - Z)\frac{\mathrm{a}}{\mathrm{b}} + Z\bar{x},$$

where $Z = \frac{\mathrm{n}}{\mathrm{b}+\mathrm{n}}$ is the credibility factor [1]

(iv)

$$\hat{\mathrm{u}} = E(\mathrm{u}|x) = \frac{\mathrm{a}+\sum x_i}{\mathrm{b}+\mathrm{n}} = \frac{9+320}{3+6} = \frac{329}{9} = 36.56$$

[2]

(v) $V(u|x) = \frac{\mathrm{a}+\sum x_i}{(\mathrm{b}+\mathrm{n})^2} = \frac{329}{9^2} = 4.06$ [2]

(vi)
The prior variance of u has changed from 9/9 = 1 to 18/36 = 0.5. With the data (and hence the likelihood) being unchanged [1]
this means that the posterior variance will also be reduced (but not necessarily halved) [1]
[Total 15]

Answered very well by most candidates. A common error in part (ii) was specifying an incorrect Gamma distribution.

[Paper Total 100]

END OF EXAMINERS' REPORT