



Institute
and Faculty
of Actuaries

EXAMINERS' REPORT

CS1 – Actuarial Statistics

Core Principles

Paper A

September 2023

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.

For some candidates, this may be their first attempt at answering an examination using open books and online. The Examiners expect all candidates to have a good level of knowledge and understanding of the topics and therefore candidates should not be overly dependent on open book materials. In our experience, candidates that spend too long researching answers in their materials will not be successful either because of time management issues or because they do not properly answer the questions.

Many candidates rely on past exam papers and examiner reports. Great caution must be exercised in doing so because each exam question is unique. As with all professional examinations, it is insufficient to repeat points of principle, formula or other text book works. The examinations are designed to test "higher order" thinking including candidates' ability to apply their knowledge to the facts presented in detail, synthesise and analyse their findings, and present conclusions or advice. Successful candidates concentrate on answering the questions asked rather than repeating their knowledge without application.

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
November 2023

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 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 in this report (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 answered inadequately in general (e.g. Q4). 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 recent papers.

Candidates are encouraged to practise more on the fundamentals of mathematical calculus and probability. For example, mixed answers in parts of Q3 and Q6 suggest that a number of candidates would benefit from additional work on differentiation and basic probability calculations.

C. Pass Mark

The Pass Mark for this exam was 60.
1508 presented themselves and 730 passed.

Solutions for Subject CS1A - September 2023**Q1**

(i)

$$\text{Expectation of } X: E[X] = \frac{1}{60} [20 + \dots + 79] \quad [1]$$

$$= \frac{1}{60} \times 99 \times \frac{60}{2} = 49.5 \quad [1]$$

(ii)

$$\text{Variance of } X: \text{Var}(X) = \text{Var}(19 + Y) \text{ with } Y \text{ uniform on } 1, \dots, 60. \quad [1]$$

$$\text{Var}(X) = \text{Var}(Y) = \frac{60^2 - 1}{12} = \frac{3599}{12} = 299.9167 \quad [1]$$

$$\text{Std}(X) = \sqrt{299.9167} = 17.3181 \quad [1]$$

Alternative answer, using result from Formulae and Tables:

$$E(X) = (20+79)/2 = 49.5$$

$$\text{Var}(X) = 1/12 *$$

$$(79-20)*(79-20+2*1) = 299.9167.$$

[Total 5]

Well answered in general.

A number of numerical errors were made, particularly when the result from the 'orange book' was used.

Q2

(i)

$$f \text{ is a density function therefore } \int_0^{20} f(x) dx = 1 \quad [1/2]$$

$$\int_0^{20} f(x) dx = a \left(\frac{x^2}{2} - 20x \right)_0^{20} = a \left(\frac{20^2}{2} - 400 \right) = -200a \quad [1]$$

$$\text{Thus } -200a = 1 \Rightarrow a = -\frac{1}{200} = -0.005 \quad [1/2]$$

(ii)

$$P(X > 16 | X > 8) = \frac{P(X > 16)}{P(X > 8)} \quad [1]$$

$$P(X > 16) = -0.005 \left(\frac{x^2}{2} - 20x \right)_{16}^{20} = 0.04 \quad [1/2]$$

$$P(X > 8) = -0.005 \left(\frac{x^2}{2} - 20x \right)_8^{20} = 0.36 \quad [1/2]$$

$$\text{Therefore } P(X > 16 | X > 8) = \frac{0.04}{0.36} = \frac{1}{9} = 0.111 \quad [1]$$

[Total 5]

This questions was generally well answered by candidates.

Q3

(i)

The random variable Y follows a negative binomial distribution [1]
 and represents the number of failures before the k th success [½]
 in trials occurring independently with probability of success $p \in (0,1)$. [½]

(ii)

Correct answer: A [3]

We want to write the probability mass function in the general form $\exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$.

Taking the log and the exponential we obtain

$$P(Y = y) = \exp\left\{y \log(1 - p) + k \log p + \log\binom{k+y-1}{y}\right\},$$

which is the expected form with

$$\theta = \log(1 - p) \text{ i.e. } p = 1 - e^\theta$$

$$\phi = 1, \text{ so that } a(\phi) = 1$$

$$b(\theta) = -k \log p$$

$$c(y, \phi) = \log\binom{k+y-1}{y}$$

(iii)

Using $p = 1 - e^\theta$ we have $E[Y] = b'(\theta)$ [½]

$$= \frac{\partial}{\partial \theta}(-k \log(1 - e^\theta))$$
 [½]

$$= -k \frac{-e^\theta}{1 - e^\theta} = k \frac{1 - p}{p}$$
 [1]

(iv)

Correct answer: C [2]

$$V[Y] = a(\phi)b''(\theta) = \frac{\partial}{\partial \theta}\left(k \frac{e^\theta}{1 - e^\theta}\right) = k \frac{e^\theta}{(1 - e^\theta)^2} = k \frac{1 - p}{p^2}.$$

[Total 9]

Parts (i) and (ii) were generally well answered.

In parts (iii) and (iv) there were mixed answers, with only well prepared candidates carrying out the required differentiation correctly.

A number of candidates did not express their answer in terms of p in part (iii).

Q4

(i)

Size of dataset [½]

Speed at which data arrive [½]

Variety of data and related sources [½]

Reliability individual data elements [½]

(ii)

Data Security.

[½]

Privacy.

[½]

[Total 3]

Responses were varied, with some of the comments being vague or unrelated, particularly in part (ii).

This is a knowledge based topic that has not been examined in recent sessions. Candidates are reminded that they should prepare for the entire breadth of the syllabus.

Q5

(i)

Correct answer: A

[2]

The definition of the cumulative distribution function is:

$$F_W(s) = P(W \leq s) = P(\max(X, Y) \leq s)$$

From the above definition, if the maximum value of X and Y has to be less than s , then both values have to be less than s . Therefore:

$$F_W(s) = P(X \leq s, Y \leq s) = P(X \leq s)P(Y \leq s) = F_X(s)F_Y(s).$$

(ii)

The definition of the cumulative distribution function is:

$$F_Z(s) = P(Z \leq s) = P(\min(X, Y) \leq s)$$

[1]

Using the above definition, if the minimum value of X and Y has to be greater than s , then both of the values have to be greater than s .

Therefore:

$$1 - F_Z(s) = P(Z \geq s) = P(\min(X, Y) \geq s)$$

[1]

$$P(X \geq s, Y \geq s) = P(X \geq s)P(Y \geq s)$$

[1]

Using the following and substituting into the equation above gives:

$$P(X \geq s) = 1 - P(X \leq s) = 1 - F_X(s)$$

[½]

$$P(Y \geq s) = 1 - P(Y \leq s) = 1 - F_Y(s)$$

[½]

$$1 - F_Z(s) = (1 - F_X(s))(1 - F_Y(s)) = 1 - F_X(s) - F_Y(s) + F_X(s)F_Y(s)$$

[1]

$$F_Z(s) = F_X(s) + F_Y(s) - F_X(s)F_Y(s)$$

Alternative solution:

$$F_Z(s) = P(Z \leq s) = P(\min(X, Y) \leq s)$$

$$= P(X \leq s \cup Y \leq s)$$

$$= P(X \leq s) + P(Y \leq s) - P(X \leq s \cap Y \leq s) \quad (\text{or } P(X \leq s) + P(Y \leq s) -$$

$$P(\max(X, Y) \leq s))$$

$$= P(X \leq s) + P(Y \leq s) - P(X \leq s)P(Y \leq s)$$

$$= F_X(s) + F_Y(s) - F_X(s)F_Y(s)$$

(iii)(a)

Correct answer: B

[3]

Using the Actuarial Tables, or otherwise:

$$F_X(s) = 1 - e^{-6s}, \quad F_Y(s) = 1 - e^{-6s}$$

Substituting these into part (ii) gives:

$$F_Z(s) = F_X(s) + F_Y(s) - F_X(s)F_Y(s) = (1 - e^{-6s}) + (1 - e^{-6s}) - (1 - e^{-6s})(1 - e^{-6s})$$

$$= (1 - e^{-6s}) + (1 - e^{-6s}) - 1 + e^{-6s} + e^{-6s} - e^{-12s}$$

$$= 1 - e^{-12s}$$

(b)

The cumulative distribution function in (iii)(a) is the exponential distribution.

[½]

The mean of $Z = \min(X, Y)$ is $\frac{1}{12}$.

[½]

[Total 11]

This questions was well answered in general.

A common error in part Q5(iii)(b) was failing to provide the mean of Z as requested in the question.

Q6

Define X as the number of trains used for a journey and D is the event that a journey is delayed.

(i)

$$P[\text{less than three trains}] = 0.60 + 0.30 = 0.9$$

[1]

(ii)

$$P[D] = P[X = 1 \cap D] + P[X = 2 \cap D] + P[X \geq 3 \cap D]$$

[½]

$$= P[D|X = 1]P[X = 1] + P[D|X = 2]P[X = 2] + P[D|X \geq 3]P[X \geq 3]$$

[½]

$$= 0.05 \times 0.6 + 0.12 \times 0.3 + 0.17 \times 0.1 = 0.083$$

[1]

(iii)

$$P[X = 1|D] = \frac{P[X=1 \cap D]}{P[D]} = \frac{P[D|X = 1]P[X=1]}{P[D]} = \frac{0.05 \times 0.60}{0.083} = 0.36144$$

[1]

$$P[X = 2|D] = \frac{P[X=2 \cap D]}{P[D]} = \frac{P[D|X=2]P[X=2]}{P[D]} = \frac{0.12 \times 0.30}{0.083} = 0.43373$$

[1]

$$P[\text{one or two trains} | D] = 0.36144 + 0.43373 = 0.7952.$$

[1]

$$\text{Alternatively: } 1 - \frac{0.17 \times 0.1}{0.083} = 1 - 0.20482 = 0.7952.$$

(iv)

The probability of randomly choosing a journey with fewer than 3 trains (out of all journeys), is higher compared to the probability of randomly choosing a journey with fewer than 3 trains out of the delayed journeys.

[1]

Given the information that a journey is delayed, the probability that the journey

involves 3 or more trains is increased.

[1]
[Total 8]

Parts (i)-(ii) were well answered.

In part (iii) there were often some numerical errors when calculating intermediate probabilities.

In part (iv) comments were often vague and unclear.

Q7

(i)

Correct answer: C

[3]

The likelihood for the parameter c given n independent randomly sample is

$$L(c) = \prod_{i=1}^n \frac{c}{y_i^{c+1}} = c^n \prod_{i=1}^n \frac{1}{y_i^{c+1}}$$

$$l(c) = n \log(c) - (c + 1) \sum_1^n \log(y_i).$$

The corresponding partial derivative is $\frac{\partial l}{\partial c} = \frac{n}{c} - \sum_1^n \log(y_i)$

$$\frac{\partial l}{\partial c} = 0 \text{ then } c = \frac{n}{\sum_1^n \log(y_i)}.$$

The MLE of c is $\hat{c} = \frac{n}{\sum_1^n \log(y_i)}$.

(ii)

The prior distribution is: $f(c) \propto c^{a-1} e^{-bc}$.

[1]

The likelihood is: $L(c) = c^n e^{-(c+1) \sum_1^n \log(y_i)} \propto c^n e^{-c \sum_1^n \log(y_i)}$

[1½]

The posterior distribution is given as:

$$P(c) \propto c^{a-1} e^{-bc} c^n e^{-c \sum_1^n \log(y_i)} = c^{n+a-1} e^{-c(b + \sum_1^n \log(y_i))}$$

[1½]

The posterior distribution of the parameter c is a $Gamma(n + a, b + \sum_1^n \log(y_i))$. [2]

(iii)

The posterior and the prior distributions are from the same family, therefore the prior is a conjugate prior.

[½]

[½]

(iv)

Under a quadratic loss, the Bayesian estimate is the posterior mean

[1]

i.e. $\frac{n+a}{b + \sum_1^n \log(y_i)}$

[1]

[Total 12]

This question was in general well answered.

In part (ii) a number of candidates failed to specify the likelihood function in the appropriate form that leads to the gamma posterior distribution.

Q8

(i)

Correct answer: B [3]

The likelihood can be written as

$$L(\theta, \mathbf{x}) = \prod_{i=1}^n f(x_i) = \theta^n \left(\prod_{i=1}^n x_i \right) e^{-\frac{\theta}{2} \sum_{i=1}^n x_i^2}.$$

The posterior density can be written as

$$f(\theta|x) \propto \theta^n e^{-\frac{\theta}{2} \sum_{i=1}^n x_i^2} \times \theta^{a-1} e^{-\theta b} = \theta^{n+a-1} e^{-\theta \left(\frac{\sum_{i=1}^n x_i^2}{2} + b \right)}.$$

(ii)

θ follows a Gamma distribution [1]

with parameters $n + a$ [1]

and $\frac{\sum_{i=1}^n x_i^2}{2} + b$. [1]

(iii)

The Bayesian estimate under all-or-nothing loss is the mode of the posterior distribution: [1]

$$\log(f(\theta|x)) = (n + a - 1)\log\theta - \theta \left(\frac{\sum_{i=1}^n x_i^2}{2} + b \right) + \text{const} \quad [1]$$

$$\frac{d}{d\theta} \log(f(\theta|x)) = \frac{n+a-1}{\theta} - \left(\frac{\sum_{i=1}^n x_i^2}{2} + b \right) \quad [1]$$

Setting equal to 0 leads to

$$\hat{\theta} = \frac{n+a-1}{\frac{\sum_{i=1}^n x_i^2}{2} + b} = 0.302. \quad [1]$$

[Total 10]

This questions was generally well answered.

Q9

(i)

Test $H_0: \rho = 0$ versus $H_1: \rho > 0$ [1]

$$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 80 \quad [1/2]$$

$$S_{yy} = \sum y_i^2 - (\sum y_i)^2/n = 330.0137 \quad [1/2]$$

$$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n = 161.3633 \quad [1/2]$$

The sample correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.993 \quad [1/2]$$

Test statistic is:

$$\frac{r\sqrt{n}}{\sqrt{1-r^2}} = 22.41 \quad [1]$$

and under the null hypothesis this should be a value from the t_7 distribution. [1/2]

Percentage point for 1-sided test is $t_{7,0.99} = 2.998$. [1/2]

Therefore, we have strong evidence against the null hypothesis. [1/2]

We reject the null hypothesis in favour of the alternative that $\rho > 0$. [1/2]

(ii)

The sample correlation and test suggest a strong positive linear relationship between X and Y . [1]
[½]
[½]

(iii)
 $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 2.017, \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1.182.$ [2]

The fitted regression line is
 $\hat{y} = 1.182 + 2.017x.$ [1]

(iv)
The point estimate is
 $\hat{y} = 1.182 + 2.017 \times 3 = 7.233$ [1]

$\hat{\sigma}^2 = \frac{1}{n-2} (S_{yy} - S_{xy}^2/S_{xx}) = \frac{1}{7} (330.0132 - 161.3633^2/80) = 0.648$ [1]

$\text{Var}(\hat{y}) = \left[\frac{1}{9} + \frac{(3 - \frac{42}{9})^2}{80} \right] \times 0.648 = 0.0945$ [1]

Percentage point is $t_{7,0.995} = 3.499$ [½]

So, the 99% CI is given as
 $7.233 \pm 3.499 \times \sqrt{0.0945}.$ i.e. (6.157, 8.309). [1½]

[Total 16]

This questions was answered in general.

There were some numerical errors in part (iv).

Q10

(i)
Assuming that the data come from a normal distribution:

$\bar{x} = \frac{733}{20} = 36.65$ [1]

$S^2 = \frac{1}{19} (29,203 - 20 \times 36.65^2) = 123.0816$ [1]

$t_{0.025,19} = 2.093$ [1]

$\bar{x} \pm t_{0.025,19} \frac{s}{\sqrt{20}} = 36.65 \pm 2.093 \sqrt{\frac{123.0816}{20}}$ [1]

$= [31.458, 41.842]$ [1]

(ii)
 $S_{xx} = 29,203 - \frac{1}{20} 733^2 = 2,338.55$ [1]

$S_{yy} = 2,009 - \frac{1}{20} 189^2 = 222.95$ [1]

$S_{xy} = 6,208 - \frac{1}{20} \times 733 \times 189 = -718.85$ [1]

$\rho = -\frac{718.85}{\sqrt{2338.55 \times 222.95}} = -0.9955457$ [1]

(iii)
Pearson's correlation coefficient measures the strength of a linear relationship

between two variables. [½]
 Since we have two numerical variables using this coefficient is justified here. [½]
 On the other hand, rank correlation coefficients are less appropriate since the numerical difference between two successive values has a meaning in this context, and numerical values therefore contain more information than just ranks. [1]

(iv)
 We have H_0 : crime rates are equal, against H_1 : crime rates are different. [1]

We see in the plot that the three high income regions are the only observed regions with an income of more than 50,000.

Mean crime rate in high earner regions = $\frac{1}{3}(3.896 + 3.958 + 2.658) = 3.504$ [½]

Mean crime rate in other regions = $\frac{1}{17}(189 - 3 \times 3.504) = 10.5$ [½]

Test statistic $T = \frac{10.5 - 3.504}{S_P \sqrt{\frac{1}{17} + \frac{1}{3}}}$ with t -distribution with 18 degrees of freedom [1]

For pooled estimate S_P^2 :

$\sum_{i:high} y_i^2 = 3.896^2 + 3.958^2 + 2.658^2 = 37.91$ [½]

$\sum_{i:others} y_i^2 = 2,009 - 37.91 = 1971.09$ [½]

$S_{high}^2 = \frac{1}{2}(37.91 - 3 \times 3.504^2) = 0.538$ [½]

$S_{others}^2 = \frac{1}{16}(1971.09 - 17 \times 10.5^2) = 6.05$ [½]

$S_P^2 = \frac{1}{18}(2 \times 0.538 + 16 \times 6.05) = 5.438$ [½]

$T = \frac{10.5 - 3.504}{\sqrt{5.438(\frac{1}{17} + \frac{1}{3})}} = 4.79$ [½]

Two-sided test, [½]

therefore need $t_{0.995,18} = 2.87844$ [½]

(or other reasonable quantiles, e.g. $t_{0.975,18} = 2.10092$).

The value of the test statistic is greater than the quantile, therefore reject null hypothesis of equal crime rates. [1]

(v)
 The assumption seems to be unjustified as the spread of values for crime rate in the plot is very small for high income regions (50k) while it is rather wide for other regions. [1]

This raises questions about the validity of the test. [1]

[Total 21]

Parts (i) and (ii) were well answered in general, with a common error in (i) being using the critical value from a standard normal distribution.

In part (iii) comments were often quite vague.

Part (iv) was not well answered, with many candidates failing to work through it appropriately.

[Paper Total 100]

END OF EXAMINERS' REPORT



Institute and Faculty of Actuaries

Beijing

14F China World Office 1 · 1 Jianwai Avenue · Beijing · China 100004
Tel: +86 (10) 6535 0248

Edinburgh

Level 2 · Exchange Crescent · 7 Conference Square · Edinburgh · EH3 8RA
Tel: +44 (0) 131 240 1300

Hong Kong

1803 Tower One · Lippo Centre · 89 Queensway · Hong Kong
Tel: +852 2147 9418

London (registered office)

7th Floor · Holborn Gate · 326-330 High Holborn · London · WC1V 7PP
Tel: +44 (0) 20 7632 2100

Oxford

1st Floor · Belsyre Court · 57 Woodstock Road · Oxford · OX2 6HJ
Tel: +44 (0) 1865 268 200

Singapore

5 Shenton Way · UIC Building · #10-01 · Singapore 068808
Tel: +65 8778 1784

