INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2019 Examinations

Subject CS1 – Actuarial Statistics Core Principles (Part 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.

Mike Hammer
Chair of the Board of Examiners
July 2019

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

1. The aim of the Actuarial Statistics 1 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 full 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.

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

1. Performance was satisfactory, with most candidates demonstrating good understanding and application of core topics in actuarial statistics.

2. Answers requiring the derivation of statistical properties contained a considerable number of errors (e.g. Question 2). Candidates are encouraged to revise corresponding parts of the Core Reading and practice on using provided definitions to derive important statistical properties.

3. The calculation of probabilities of certain events is fundamental for the understanding and use of actuarial statistics. Candidates are advised to practice on this topic (e.g. Question 4), under scenarios of varying complexity.

4. Attention is also drawn on providing full and mathematically precise definitions or statistical statements (e.g. Questions 5, 6).

C. Pass Mark

The combined pass mark for CS1 in this exam diet was 58.

@Institute and Faculty of Actuaries

Solutions Subject CS1 – A

Q1

(i) $E(X) = \frac{1}{\lambda} = 15$ so $\lambda = \frac{1}{15} = 0.06666 \ldots$ [1]

So P(X > 20) = 1 – F(20) = 1 – (1 – exp(-0.06666 x 20)) = 0.26360 [1]

(ii) $P(X > 20 \mid X > 15) = \frac{P(X > 20 \cap X > 15)}{P(X > 15)}$ [1]

$= \frac{P(X > 20)}{P(X > 15)} = P[X > 5]$ (using memoryless property)

[1]

Numerator as calculated above for Part (i), denominator is:

$P(X > 15) = 1 - F(15) = 1 - (1 - \exp(-0.06666\, x\, 15) = \exp(-1)$

$So\ P(X > 20 \mid X > 15) = 0.26360/0.36788 = 0.71653$ [1]

Alternatively, using property of exponential distribution:

$$P(X > 20 \mid X > 15) = P(X > 20 - 15) = 0.71653$$

The question was answered generally well by most candidates. Some candidates were unable to recall and apply the memoryless property of the exponential distribution in part (ii).

Q2

(i) $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$ [1]

(ii) $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$ [1]

(iii) $\text{MSE}(\hat{\theta}) = E\left[\left(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right)^2\right]$ [1]

$= E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2\right] + 2E[\hat{\theta} - E(\hat{\theta})]\,[E(\hat{\theta}) - \theta] + [E(\hat{\theta}) - \theta]^2$ [1]

$= V(\hat{\theta}) + \text{bias}^2(\hat{\theta}).$ [1]

Parts (i) and (ii) require standard definitions and were answered well by most candidates. Part (iii) was answered poorly. A number of candidates repeated the answer in parts (ii) and (iii), failing to properly derive the required expression.

Q3

The mean number of claims per policy is $\bar{x} = \frac{82}{200} = 0.41$ [1]
Using the normal approximation to the Poisson distribution,
the approximate 95% CI [1]
for λ is $\bar{x} \pm 1.96\sqrt{\frac{\bar{x}}{n}}$ which gives
$0.41 \pm 1.96\sqrt{\frac{0.41}{200}} = 0.41 \pm 0.0887$, i.e. (0.321, 0.499). [2]

Candidates performed strongly on this question, with most applying correctly the normal approximation to the Poisson distribution. A common error was to use the incorrect mean, i.e. 82 rather than 82/200. Answers working with the alternative statistic $\frac{\sum X_i - n\lambda}{\sqrt{n\hat{\lambda}}}$ *were given full credit when used correctly.*

Q4

The initial step is to define the sample space:
Sample space =

$$\Omega = \{(1,1),(1,2),(1,3),\dots,(6,6)\} = (i,j) \mid i,j = 1,2,3,4,5,6$$

Each outcome is equally likely with probability 1/36.
$A = \{(1,2),(2,1)\}$
$B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$
$C = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)\}$

(i) P(C) = 11/36 [1]
(ii) P(A|C) = $\frac{P(A\cap C)}{P(C)}$ [½]
$P(A \cap C) = 2/36$ [1]
P(A|C) = $\frac{2/36}{11/36}$ = 2/11 [½]

(iii) P(B|C) = $\frac{P(B\cap C)}{P(C)}$ = $\frac{2/36}{11/36}$ = 2/11 [2]

(iv) P(A) = 2/36 ≠ P(A|C) -> So they are not independent. [1]

(v) P(B) = 6/36 ≠ P(B|C) -> So they are not independent. [1]

The question was generally well answered. Candidates that took a methodical approach in setting out the sample space scored well on parts (i), (ii) and (iii). Most candidates were able to demonstrate correctly the lack of independence for parts (iv) and (v). Common errors occurred in parts (ii) and (iii) were P(C) or P(B and C) etc. were calculated incorrectly.

Q5

(i) If $X_1, X_2, \dots, X_n$ is a sequence of independent, identically distributed random variables with finite mean μ and finite (non-zero) variance σ^2, then the distribution of

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

[1]

approaches the standard normal distribution, $N(0,1)$, [½]
as n tends to infinity. [½]

(ii) If $X_1, X_2, \dots, X_n$ are independent and each follows a Bernoulli(p) distribution with mean p and variance p(1-p), then $B = \sum_{i=1}^{n} X_i$ follows a Binomial(n,p) distribution.

[2]

The CLT from part (i) can also be expressed as follows: the distribution of

$$\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$$

approaches the standard normal distribution, $N(0,1)$, as n tends to infinity. [1]

Therefore, we have that $\frac{B-np}{\sqrt{np(1-p)}}$ has the standard normal distribution for large n.

[1]

(iii) We can use $\hat{p} = 0.57$ in the variance (denominator). [1]

So the CI is given as $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, i.e. [1]

$0.57 \pm 1.96\sqrt{\frac{0.57(1-0.57)}{100}}$ = (0.473, 0.667) [2]

Parts (i) and (ii) were reasonably well attempted. In part (i) full credit was given for providing the answer in terms of of $\frac{\sum X_i - n\mu}{n\sigma^2}$ *or equivalent. A number of candidates were not precise enough in their statement of the central limit theorem, for example, missing out the requirement for large sample size. Part (iii) was well answered, although a number of arithmetic slips were made in the calculation of the confidence interval.*

Q6

(i)

The random variables X and Y are independent if, and only if, the joint pdf is the product of the two marginal pdfs for all (x,y) in the range of the variables, i.e.
$f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$ for all (x, y) in the range. [2]

(ii)

(a) $f_X(x) = \int_x^1 8xy\,dy = 8x[y^2/2]_x^1 = \frac{8x(1-x^2)}{2} = 4x(1-x^2), \quad 0 < x < 1$

[1]

$f_Y(y) = \int_0^y 8xy\,dx = 8y[x^2/2]_0^y = 4y^3, \quad 0 < y < 1$

[1]

(b) Here, $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$ so X and Y are not independent.

[1]

(iii)

$E(X|Y=y) = \int_0^y \frac{x\,f_{X,Y}(x,y)}{f_Y(y)}\,dx = \int_0^y x\,\frac{8xy}{4y^3}\,dx$

[2]

$= \frac{2}{3y^2}[x^3]_0^y = \frac{2y}{3}, \quad 0 < y < 1$

[1]

In part (i) many candidates failed to give the full definition, for example mentioning that the property must hold for all (x, y). Part (ii) was well answered by almost all candidates. Part (iii) was answered successfully by only the strongest candidates, with many candidates getting confused with the integral limits or integrating with respect to y instead of x.

Q7

(i) Variance is known, so $\bar{X} \sim N(3, \frac{4}{9})$ [1]

$$P[\bar{X} > 4] = 1 - P\left[\frac{\bar{X}-3}{\frac{2}{3}} < \frac{4-3}{\frac{2}{3}}\right] = 1 - P[Z < 1.5]$$
$$= 1 - 0.93319 = 0.06681$$

[1]

(ii)

$\bar{X}$ and $\bar{Y}$ are independent and both are normally distributed. So, $\bar{Y} - \bar{X}$ is normally distributed,

[1]

$\bar{Y} - \bar{X} \sim N\left(4-3, \frac{4}{9} + \frac{10}{18}\right) = N(1,1)$ [1]

$P[\bar{X} > \bar{Y}] = P[\bar{Y} - \bar{X} < 0] = P[Z < -1] = 1 - P[Z \leq 1]$

$= 1 - 0.84134 = 0.15866$ [1]

(iii)

$P[S_X^2 > 4] = 1 - P[S_X^2 \leq 4]$ [½]

$= 1 - P\left[\frac{8 \times S_X^2}{4} \leq 8\right]$ [½]

$= 1 - P[\chi_8^2 \leq 8]$ [½]

$= 1 - 0.5665 = 0.4335$ [½]

(iv)

S_X^2 and S_Y^2 are independent, and therefore, $\frac{S_X^2/4}{S_Y^2/10} \sim F_{8,17}$ [1]

$P[S_X^2 > S_Y^2] = P\left[\frac{S_X^2}{S_Y^2} > 1\right] = P\left[\frac{S_X^2/4}{S_Y^2/10} > \frac{10}{4}\right] = P\left[F_{8,17} > 2.5\right]$ [1]

Checking the 5% probabilities for the $F_{8,17}$ distribution in the "Formulae and Tables" we find that

$P\left[F_{8,17} > 2.5\right] \approx P\left[F_{8,17} > 2.548\right] = 0.05$ [1]

(v)

Actually, we have $P\left[F_{8,17} > 2.5\right] > P\left[F_{8,17} > 2.548\right]$. [1]

So, the probability we are looking for is slightly greater than 5%. [1]

The question was well attempted. In parts (ii) and (iv), reference to independence is required for full marks.

Q8

(i) (a)

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}$$

Taking logs of both sides:

$$\log\big(f(y)\big) = y \log \mu - \mu - \log y!$$

[1]

Then taking exponentials:

$f(y) = \exp(y \log \mu - \mu - \log y!)$ [½]

Comparing this to the generalised form of the exponential family of distributions:

$$f(y;\theta;\varphi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right\}$$

We see that:

$\theta = \log\mu$ [½]
$b(\theta) = \mu = e^{\theta}$ [½]
$\varphi = 1$ [½]
$a(\varphi) = 1$ [½]
$c(y,\varphi) = -\log y!$ [½]

(i) (b)

Using the properties of exponential distributions:

$E(Y) = b'(\theta) = \frac{d}{d\theta}e^{\theta} = e^{\theta} = \mu$

[1]

$\text{Var}(Y) = a(\varphi)\, b''(\theta) = 1 \times \frac{d^2}{d\theta^2}e^{\theta} = e^{\theta} = \mu$ [1]

(ii) (a) Using the model output, we can see that:

$\beta > 2\ \times$ standard error (β)

i.e. 0.42408 > 2 x 0.09352 = 0.18704 [1]

This is a two-tailed test for significance at the 5% significance level, with z-value 1.96 (approximated by 2) – which is based on the null hypothesis of $\beta \sim N(0, 0.09352^2)$. Since $\beta > 2\ \times$ standard error (β), we can conclude that the parameter β for the variable 'tree density' is significant in the model.

[2]

(ii) (b) Using the Poisson canonical link function, we have:

$\eta = \log\mu = \alpha + \beta t$ [1]

So for $t = 12,\ \alpha = -1.54520$, and $\beta = 0.42408$:
$\log\mu = -1.54520 + (0.42408\ \times 12) = 3.54376$ [1]

So the expected number of bears that would be detected is:

$\mu = e^{3.54376} = 34.6$ bears ≈ 35 bears [1]

Part (i) was very well answered. Answers in part (ii) were problematic, with many candidates failing to apply their knowledge to correctly interpret the given model output.

Common errors included applying a 1-sided test in part (ii)(a) and not using the canonical link function correctly in part (ii)(b).

Q9

(i) $l(p) = constant + x\log p + (n-x)\log(1-p)$ [1]
$l'(p) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x(1-p)-(n-x)p}{(1-p)p} = 0$ [1]

$0 = x(1-p) - (n-x)p = x - np$
$\hat{p} = \frac{x}{n}$ [1]

(ii) $\pi(p) \propto L(p)f(p)$ [1]
$\propto p^x(1-p)^{n-x}\, p^{\alpha-1}(1-p)^{\beta-1} = p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$
$= p^{x+\alpha-1}(1-p)^{999-x+\beta}$ [1]

(iii) The posterior distribution is a beta distribution [1]
with parameters $x+\alpha$ and $n-x+\beta$. [1]

(iv) The prior distribution and the posterior distribution are of the same type. [1]
The beta distribution is the conjugate prior for the binomial distribution. [1]

(v) $\hat{p} = \frac{50}{1000} = 0.05$ [1]

(vi) Under quadratic loss, the Bayesian estimator is the expectation of the posterior distribution. [1]
In our case, this is $\hat{p} = \frac{x+\alpha}{x+\alpha+n-x+\beta} = \frac{x+\alpha}{n+\alpha+\beta}$ [1]
And for the given parameters we obtain $\hat{p} = \frac{52}{1004} = 0.0518$ [1]

(vii) The two estimates are almost identical meaning that the impact of the prior distribution is very limited and the Bayesian estimator is mainly determined by the data. [1]

(viii) We can write the posterior mean in credibility form as
$\hat{p} = \frac{x+\alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta} \times \frac{x}{n} + \frac{\alpha+\beta}{n+\alpha+\beta} \times \frac{\alpha}{\alpha+\beta}$ [1]

$\hat{p} = ZE[X] + (1-Z)E[p]$ [1]

with credibility factor $Z = \frac{n}{n+\alpha+\beta}$ [1]

Alternatively, the numerical answer can be given as

$$\frac{X+2}{1004} = \frac{1000}{1004} \times \frac{X}{1000} + \frac{4}{1004} \times \frac{2}{4}, \quad \text{where } Z = \frac{1000}{1004}.$$

The question was very well answered with many candidates achieving high marks across the different parts. In part (vii) valid comments different from those presented here, were also given full credit. There were some numerical slips in parts (v) and (vi).

Note that the wording "credibility interval" that is used in the part (viii) is not accurate and should have been "credibility estimate". The surrounding wording in the same part of the question is such that the possibility of misunderstanding is minimised. The examiners did not find evidence of this ambiguity having a negative impact on candidates' performance.

Q10

(i)
Start by calculating the sum of squares:

$S_{xx} = 59{,}054 - \frac{828^2}{12} = 1{,}922$ [½]

$S_{xy} = 1{,}334$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1{,}334}{1922} = 0.69407$ [1]

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 72 - 0.69407 \text{ x } 69 = 24.10926$ [1]

Hence, the fitted regression equation of y on x is: $\hat{y}$ = 24.10917 + 0.69407x . [½]

(ii) (a) $S_{yy} = 63{,}362 - \frac{864^2}{12} = 1{,}154$, [½]

so:

$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$ [½]

$= \frac{1}{10}\left(1{,}154 - \frac{1{,}334^2}{1{,}922}\right)$ [½]

$= 22.81124$ [½]

(ii) (b) Now $\frac{10\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{10}$, which gives a confidence interval for σ^2 of: [1]

$\left(\frac{10 \text{ x } 22.81124}{18.31}, \frac{10 \text{ x } 22.81124}{3.94}\right) = (12.46\,, 57.90)$ [1]

(iii) We test $H_0: \beta = 0 \; vs \; H_1 : \beta > 0$. [1]

Now $\frac{\hat{\beta}-\beta}{\sqrt{(\hat{\sigma}^2/S_{xx})}} \sim t_{10}$. The observed value here is:

$$\frac{0.69407 - 0}{\sqrt{\left(\frac{22.81124}{1922}\right)}} = 6.371$$

[1]

This is a significant result which exceeds the 0.5% critical value of 3.169. [1]

So there is sufficient evidence at the 0.5% level to reject H_0 , and conclude that $\beta > 0$ (i.e. that the data is positively correlated). [1]

(iv) The variance of the distribution of the second-part exam score corresponding to a first-part exam score of 57 is:

$$\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{12} + \frac{(57-69)^2}{1922}\right] \text{ x } 22.81124 = 3.610$$

[1]

The predicted value is 24.10917 + 0.69407 x 57 = 63.67116.

[1]

Using the t_{10} distribution, the 95% confidence interval is:

$$63.67116 \pm 2.228 \text{ x } \sqrt{3.610} = (59.44\,, 67.90)$$

[1]

(v) We are testing $H_0: \rho = 0.75 \; vs \; H_1 : \rho \neq \; 0.75$

If H_0 is true, then the test statistic Z_r has a $N\left(z_\rho, \frac{1}{9}\right)$ distribution, where:

$$z_\rho = \frac{1}{2}\log\frac{1+0.75}{1-0.75} = 0.9729551\,.$$

[½]

Pearson's correlation coefficient can be calculated as

$$r = \frac{1334}{\sqrt{1922 \times 1154}} = \; 0.89573$$

[1]

The observed value of this statistic is $z_r = \frac{1}{2}\log\frac{1+0.89573}{1-0.89573} = 1.45018$,

[½]

which corresponds to a value of $\frac{1.45018 - 0.9729551}{\sqrt{\frac{1}{9}}} = 1.431435$

[1]

on the N(0 , 1) distribution.

This is less than 1.96, the upper 2.5% point of the standard normal distribution.

So there is insufficient evidence at the 5% level to reject H_0 ,
i.e. the data do not provide enough evidence to conclude that the correlation parameter is different from 0.75.

[1]

(vi) The proportion of variability explained by the model is by R^2 :

$$R^2 = r^2 = \left(\frac{1{,}334}{\sqrt{1922 \text{ x } 1154}}\right)^2 = 0.802329 = 80.2\%$$

[1]

(vii) 80.2% of the variation is explained by the model, which indicates that the fit is very good.

[1]

The question was generally well answered. In part (iii) the test needs to involve the slope parameter for full marks (rather than, say, the correlation coefficient). Part (iv) asks for the "mean" response – a common error here was to provide the individual response.

END OF EXAMINERS' REPORT

@Institute and Faculty of Actuaries