Institute and Faculty of Actuaries

EXAMINERS' REPORT

CS1 - Actuarial Statistics

Core Principles

Paper B

April 2024

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
April 2024

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.

In particular, the CS1B paper is a problem-based examination and focuses on the assessment of computer-based data analysis and statistical modelling skills.

For the CS1B exam candidates are expected to include the R code that they have used to obtain the answers, together with the main R output produced, such as charts or tables.

When a question requires a particular numerical answer or conclusion, this should be stated explicitly and clearly, separately from, and in addition to, the R output that may contain the relevant numerical information.

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. In particular, there are variations of the R code presented here, which are valid and can produce the correct output. All mathematically and computationally valid solutions or answers received credit as appropriate.

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

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

Overall performance in CS1B was satisfactory. Well prepared candidates were able to achieve high marks.

Most candidates demonstrated sufficient knowledge of the key R commands required for the application of the statistical techniques involved in this subject.

In some occasions candidates failed to provide appropriate and informative annotation on produced graphs (e.g. parts of Q1, Q2).

Also, full R code was lacking in some cases. Candidates must include the R code, in full, used to obtain their answers, together with the main R output produced in their answers.

C. Pass Mark

The Pass Mark for this exam was 59.
1613 presented themselves and 771 passed.

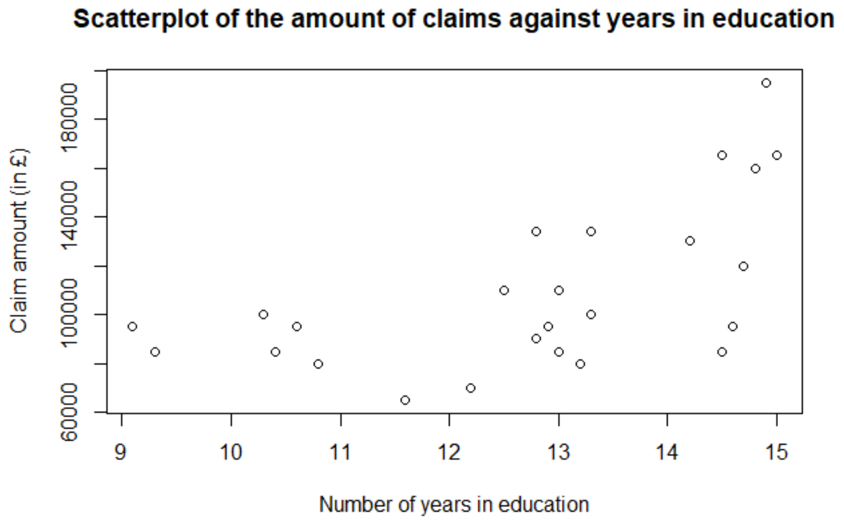
Solutions for Subject CS1B – April 2024

Q1

```
load("AmountYears.RData")
```

(i)

```
plot(EducationYears, ClaimAmount, xlab = "Number of
years in education", ylab = "Claim amount (in £)", main
= "Scatterplot of the amount of claims against years in
education")
```
[1]



[2]

(ii) (a)

```
m1 <-lm(ClaimAmount~EducationYears)
```
[1]

```
summary(m1)

Call:
lm(formula = ClaimAmount ~ EducationYears)

Residuals:
   Min     1Q Median     3Q    Max
-43747 -19875   1768  18575  61813

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      -32220      40159  -0.802  0.43058
```

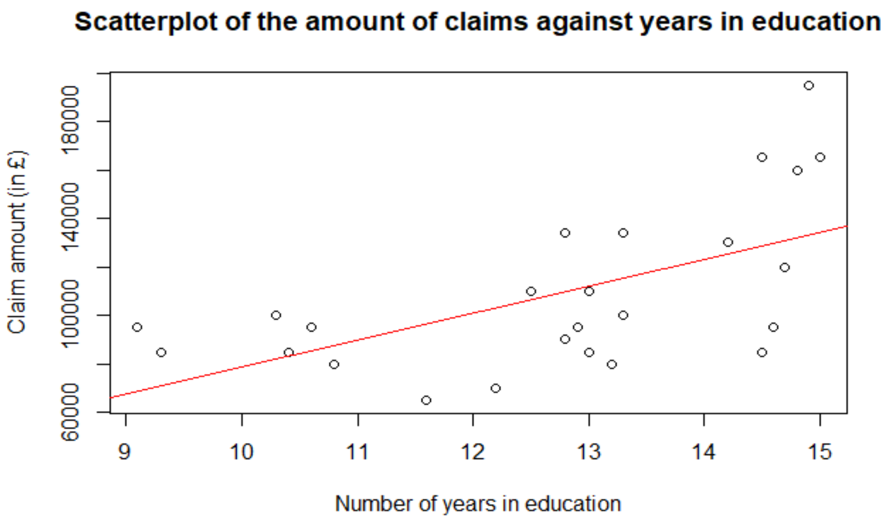
```
EducationYears    11101       3124   3.553  0.00169 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 27450 on 23 degrees of freedom
Multiple R-squared:  0.3544,    Adjusted R-squared:  0.3263
F-statistic: 12.62 on 1 and 23 DF,  p-value: 0.001695
```

[1]

(b)

```
abline(m1, col = "red")
```

[1]



[1]

(iii)(a)

```
m2 <- lm(ClaimAmount~EducationYears + I(EducationYears^2))
```

[2]

```
summary(m2)

Call:
lm(formula = ClaimAmount ~ EducationYears + I(EducationYears^2
))

Residuals:
   Min      1Q  Median      3Q     Max
```

```
-51368 -15047  -2360  14008  45625

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            609272     242570   2.512   0.0199 *
EducationYears         -95999      40154  -2.391   0.0258 *
I(EducationYears^2)      4371       1635   2.674   0.0139 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24380 on 22 degrees of freedom
Multiple R-squared:  0.5127,    Adjusted R-squared:  0.4684
F-statistic: 11.57 on 2 and 22 DF,  p-value: 0.000368
```

[1]

(b)

The fitted model is:

$\text{ClaimAmount} = 609272 - 95999 \times \text{EducationYears} + 4371 \times \text{EducationYears}^2$ [1]

(iv)

The quadratic model in part (iii) seems more suitable. [1]

The adjusted R-squared has increased. [1]

The coefficient of the quadratic term is significantly different from 0. [1]

[Total 14]

Commentary:
(i), (ii): Well answered in general. A number of candidates failed to provide informative annotation on the graphs. A small number of candidates presented the graphs with the axes reversed. *(iii): Mixed answers with some candidates not showing appropriate output.* *(iv): Not very well answered. A common error was referring to the unadjusted R-squared measure.*

Q2

(i)

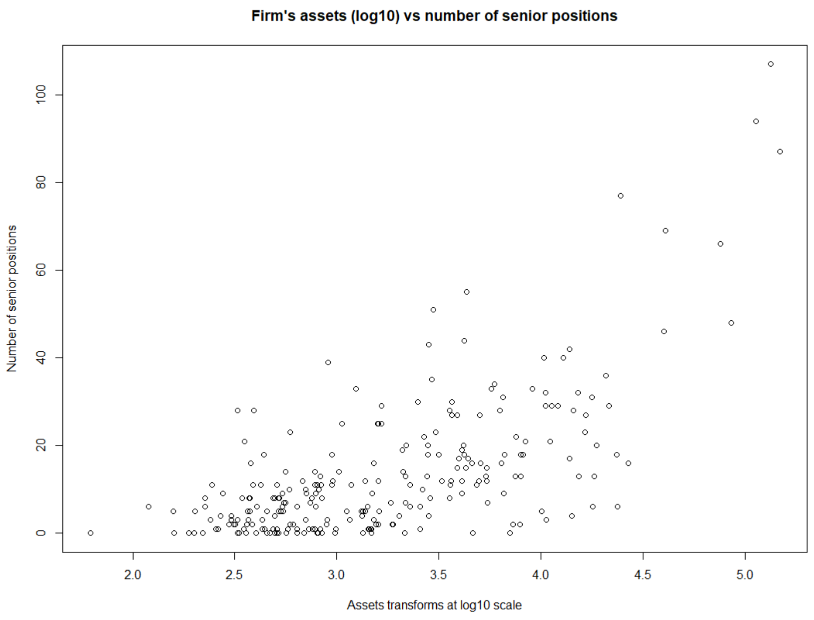
plot(assets,sn_positions, xlab="Assets", ylab="Number of senior positions", main="Firm's assets vs number of senior positions") [1]

[2]

(ii)

log10_assets = log10(assets) [½]

plot(log10_assets,sn_positions, xlab="Assets transforms at log10 scale", ylab="Number of senior positions", main="Firm's assets (log10) vs number of senior positions") [1½]



[2]

(iii)

The variable assets is extremely widespread in part (i). [1]

Taking the log10 of the assets in (ii) helps to show an increasing relationship between the variables. [1]

(iv)

mu = mean(sn_positions) [½]

mu

13.58065 [½]

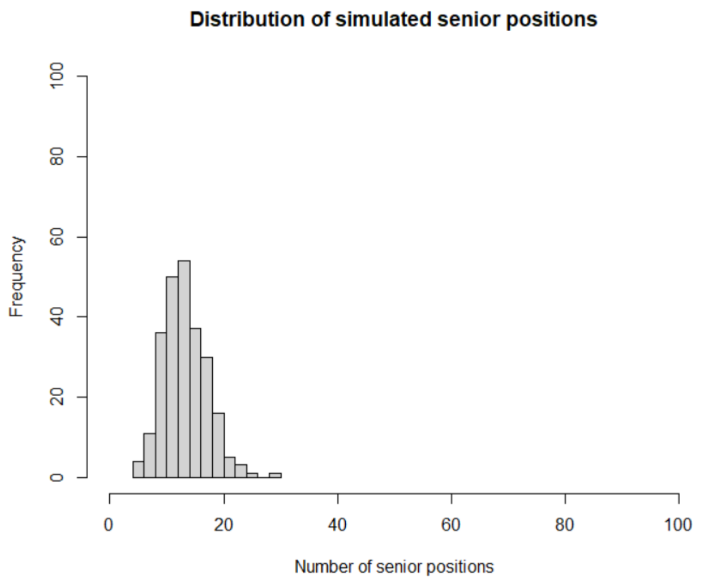
(v)

set.seed(222)

theoretical_samp = rpois(length(sn_positions), mu) [2]

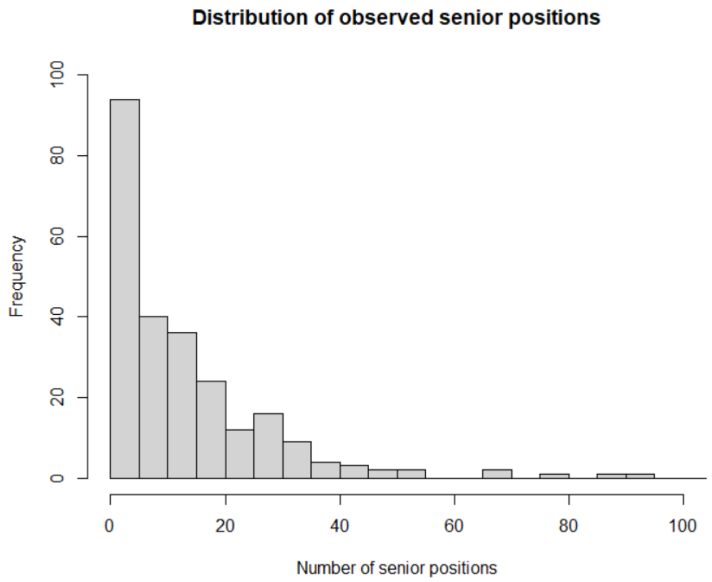
(vi)

hist(theoretical_samp,xlim=c(0,100), ylim=c(0,100), main = "Distribution of simulated senior positions", xlab="Number of senior positions") [1]



[2]

hist(sn_positions, breaks=20,xlim=c(0,100), ylim=c(0,100), main="Distribution of observed senior positions", xlab="Number of senior positions") [1]



[2]

(vii)

We can see that the two distributions are very different. [1]

For example, we can see an over-representation of zero values in the observed counts. [1]

Therefore, the number of senior positions sn_positions variable does not follow a Poisson distribution with parameter 13.58. [1]

[Total 21]

Commentary:
(i), (ii) Well answered overall. There were some answers where the graph annotation was not informative, and some cases with reversed axes. *(iii) Well answered, with most candidates receiving full or partial credit.* *(iv), (v) Very well answered.* *(vi) Mixed answers. While most candidates produced the two plots, in many cases the axes were not matched. Suboptimal titles and/or axis labels were used in some cases.* *(vii) Most candidates correctly observed that the distributions are different, but failed to comment specifically on the data not following the Poisson distribution that we have here.*

Q3

(i)

```
p = sum(CorrectOutOf20Questions)/(50 * 20)        [1]
p  # 0.568                                        [1]
```

•

(ii)

probability to pass the test is P[correct answers > 15]

```
1 - pbinom(15, 20, p)                             [1]
# 0.02747446                                      [1]
```

(iii)

```
sum(CorrectOutOf20Questions>15)/50                [½]

# 0.18                                            [½]
```

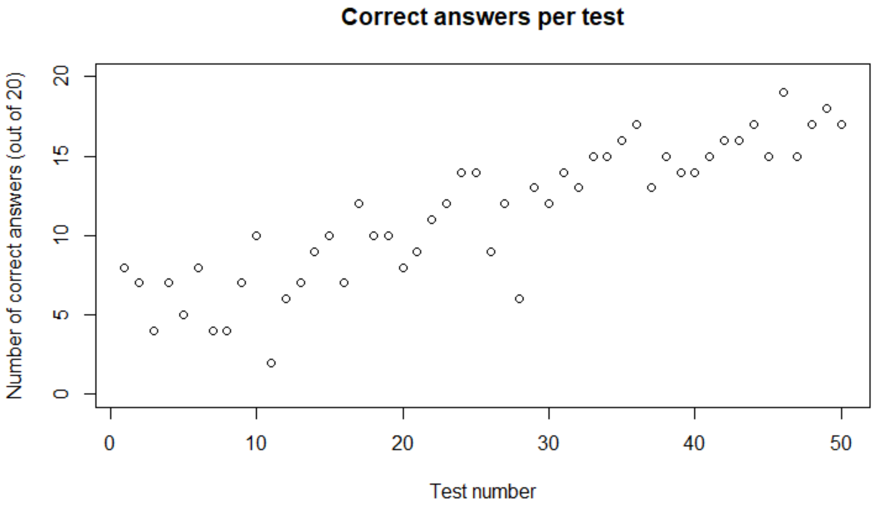
(iv)

The results in (ii) and (iii) are both estimates of the success probability in a MC test.
The difference is that the answer in part (ii) relies on the estimate from part (i) and therefore takes the whole distribution of test results (the actual points achieved) into account, [2]

while the answer in part (iii) turns the test results into a Bernoulli RV (1 for pass and 0 for fail). Therefore, in part (iii) it does not matter by how far away the result is from the pass mark while it does matter for the answer in part (ii). [2]

(v)

```
plot(TrialNumber, CorrectOutOf20Questions ylim=c(0,20),
main = "Correct answers per test",

ylab="Number of correct answers (out of 20)", xlab="Test
number")
```
[1]



[2]

(vi)

The plot in part (v) shows an upward trend, or positive correlation between test number and the number of correct answers. [1]

Suggesting that the student is improving their results with practicing. [1]

(vii)

Estimated probability of answering a question correctly:

```
p = 18.085/20
p # 0.90425
```
[1]
[1]

```
# prob for passing test

1 - pbinom(15, 20, p)   # 0.9632752
```
[2]

[Total 18]

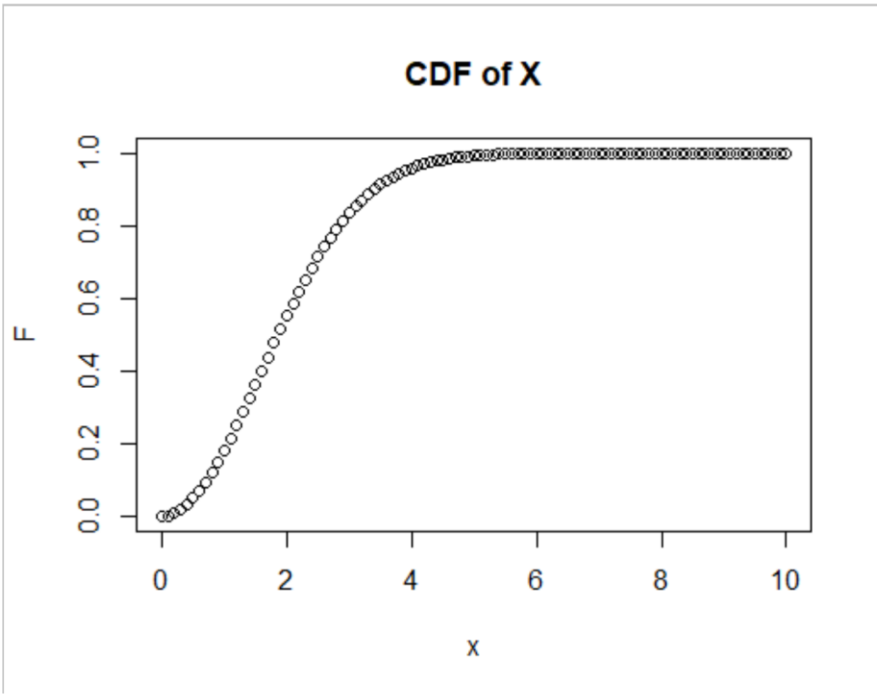
Commentary:
(i) Very well answered. *(ii) Well answered in general. A common error was using 16 instead of 15 in the CDF (pbinom) calculation.* *(iii) Well answered overall.* *(iv) Not answered well. A number of candidates failed to provide meaningful comments, with many repeating the numerical results.* *(v) Good answers. As in previous questions involving graphs, some candidates did not provide informative titles and/or axis labels.* *(vi) There were mixed answers in this part. Most candidates described the positive correlation, but few suggested improvement as a result of practice.* *(vii) Answers were mixed, with a number of candidates not attempting this part.*

Q4

(i)

```
cdf

x = seq(0, 10, by = 0.1)
lambda = 0.2
F = 1-exp(- lambda * x^2)
plot(x,F,main= "CDF of X")
```

[½] [½] [1] [1]



[1]

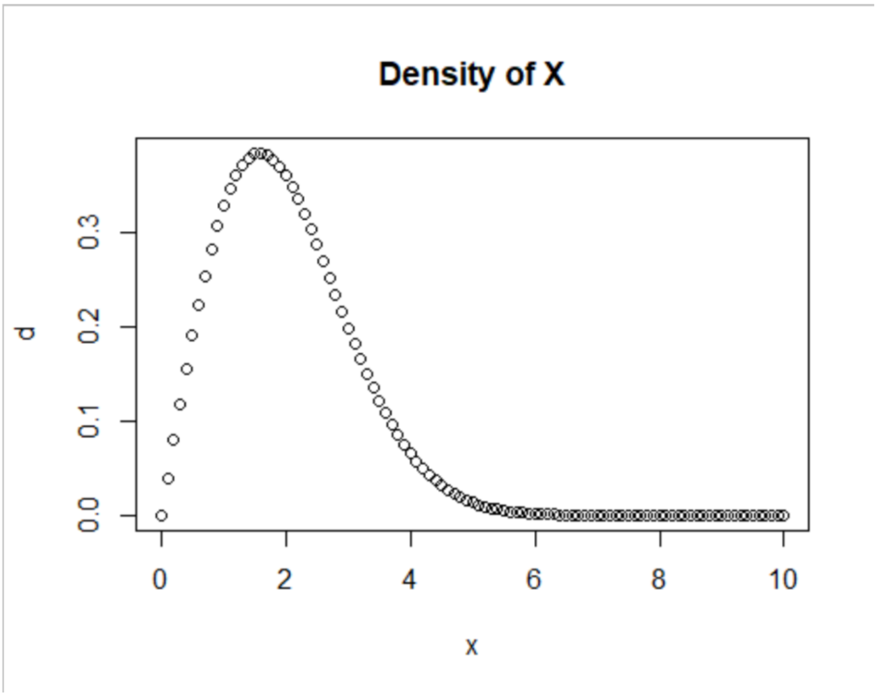
(ii) density

Differentiate F: $f(x) = \begin{cases} 0, & x < 0 \\ 2\lambda x \exp(-\lambda x^2), & x \geq 0 \end{cases}$ [2]

```
x = seq(0, 10, by = 0.1)
d = 2 * lambda * x * exp(-lambda * x^2)
plot(x,d,main="Density of X")
```

[1]

[1]



[1]

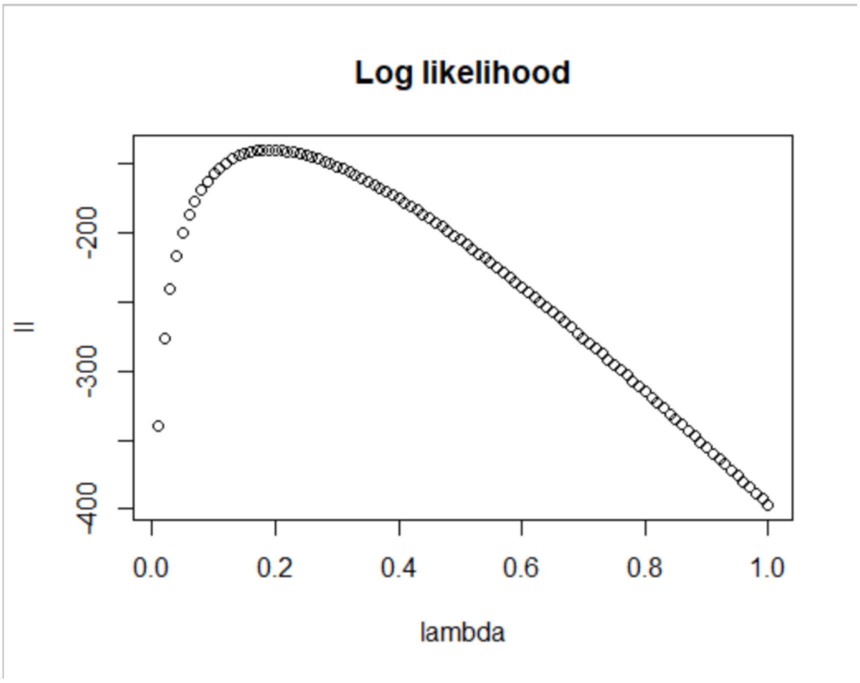
(iii)

```
#log likelihood
load("randomSample.Rdata")
d = 2 * lambda * x * exp(-lambda * x^2)          [1]
ll = sum(log(d))                                 [1]
ll    #   -139.7437                              [1]
```

(iv)

```
#log likelihood for 100 values of lambda
ll = 1:100                                       [1]
lambda = 1:100
for (k in 1:100) {                               [1]
lambda[k] = 0.01 * k                             [1]
ll[k] = sum(log(2 * lambda[k] * x * exp(-lambda[k] * x^2)))}[2]
```

```
plot(lambda,ll,main="Log likelihood")
```
[1]



[1]

(v)

```
N = length(x)   # sample size
lambda.hat = N/sum(x^2)
lambda.hat      #  0.191256
```
[1] [1] [1]

(vi)

The plot of the log likelihood function shows a maximum near 0.2, which is the MLE. [1]

That is consistent with the estimate obtained in part (v). [1]

[Total 24]

Commentary:
(i) Well answered in general. *(ii) Mixed answers. A common issue was errors in the differentiation.* *(iii) Well answered in general.* *(iv) There were mixed answers here, with many candidates producing the correct plot, but*

there were various errors in many cases.
(v) Well answered in general. A small number of candidates failed to square x.
(vi) Well answered in general.

Q5

Load the data:

```
> load("ClaimsData.Rdata")
```

(i)

```
> model_lin <- lm(claim_count ~ age)                    [1]
> model_lin$coefficients
(Intercept)         age
 4.29414532 -0.05026053
```

The intercept is 4.29 and the slope is -0.05. [2]

(ii)

```
> model_glm_age <- glm(claim_count ~ age, family = poisson)  [2]
> model_glm_age
#[1] Call:  glm(formula = claim_count ~ age, family = poisson)


Coefficients:
(Intercept)          age
    1.70492     -0.02323


Degrees of Freedom: 95 Total (i.e. Null);  94 Residual
Null Deviance:     149.5
Residual Deviance: 128.2      AIC: 349.8
```

[2]

(iii)

A comparison using scaled deviances is only valid for nested models. [1]

The model in (i) assumed a Normal distribution, whilst the model in (ii) assumed a Poisson distribution. Therefore, neither model is a subset of the other. [2]

(iv)

```
> model_lin_glm <- glm(claim_count ~ age, family = gaussian)
```
[2]

```
> model_lin_glm

#[1] Call:  glm(formula = claim_count ~ age, family =
gaussian)


Coefficients:

(Intercept)          age

    4.29415     -0.05026


Degrees of Freedom: 95 Total (i.e. Null);  94 Residual

Null Deviance:     295.8

Residual Deviance: 249.3      AIC: 370
```
[2]

(v)

The AIC of the Poisson model in part (ii) is lower than that of the Normal model in part (iv). [1]

Therefore, the Poisson GLM provides a better fit. [1]

(vi) (a)

```
> model_glm_full <- glm(claim_count ~ age*gender, family =
poisson)
```
[2]

```
> model_glm_full
```

```
#[1] Call:  glm(formula = claim_count ~ age * gender, family =
poisson)


Coefficients:

(Intercept)          age      genderM  age:genderM

    1.84642     -0.03459     -0.18217      0.01805


Degrees of Freedom: 95 Total (i.e. Null);  92 Residual

Null Deviance:     149.5

Residual Deviance: 113.5      AIC: 339
```

[2]

(b)
H0: the bigger (full) model is NOT a significant improvement over the simpler model; against H1: it is a significant improvement. [½]

```
> anova(model_glm_age, model_glm_full, test="Chisq")
```

[½]

```
Analysis of Deviance Table

Model 1: claim_count ~ age
Model 2: claim_count ~ age * gender

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        94     128.25
2        92     113.46  2   14.789 0.0006148 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[½]

The p-value is clearly significant (less than 0.1%). [½]

Therefore, the GLM with age*gender is a significant improvement over that with gender alone. [1]

[Total 23]

Commentary:
(i) Well answered. A fair number of candidates failed to pull out the intercept and slope estimates explicitly. *(ii) Well answered. As in part (i), a fair number of candidates failed to pull out the coefficient estimates and AIC explicitly.* *(iii) Not well answered. Only a relatively small number of candidates referenced nesting.* *(iv) Well answered in general. A number of candidates failed to present the coefficient estimates and the value of the AIC explicitly.* *(v) Very well answered.* *(vi)(a) Well answered in general. A number of candidates failed to pull out the coefficient estimates and AIC value explicitly.* *(vi)(b) Not well answered. Only a small number of candidates performed a proper comparison (e.g. anova) with reference to statistical significance in the difference of the deviances.*

[Paper Total 100]

END OF EXAMINERS' REPORT

Institute and Faculty of Actuaries

www.actuaries.org.uk