INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2019 Examinations

Subject CS1 – Actuarial Statistics Core Principles (Part B)

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
September 2019

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

3. 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.

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

5. Annotated plots and relevant comments should be provided when instructed to do so in the question.

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

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

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

9. In cases where a question is based on simulations, all numerical answers provided in this document are examples of possible results. The numerical values presented here will be different if the simulations are repeated.

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

1. Overall performance in CS1B was satisfactory and slightly stronger compared to the April 2019 session. Well prepared candidates were able to score highly.

2. Candidates demonstrated a good knowledge of the key R commands required for the application of the statistical techniques involved in this subject.

3. The quality of the commentary given alongside the R output was not particularly strong and varied significantly among candidates.

C. Pass Mark

The combined pass mark for CS1 in this exam diet was 55.

Solutions Subject CS1 – B

Q1

```
x= c(0.22, 0.38, 1.28, 0.54, 0.56, 1.36, 0.55, 0.37, 0.43,  0.46,
0.62, 0.54, 0.54, 0.51, 0.44, 0.68, 0.55, 0.30)
```

(i) [2]

```
mean(x)
```

0.5738889
or use answer from t.test in part (ii)

(ii) [3]

```
t.test(x, conf=0.95)
```

[0.4276654, 0.7201123]

(iii) [3]

```
y = replicate(10000,mean(sample(x,replace =TRUE)))
```

[2]

```
quantile(y,prob=c(0.025,0.975))
```

[0.4572222, 0.7200000]

(iv) The CIs are almost identical with the CI in part (iii) being narrower. [1]

The reason might be that the distribution of X is not normal, and therefore the distribution of the mean is not normal for the small sample size we have in this question. [1]

The bootstrap method provides a good approximation of the distribution of the mean independently of the type of distribution. [1]

[Total 13]

Parts (i) and (ii) were very well answered. In part (ii) many candidates calculated the CI using the algebraic route rather than the t.test() function. This is valid but more time-consuming. Answers in part (iii) were mixed, with a number of candidates applying parametric bootstrap. In part (iv), most candidates noted the similarities or differences between the two intervals, but did not comment on possible reasons.

Q2

(i) (a)

```
# read the data

> interest_rates<-read.csv("Interest_rates.csv")

# calculate the Pearson correlation coefficients
```

```
> C<-cor(interest_rates, method="pearson")
```

[2]

Alternative solution

```
cor(interest_rates)

> C
           X1.year   X5.year  X10.year  X15.year  X20.year  X30.year
X1.year  1.0000000 0.8760513 0.7827029 0.7380130 0.6146684 0.4094968
X5.year  0.8760513 1.0000000 0.9666006 0.9356486 0.8269608 0.6020591
X10.year 0.7827029 0.9666006 1.0000000 0.9907851 0.9339235 0.7657966
X15.year 0.7380130 0.9356486 0.9907851 1.0000000 0.9696405 0.8315513
X20.year 0.6146684 0.8269608 0.9339235 0.9696405 1.0000000 0.9379375
X30.year 0.4094968 0.6020591 0.7657966 0.8315513 0.9379375 1.0000000
```

[2]

(b)
The correlation matrix shows that there is strong (positive) correlation between returns on bonds of similar maturity. [2]

It also shows that the correlation between returns is weaker as the length of maturity between bonds increases. [2]

(ii) (a)

```
# carry out principal component analysis using SVD method

> pca<-prcomp(interest_rates)
```

[3]

```
# review the results of the principal component analysis

> summary(pca)
Importance of components:
                           PC1      PC2      PC3       PC4       PC5
Standard deviation     0.00945 0.003419 0.001488 0.0002555 0.0002061
Proportion of Variance 0.86432 0.113130 0.021430 0.0006300 0.0004100
Cumulative Proportion  0.86432 0.977440 0.998880 0.9995100 0.9999200
                             PC6
Standard deviation     8.986e-05
Proportion of Variance 8.000e-05
Cumulative Proportion  1.000e+00
```

[3]

```
> pca<-princomp(interest_rates)

> summary(pca)
Importance of components:
                            Comp.1      Comp.2      Comp.3       Comp.4
```

```
Standard deviation     0.009317648 0.003370953 0.001467303 0.0002519268
Proportion of Variance 0.864317887 0.113127014 0.021433880 0.0006318435
Cumulative Proportion  0.864317887 0.977444901 0.998878781 0.9995106243
                             Comp.5       Comp.6
Standard deviation     0.0002032373 8.860665e-05
Proportion of Variance 0.0004112140 7.816163e-05
Cumulative Proportion  0.9999218384 1.000000e+00
```

(b)
The R-output shows that the proportion of variance explained by the first two principal components is c.98%, and by the first three components c.99%. [2]

Therefore it would be reasonable to reduce the dimensions of the dataset by using the first two (or three) principal components. [2]

[Total 18]

Candidates performed generally well in this question. Part (i) was very well answered, with some partial answers in (i)(b) where many candidates observed a relationship in individual years without drawing out the overall trend. Answers in part (ii) were also satisfactory. Note that in part (ii) the princomp() function can alternatively be used in R.

Q3

```
set.seed(2019)
```

(i) The distribution of Y is a Gamma distribution, [1]

$Y \sim Gamma(n, \lambda)$ [2]

(ii)

```
n = 15
lambda = 2
x = rexp(n, lambda)                                   [2]
```

(iii)

```
y = sum(x)                                            [1]
```

y= 10.7205

(iv)

```
y = 0*(1:1000) # generate a vector of size 1000       [1]
   for (i in 1:1000){                                 [4]
     y[i] = sum(rexp(n, lambda))                      [3]
   }
```

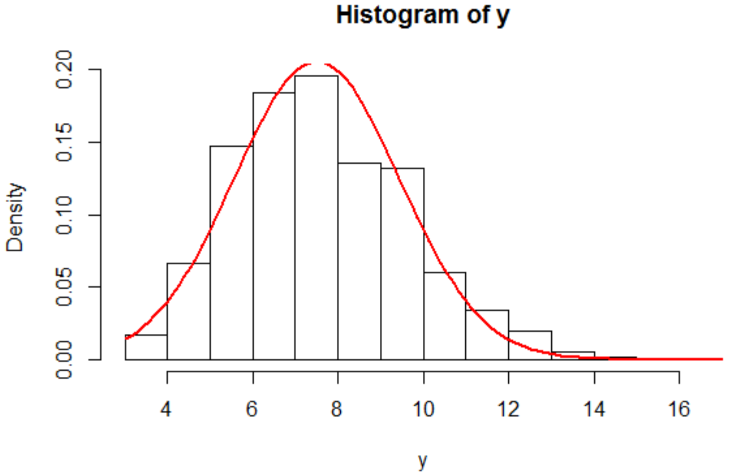
(v)

```
hist(y, prob=TRUE)                                    [2]
```

(vi) (a)

```
curve(dnorm(x,mean=n/lambda,
 sd=sqrt(n/(lambda^2))), add=TRUE, lwd=2, col="red")
```

[2]



(b) In contrast to a normal distribution, the histogram is clearly not symmetrical. [1]

This comes from the fact that Y can take only positive values.

For a larger sample size of n of $x_1, \ldots, x_n$ (not a larger B) the CLT ensures that the distribution of Y becomes approximately normal. [2]

[Total 21]

Parts (i)-(iii) were generally very well answered. In part (iv) there was wide variation in the quality of the answers with various errors in the details of the code. A common error in part (v) was omitting the (prob = TRUE) part of the code, which is required for relative frequencies. Part (vi)(a) was reasonably well attempted, while many candidates did not attempt part (vi)(b), with some giving incomplete comments. Part (vi)(a) can alternatively be answered using R code based on the lines() command.

Q4

(i)

```
> x = c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)

> y = c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)

> plot(x,y)      # make a plot
```

[1]

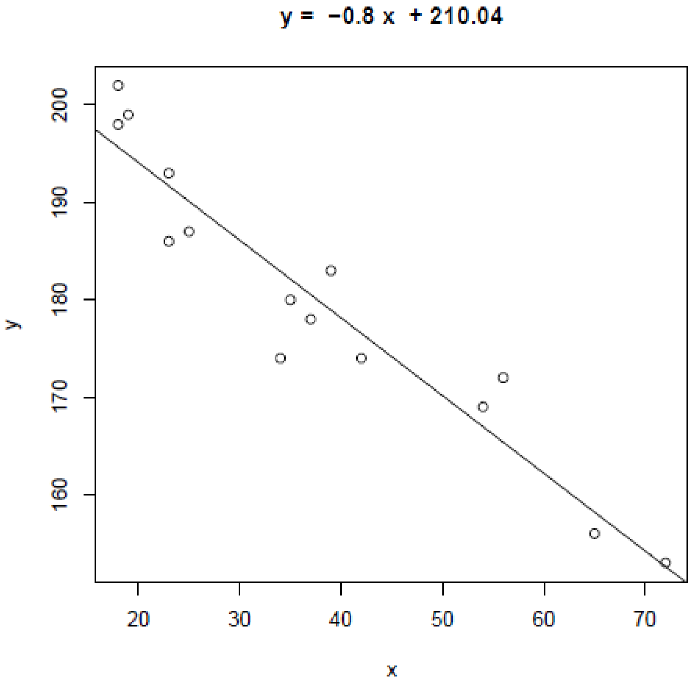
```
# Obtain the basic values of the regression analysis
```

```
> lm.result = lm(y ~ x)
> lm.result
     Call:
     lm(formula = y ~ x)
     Coefficients:
     (Intercept) x
     210.0485 - 0.7977
```

[2]

```
> abline(lm(y ~ x))   # plot the regression line
```

[1]



[1]

(ii) Max heart rate reduces as age increases. The fit of the model seems good. [2]

(iii) Can do a test to see if the slope of -1 is correct. Let H_0 be that $\beta = -1$, and H_A be that $\beta \neq -1$. Then we can create the test statistic and the p-value as follows:

```
> es = resid(lm.result)     # the residuals lm.result
> b1 =(coef(lm.result))[['x']] #the x part of the coefficients
> n = 15
> s = sqrt( sum( es^2 ) / (n-2) )
> SE = s/sqrt(sum((x-mean(x))^2))
> t = (b1 - (-1) )/SE
# find the right tail for this value of t with 15-2 d.f.
> pt(t,13,lower.tail=FALSE)
# [1] 0.006310157
```

[6]

The p-value is twice this as the problem is two-sided,
i.e. 2*0.006310157 = 0.01262031 [1]

(iv) The null hypothesis is rejected at the 5% level of significance. The slope may not be equal to -1 for these data. (Which is the slope predicted by the original formula 220 - Age.) [2]

[Total 16]

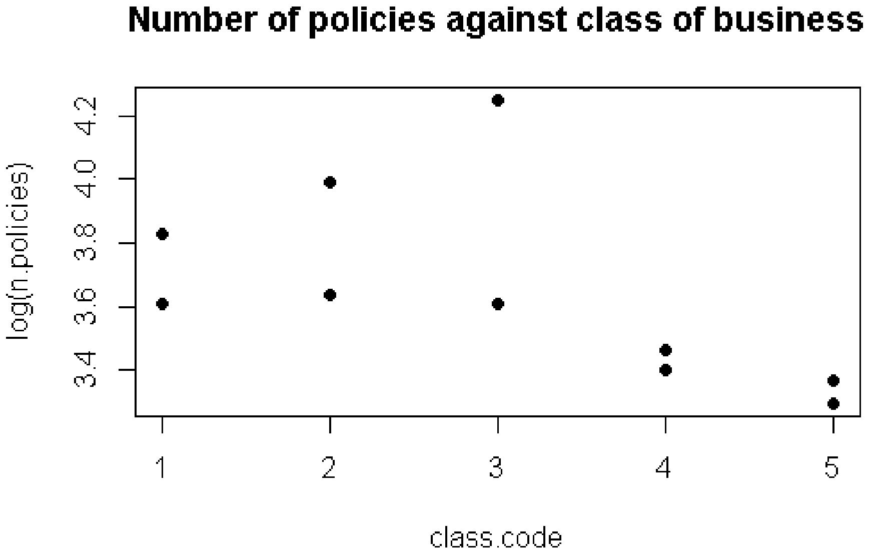
Well answered overall. In part (i) a number of candidates did not show the fitted line on the graph as required. In part (ii) many candidates failed to comment on the relationship between the two variables. Part (iii) was answered well only by well prepared candidates. In part (iv) answers using a different level of significance and consistent conclusion were given credit as appropriate.

Q5

(i) (a)

```
load("policies_data.RData")

plot(log(n.policies) ~ class.code, pch=19,main = "Number
of policies against class of business")
```
[1]



[1]

(b)

There seems to be some dependence of number of claims on class of business [1]

with lower numbers for classes 4 and 5. [1]
The relationship is not clear though. [1]

(ii) It now seems that the number of claims also depends on the gender of policyholders. [1]
The numbers are generally higher for males. [1]

(iii) R code:

```
class.code = as.factor(class.code)
sex.code = as.factor(sex.code)

glm1 = glm(n.policies ~ class.code, family = "poisson")
```
[2]
```
summary(glm1)
```
[1]

```
#Coefficients:
#             Estimate    Pr(>|z|)
# (Intercept)   3.7257     <2e-16 ***
# class.code2   0.1029     0.4965
# class.code3   0.2540     0.0825 .
# class.code4  -0.2917     0.0822 .
# class.code5  -0.3935     0.0229 *
```

Parameter estimates and their p-values are shown above.

Business class 1 is used as the baseline category (intercept level). [1]

The effect of class 5 on the number of policies appears to be significantly different from that of class 1, and there is some (weak) evidence that classes 3 and 4 also have a different effect. [4]

(iv) R code:

```
glm2 = glm(n.policies ~ class.code + sex.code, family =
"poisson")
```
[2]
```
summary(glm2)
```
[1]

```
#Coefficients:
#             Estimate    Pr(>|z|)
#(Intercept)   3.8611     < 2e-16 ***
#class.code2   0.1029     0.49648
#class.code3   0.2540     0.08248 .
#class.code4  -0.2917     0.08225 .
#class.code5  -0.3935     0.02288 *
```

```
#sex.code2    -0.2921     0.00386 **
```

Numbers of policies depend on both business class and gender of policyholder. [1]

Business class 5 has the strongest effect on number of policies when compared to class 1, and this effect is negative (reducing number of policies). Male policyholders give the baseline here, so being female has a significant negative effect on number of policies. [4]

(v) The null hypothesis is that the second model (including both factors) is *not* an improvement over the first model. [1]

R code:

```
anova(glm1,glm2,test = "Chisq")
```

[2]

```
#Model 1: n.policies ~ class.code
#Model 2: n.policies ~ class.code + sex.code
#  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
#1         5    14.2560
#2         4     5.8163  1   8.4397 0.003671 **
```

[1]

The p-value is 0.003671. [1]

Therefore we have strong evidence against the null hypothesis. We conclude that the second model gives significant improvement. [1]

(vi) R code:

```
predict(glm2, data.frame(class.code="2", sex.code="1"),
type="response")
```

[2]

Based on model 2 we predict 52.67 policies. [1]

[Total 32]

The performance in this question was mixed. Part (i) was generally well answered, but the comments in (b) were often insufficient. In part (ii) there was a range of comments of varying validity. In parts (iii) and (iv) many candidates fitted the GLM successfully but produced weak comments. Note that part (iv) can alternatively be answered using the update() command. Parts (v) and (vi) were not well answered, with a number of candidates failing to use the deviance as requested and instead opting for the AIC.

END OF EXAMINERS' REPORT