INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2022

Subject CS1B – Actuarial Statistics
Core Principles

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 Specialist Advanced (SA) and Specialist Principles (SP) 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.

Sarah Hutchinson
Chair of the Board of Examiners
July 2022

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 explicitly and clearly stated, 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.

In cases where a question is based on simulations, and no seed was specified in the question, 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 *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 certain parts of the exam paper, candidates provided answers to particular question parts while answering different parts of the question (e.g. in Question 3). Cross marking was used to give credit where appropriate.

Candidates scored lower in questions with unusual style (e.g. Question 3(i)-(iv)) despite the tested topics being standard basic statistical concepts from the CS1 Core Reading. This highlights the need for candidates to cover the whole syllabus when they revise for the exam and not rely heavily on questions appearing in recent papers.

C. Pass Mark

The Pass Mark for this exam was 59.
1311 presented themselves and 579 passed.

Solutions for Subject CS1B – April 2022

Q1

(i)
Let *X* be the total weight of 8 people. By the assumption: $X \sim Normal(560\,, 57^2)$. We are interested in the probability P(X > 650):

```
1 - pnorm(650,560,57)                                    [1]
# 0.05717406
```

The probability that the total weight of 8 people exceeds 650kg is equal to 0.05717406. [1]

(ii)
Let *Y* be the total weight of 9 people. By the assumption: $Y \sim Normal(630\,, 61^2)$. We are interested in the probability P(Y > 650):

```
1 - pnorm(650,630,61)                                    [1]
# 0.3715054
```

We get that the probability that the total weight of 9 people exceeds 650kg is much higher and is equal to 0.3715054. [1]

(iii)
Parts (i) & (ii) show that as more people enter the lift, the probability of exceeding 650kg increases. [½]

While the probability of exceeding 650kg is small with 8 people, [½]
exceeding the maximum weight is considerably more likely with 9 people. [1]

(iv)
Again, $X \sim Normal(560, 57^2)$, where *X* is the total weight of 8 people.
In order to find the central region that contains 80% of the distribution we need to identify the 10%-percentile and the 90%-percentile of *X*.

We use the function "qnorm" in the code:

```
> qnorm(0.1,560,57)                                      [1]
# 486.9516
```

```
> qnorm(0.9,560,57)                                      [1]
# 633.0484
```


The requested interval is [486.9516, 633.0484]. [1]

(v)(a)
We now have $Y \sim Gamma(96.5220, 0.1724)$, where Y is the total weight of 8 people. The computation this time produces:

qgamma(0.1, 96.5220, 0.1724) [½]
488.195

> qgamma(0.9, 96.5220, 0.1724) [½]
634.0333

and the interval is [488.195, 634.0333]. [½]

(b)
The intervals are very similar. [½]
The mean and standard deviation of the two distributions are (approximately) equal. [½]
As the first (shape) parameter of the gamma distribution is large, the distribution is close to a normal distribution. [½]

[Total 12]

Candidates overall answered well this question.

In parts (i) and (ii) a common error was calculating the probability at the wrong tail, e.g. using pnorm() instead of 1-pnorm().

In part (iv) some candidates provided a wrong answer giving a single value with qnorm(0.8, …).

In part (v) (b), well prepared candidates referenced the gamma shape parameter.

Q2
(i)(a)
R code:

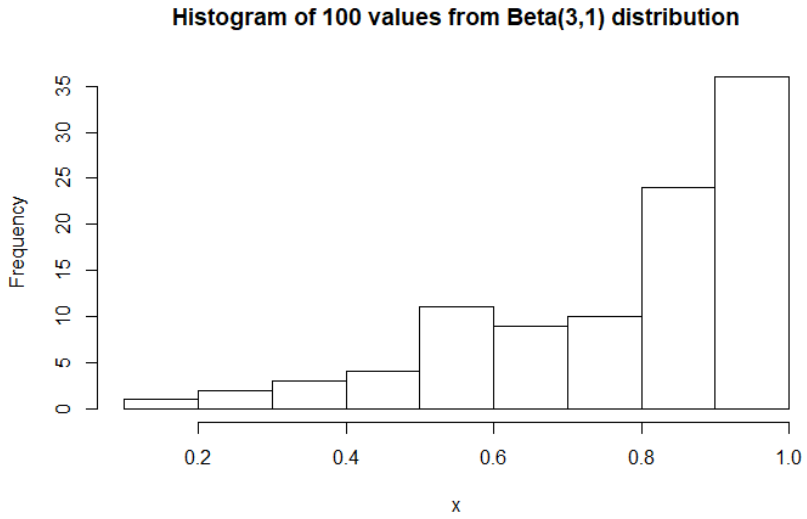
```
set.seed(12345)
x = rbeta(100,3,1)
```

[1]

(b)
R code:

```
hist(x, main="Histogram of 100 values from Beta(3,1) distribution")
```

[1]



[2]

(c)
The histogram is heavily skewed. [1]
This is consistent with the skewness of a Beta(3,1) distribution, which is negative for ($\alpha > \beta$), see "Formula and Tables …", page 13. [1]

(ii)(a)
R code:

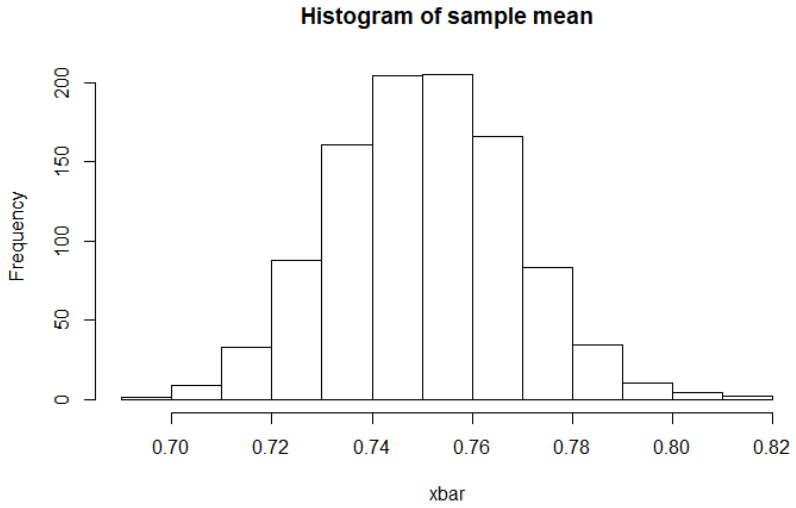
```
set.seed(12345)
nsim = 1000                                  [½]
xbar = numeric(nsim)                         [1]
for (i in 1:nsim){                           [1½]
     x = rbeta(100,3,1)                      [1]
     xbar[i] = mean(x)                       [1]
```

Alternative code may be used. For example, without using a loop:

```
set.seed(12345)
xbar <- replicate(1000,mean(rbeta(100,3,1)))
```

(b)

```
hist(xbar, main="Histogram of sample mean")  [1]
```

[1]

(c)
The distribution of the sample mean is roughly symmetrical. [1]
This demonstrates the CLT where the distribution of the sample mean is approximately normal for large sample size. [1]

(iii)
R code:

```
y = c(4.9, 3.3, 2.2, 2.3, 1.6, 2.4, 4.7, 1.4, 1.7, 5.1)
t.test(y, conf.level =0.90)                                    [1]
# 2.124776 3.795224
```

Alternative R code:

```
mean(y)-qt(0.95, length(y)-1)*sd(y)/sqrt(length(y))
mean(y)+qt(0.95, length(y)-1)*sd(y)/sqrt(length(y))
```

We have assumed that the data come from a normal distribution. [1]
CI is given by (2.125, 3.795) [1]

(iv)
R code:

```
se.t = sd(y)/sqrt(length(y)); se.t                             [1]
# 0.4556314
```

Standard error of sample mean = 0.456. [1]

(v)
R code:

```
set.seed(12345)
nsim = 10^4                                                    [½]
ybar.sim = numeric(nsim)                                       [½]
for (i in 1:nsim){                                             [1]
y.sim = sample(y, replace=T)                                   [2]
ybar.sim[i] = mean(y.sim)                                      [1]
```



```
se.boot = sd(ybar.sim); se.boot                                    [1]
# 0.4318923
```

Bootstrap standard error of sample mean = 0.432. [1]

Alternative code may be used. For example, without using a loop:

```
set.seed(12345)
sd(replicate(10000, mean(sample(y, replace = T))))
```

(vi)(a)
We can use the output from part (v) and the R code:

```
boot.ci.1 = quantile(ybar.sim, c(0.05,0.95)); boot.ci.1      [2]
# 2.27 3.69
```

The 90% CI is (2.27, 3.69). [1]

(b)
The CI is now narrower; [1]
this suggests that the data may not be from a Normal distribution [1]
and the statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ may not follow a t_9 distribution, as suggested in part (iii). [1]

[Total 33]

Candidates overall answered well this question.

In certain parts requiring the use of the seed() command, some candidates presented numerical and graphical answers that differed from those shown in the solutions, despite following the correct procedure. This suggests that the seed() command was not used properly throughout the question.

In part (i)(c) alternative comments, e.g. relating to the mean or variance of the distribution, received credit as appropriate.

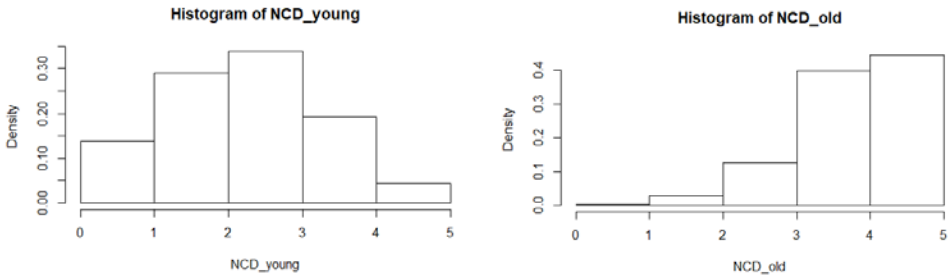
In part (ii)(b) (and similar parts), presenting both the code and the resulting graph is required for full marks.

In part (iv) a common error was to give the standard deviation of y as the final answer. In part (v) a range of coding variations were presented and received appropriate credit where correct. Approaches using parametric bootstrap assuming a normal distribution were also given credit.

In part (vi)(b), similar reasonable comments were given credit as appropriate.

Q3
(i)
Plot histograms for no claims discount years:

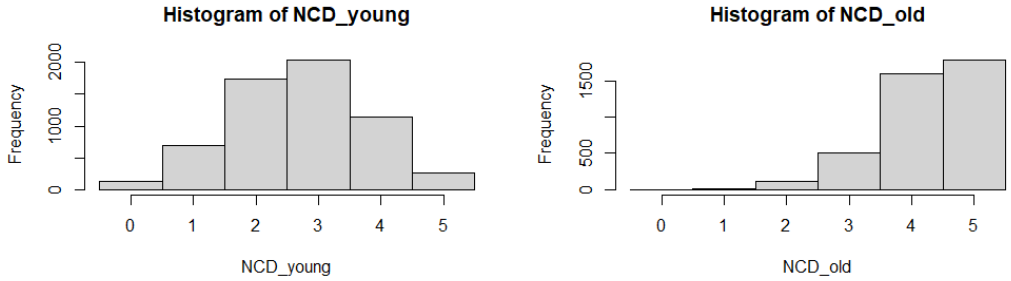

```
NCD_old = NCD[age==1]                                    [1]
NCD_young = NCD[age==0]                                  [1]
hist(NCD_old, prob=TRUE, breaks = c(0,1,2,3,4,5))        [1]
hist(NCD_young, prob=TRUE, breaks = c(0,1,2,3,4,5))      [1]
```



[1]

Alternative R code (this produces more accurate plots):

```
hist(NCD_young, breaks=seq(-0.5,5.5,by=1))
hist(NCD_old, breaks=seq(-0.5,5.5,by=1))
```



(ii)

The distribution for old policyholders is more concentrated on the right, [1]
so the number of years of NCD seems to be higher for old policyholders. [1]

(iii)(a)

Calculate proportions:

```
prop_young = sum(NCD_young > 2)/length(NCD_young)        [1½]
prop_young
```

Proportion is 0.5736150 [½]

Alternative R code:

```
length(NCD_young[NCD_young>2])/length(NCD_young)
```

(b)

```
prop_old = sum(NCD_old > 2)/length(NCD_old)              [1½]
prop_old
```

Proportion is 0.9704187 [½]

Alternative R code:

```
length(NCD_old[NCD_old>2])/length(NCD_old)
```


(iv)
Test the hypothesis that prop of NCD >2 is equal for young and old, against alternative that proportions are not equal: [1]

```
NCDgreaterTwo = c(sum(NCD_young > 2), sum(NCD_old > 2))      [1]
YoungOld = c(length(NCD_young), length(NCD_old))             [1]
```

Two-sample test for equality of proportions:

```
prop.test(NCDgreaterTwo, YoungOld, correct=F)                [2]

data:  NCDgreaterTwo out of YoungOld
X-squared = 1924, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.4103664 -0.3832408
sample estimates:
   prop 1    prop 2
0.5736150 0.9704187
```

The *p*-value is very small, therefore we reject the null hypothesis of equal proportions. [1]

(v)
R code:

```
glm(claims ~ age + LY + NCD, family = poisson)
```

or

```
glm(claims ~ factor(age) + LY + NCD, family = poisson)       [2]

summary(glm(claims ~ age + LY + NCD, family = poisson))

Call:
glm(formula = claims ~ age + LY + NCD, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7864  -1.1705  -0.0495   0.5274   3.9661

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.467231   0.034983  13.356  < 2e-16 ***
age         -0.009214   0.033348  -0.276    0.782
LY          -0.113913   0.021145  -5.387 7.16e-08 ***
NCD         -0.095156   0.019703  -4.829 1.37e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 11574  on 9999  degrees of freedom
Residual deviance: 11015  on 9996  degrees of freedom
AIC: 23104
```


[1]

(vi)
We need to fit the three possible models with two variables each and compare the AIC values:

```
glm(claims ~ LY + NCD, family = poisson)$aic
# 23102.41
```
[2]

```
glm(claims ~ age+LY, family = poisson)$aic
# 23125.08
```
[2]

```
glm(claims ~ age+NCD, family = poisson)$aic
# 23131.97
```
[2]

Model LY + NCD has the smallest AIC value and should be preferred. [2]

(vi)
The AIC for model LY + NCD (23102) is slightly lower than the AIC for age + LY + NCD (23104), but the AIC values are very close. [1]
We would prefer model LY + NCD as it also involves fewer variables (it is a smaller model). [1]

[Total 30]

Candidates overall answered well this question. A number of candidates did not attempt parts (i)-(iv).

In part (ii), alternative comments, for example in terms of the skewness or other summary statistics, received credit as appropriate.

In part (iv), using the continuity correction results in the same p-value. Also in part (iv), a variety of alternative tests were attempted and received credit where appropriate. The later parts were well attempted.

In part (vi) common issues included candidates failing to compare models, focusing instead on variable age not being significant (this only checks the significance of one variable in the presence of others, rather than comparing models). Also in part (vi), alternative answers based on analysis of deviance, received credit when applied correctly.

Q4

(i)
First set up the calculations required for the EBCT 2 model:

```
> load("claims.Rdata")
> n<-ncol(policies_matrix)
```
[½]

```
> N<-nrow(policies_matrix)
```
[½]

```
> X<-claims_matrix/policies_matrix
```
[½]

```
> X
```



```
          [,1]       [,2]       [,3]       [,4]       [,5]
[1,] 0.05208526 0.04971335 0.03587963 0.03906162 0.03289183
[2,] 0.02625538 0.02996795 0.03163265 0.03073171 0.03692308
[3,] 0.04502762 0.04240677 0.04468625 0.04433801 0.04377199
[4,] 0.04784254 0.08797386 0.05991678 0.06325707 0.05767123

> Xibar<-rowSums(claims_matrix)/rowSums(policies_matrix)
> Xibar
# 0.04042581 0.03078187 0.04407590 0.06140187                [1]

> Pi <-rowSums(policies_matrix)
> Pi
#  7609  3044 14019  4280                                     [½]

> P <-sum(Pi)
> P
# 28952                                                      [1]

> Pstar <-sum(Pi*(1-Pi/P))/(N*n-1)
> Pstar
# 1011.12                                                    [2]

> m <-sum(claims_matrix)/P
> s<-mean(rowSums(policies_matrix*(X-Xibar)^2)/(n-1))

> m
# 0.04428019
```

So, E[m(theta)] = 0.04428019 [1]

```
> s
# 0.07732984
```

E[s^2(theta)] = 0.07732984 [1]

```
> v<-(sum(rowSums(policies_matrix*(X-m)^2))/(n*N-1)-s)/Pstar
> v
# 8.801906e-05
```

Var[m(theta)] = 8.801906e-05 [2]

(ii)
Now calculate the credibility factors:

```
> Zi<-Pi/(Pi+s/v)                                            [4]
> Zi
# 0.8964887 0.7760242 0.9410267 0.8296892
```

Credibility factors are: (0.8964887, 0.7760242, 0.9410267, 0.8296892) [1]

(iii)
Then calculate the credibility premiums per unit of risk volume:

```
> premiums<-Zi*Xibar +(1-Zi)*m                               [4]
```



```
> premiums
# 0.04082478 0.03380516 0.04408794 0.05848586
```
[1]

```
> year6sales<-c(1920, 575, 2820, 798)
```
[1]

Finally calculate the credibility premium by multiplying by the sales for the upcoming year:

```
> premiums*year6sales
#  78.38358  19.43797 124.32800  46.67172
```
[3]

Therefore, the expected payout for the insurers in the coming year is as follows:
Insurer A: £78.4m, Insurer B: £19.4m, Insurer C: £124.3m, Insurer D: £46.7m [1]

[Total 25]

Candidates overall answered well this question.

Candidates demonstrated this kind of calculations had been well practised. Some numerical slips were common, however candidates who showed intermediate steps were able to pick up most of the available partial marks.

[Paper Total 100]

END OF EXAMINERS' REPORT