Institute
and Faculty
of Actuaries

EXAMINERS' REPORT

CS1B – Actuarial Statistics

Core Principles

Paper B

September 2022

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.

Sarah Hutchinson
Chair of the Board of Examiners
December 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, 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 score highly.

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

The quality of the commentary given alongside the R output was not always strong or insufficient (e.g. in Question 2). Performance in questions with atypical style (e.g. Q4(iv)) was relatively weak, 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.

In some cases, the layout of the provided answers was not satisfactory, with input and output being separated and presented with large gaps, and with significant duplications.

C. Pass Mark

The Pass Mark for this exam was 55
1302 presented themselves and 539 passed.

Solutions for Subject CS1B – September 2022

Q1

(i)

```
set.seed(2022)

x_bar <- rep(0,5000)                                    [1]
for (i in 1:5000)                                       [1]
+ {x <- rpois(150,6)                                    [2]
+ x_bar[i] <- mean(x)}                                  [1]
mean(x_bar)                                             [½]
[1] 5.999752                                            [½]
> var(x_bar)                                            [½]
[1] 0.03958983                                          [½]
```

(ii)

From the Central Limit Theorem, the distribution of the sample means will be approximately:

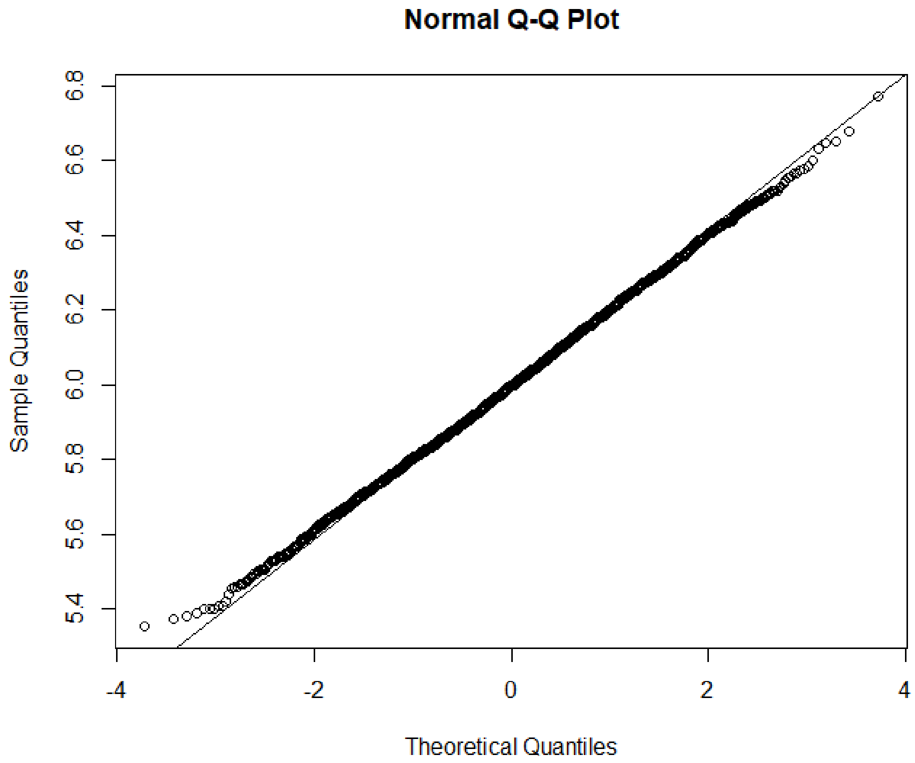
$N\left(6, \frac{6}{150}\right)$, i.e. $N(6, 0.04)$ [1]

(iii)

The mean and variance from part (i) are very close to the approximate Normal distribution in part (ii) [1]

(iv)

```
qqnorm(x_bar)                                           [2]
qqline(x_bar)
```



[1]

(v)
The QQ plot shows that the quantiles of x_bar and the Normal distribution are very similar [1]
Therefore, the Normal distribution is a good approximation for x_bar [1]
x_bar has a slightly lighter upper tail than the Normal distribution [1]
x_bar also has a slightly lighter lower tail than the Normal distribution [1]
The approximation will improve with larger sample sizes [1]

[Mars available 5, maximum 4]
[Total 16]

This question was answered well by most candidates.

There were some errors in specifying the variance in part (ii).

In part (iv), well prepared candidates were able to produce a QQ plot along with reasonable comments. A common error in part (iv) was using random values rather than the quantiles of the distribution to produce the plot.

In part (v), a relatively small proportion of candidates were able to identify and interpret differences in the tails.

Q2
(i)
X follows a Binomial (n, θ) distribution, where n is the number of clients surveyed and θ is the probability of getting a positive approval of the product from a client [1]

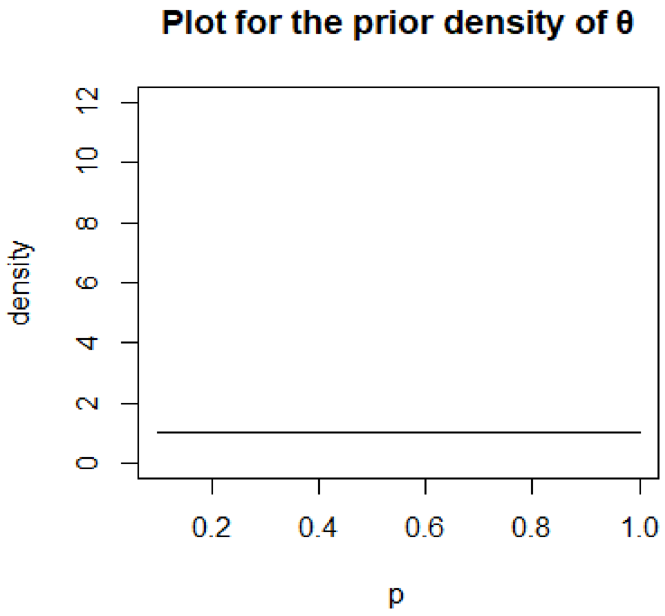
(ii)
The posterior distribution of θ is Beta with parameters
$\alpha + x = \alpha + 101$ and $n - x + \beta = 59 + \beta$. [2]

(iii)
The prior and posterior are of the same family. So, we have a conjugate prior [1]

(iv)
Beta(1, 1) is the same as a Uniform(0,1) distribution: $\alpha = 1, \beta = 1$ [1]

(v)(a)

```
n = 160
x = 101
alpha = 1
beta = 1
p = seq(0,1, by=0.01)                                              [½]
plot(p, dbeta(p, alpha, beta), ylab="density", type ="l",
ylim = c(0,12), main = "Plot for the prior density of θ")          [1]
```

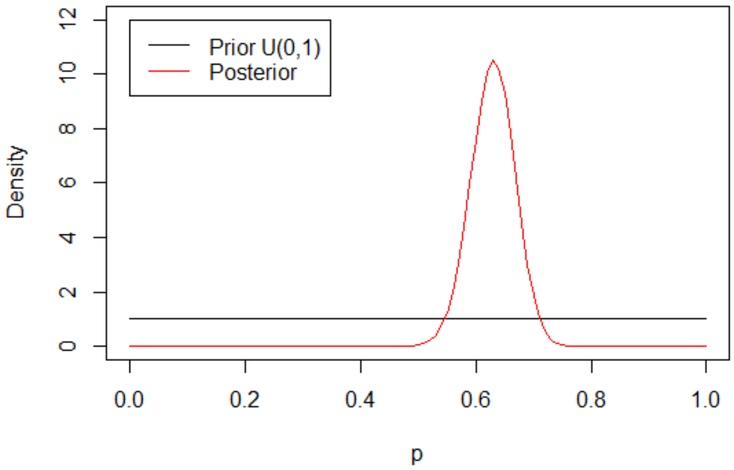


[½]

(b)

```
lines(p, dbeta(p, x+alpha, n-x+beta), type ="l", col="red") [1]

legend(0.,12, c("Prior U(0,1)",
"Posterior"),lty=c(1,1),col=c("black", "red"))                [½]
```



[½]

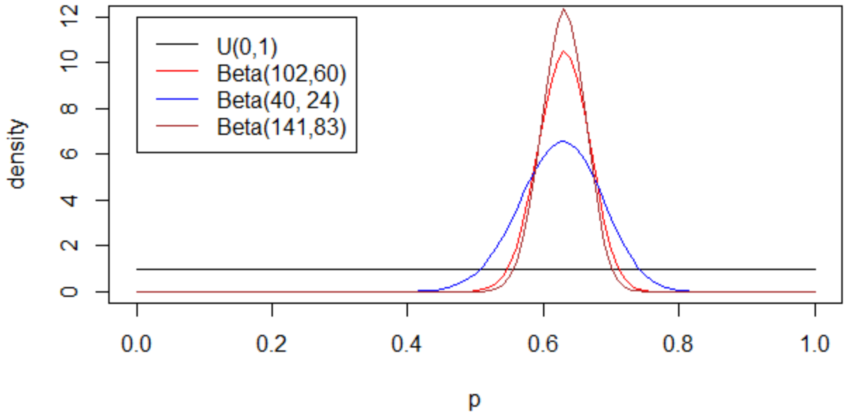
(vi)

```
alpha1 = 40
beta1 = 24

p = seq(0,1, by=0.01)
lines(p, dbeta(p, alpha1, beta1), col="blue")                    [1]

lines(p, dbeta(p, x+alpha1, n-x+beta1), type ="l",
col="brown")                                                     [1]

legend(0.,12, c("U(0,1)", "Beta(102,60)", "Beta(40, 24)",
"Beta(141,83)"),lty=c(1,1),col=c("black", "red", "blue",
"brown"))
```



[1]

(vii)
The Beta(40, 24) prior contains more information than the uniform (0, 1) prior [1]
and this is reflected in the posterior densities obtained as the posterior in (vi) is narrower than the one in (v) [1]

(viii)(a)
Using the Uniform(0, 1) prior:

```
1 - pbeta(3/5, x+alpha, n-x+beta)
```
[1½]

0.7840908 [½]

Using the Beta(40, 24) prior:

```
1 - pbeta(3/5, x+alpha1, n-x+beta1)
```
[1½]

0.8199189 [½]

(b)
The Beta(40, 24) prior provides a slightly higher posterior probability than the proportion of clients positively perceiving the new product [1]

As the difference between the two priors is considerable and the difference between the two posteriors is small, we can conclude that most information is coming from the data rather than the priors [1]

[Total 20]

This question was not well answered, with many candidates not attempting parts of it.

In parts requiring comments, credit was awarded for valid comments that are not shown here.

In parts where a legend is shown on the graph, credit was given when candidates provided a relevant description or comment in the answer instead of the legend.

In part (iv), a common error was to give (0,1) as the parameters of the distribution.

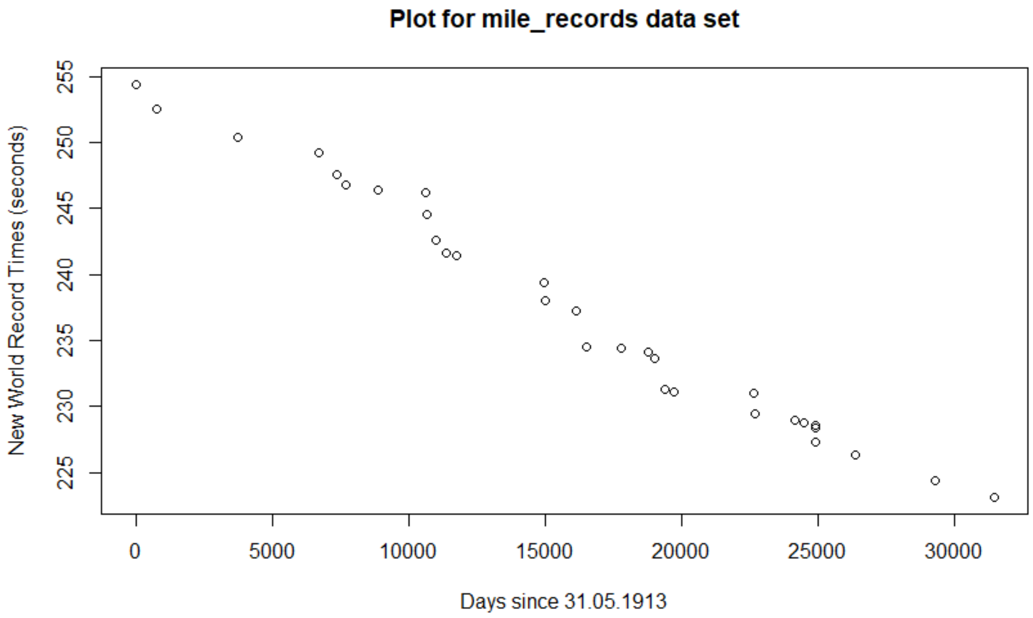
In part (viii)(a), a common error was using the prior instead of the posterior distribution.

Q3

```
# load data
load("mile_records.Rdata")
```

(i)

```
plot(record.date, record.time, ylab="New World Record Times
(seconds)", xlab="Days since 31.05.1913",main="Plot for
mile_records data set")
```
[1]



[1]

(ii)

```
cor(record.date, record.time)
```
[1]

-0.9885164 [1]

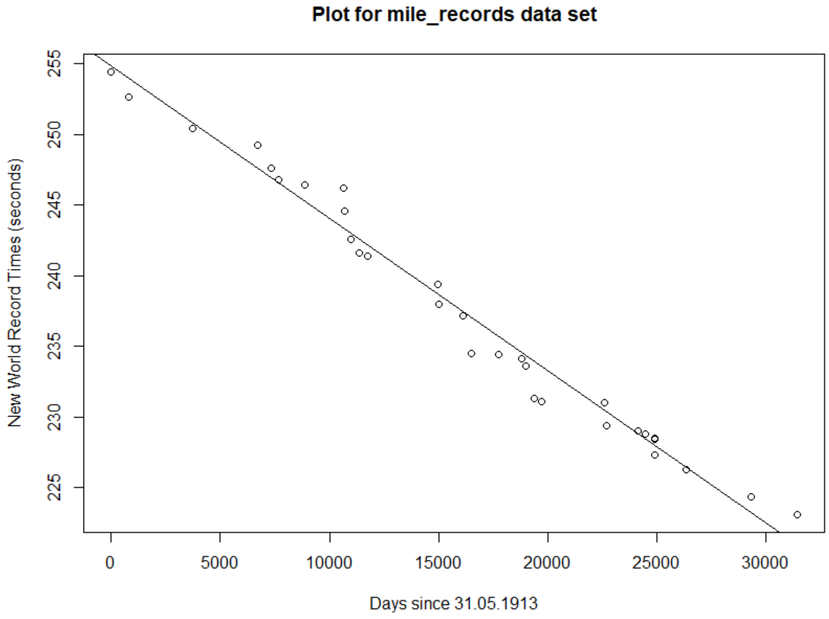
(iii)

```
model = lm(record.time ~ record.date)
coefficients(model)
```
[1]
[1]

The intercept of the regression line is 254.8 [½]
and the slope is -0.001076 [½]

(iv)

```
abline(model)
```
[1]



[1]

(v)

H0: beta = 0, versus H1: beta ≠ 0 [1]

summary(model) [1]

```
Call:
lm(formula = record.time ~ record.date)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6103 -0.7494 -0.1780  0.6999  2.8474 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.548e+02  5.409e-01  471.08   <2e-16 ***
record.date -1.076e-03  3.004e-05  -35.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.379 on 30 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9764 
F-statistic:  1284 on 1 and 30 DF,  p-value: < 2.2e-16
```

P-value is very small (i.e. p-value: < 2.2e-16) [1]

So conclude that negative slope is significant [1]

(vi)

The plot clearly shows a negative relationship between days since May 1913 and the current one-mile world record. [1]

The correlation coefficient is very close to -1, and the slope is negative and significant, confirming the relationship [1]

However, the response variable cannot increase with time (it is a world record).
So we cannot observe a positive relationship, meaning that the above results are to be expected [2]

(vii)

```
x = record.date[32]+365*100                                  [1]
predict(model, data.frame(record.date=x))                    [2]
# Or
coefficients(model)[1] + coefficients(model)[2]*x
```

181.66 seconds [1]

(viii)

```
twoMinuteDate = (120 -
coefficients(model)[1])/coefficients(model)[2]               [1]

twoMinuteDate = twoMinuteDate - record.date[32]              [1]
twoMinuteDate/365                                            [1]
```

This gives: 256.97 years. [1]

(ix)
The model is suitable for modelling the past observations, and we would also consider this a good model for predicting records in the near future as the model is significant, R^2 is close to one and the correlation is clearly negative. We have no reason to assume that the linear relationship will break down in the near future [1]

However, in the long run the model predicts unrealistically low times for the one mile run, eventually predicting even negative times [1]

[Total 27]

The quality of answers given in this question was mixed.

Part (i) was well answered generally, but marks were lost for not providing appropriate annotation on the plot.

In part (v) a number of candidates failed to state the hypotheses and explicitly state a full conclusion.

In parts (vii), (viii), a relatively small proportion of candidates provided fully correct answers. In parts (vi) and (ix) credit was awarded for other reasonable comments.

Q4

```
# load data
load("employee.RData")
```

(i)
Categorical variables: gender, job.type and job.location [1]

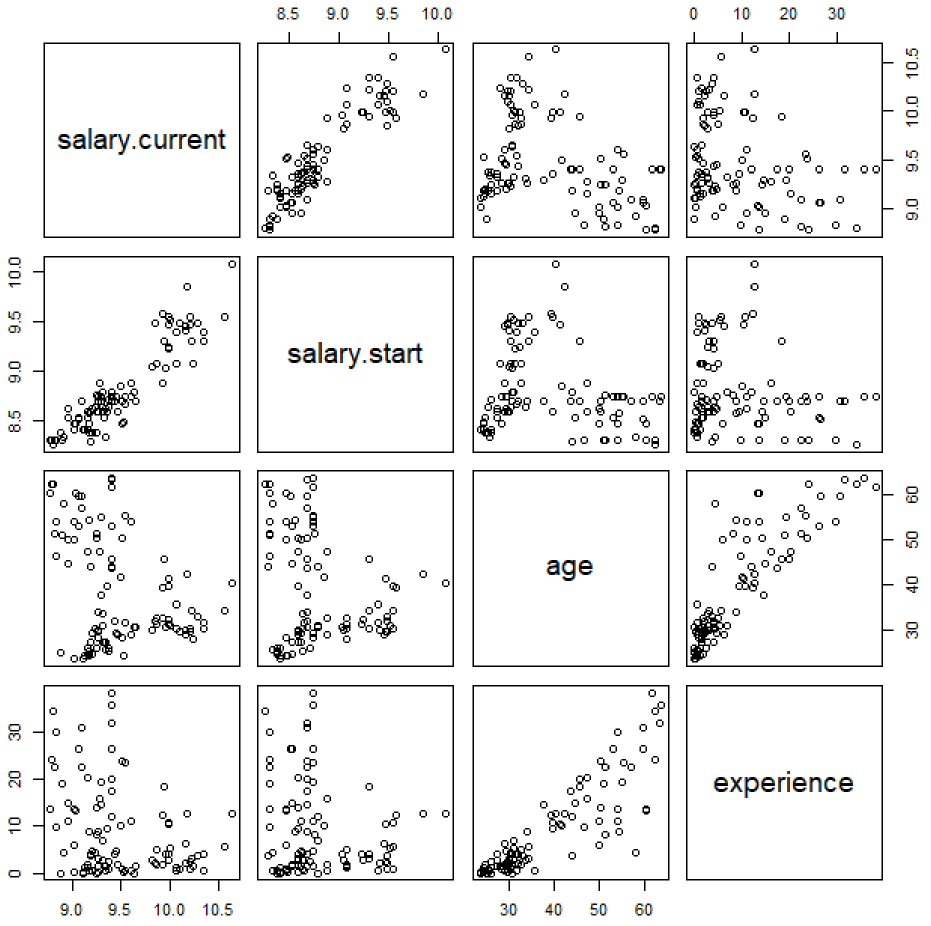
Numerical variables: salary.current, salary.start, age and experience [1]

(ii)
Scattergraphs between each pair of continous variables:

```
data_continuous <- data.frame(salary.current, salary.start,
age, experience)                                         [1]
pairs(data_continuous)                                   [1]
```

or

```
plot(data_continuous)
```



[1]

(iii)
The first column shows the relationship between salary.current and the other continuous variables. We can see that:

there appears to be a positive linear relationship between the current and the starting salary [1]
there appears to be a weak negative linear relationship between current salary and the employee age and the employee experience [1]
age has an increasing relationship with salary at the very start but then it becomes unclear [1]

[Marks available 3, maximum 2]

(iv)(a)
The lower quartile, median, upper quartile of salary.current are given as:

```
quantile(salary.current, c(0.25,0.5,0.75))                    [½]
25%    50%    75%
9.1750 9.3650 9.8625                                          [1½]
```

and the mean:

```
mean(salary.current)                                          [1]
# 9.4778
```

```
## Alternatively, candidates can use the function summary
summary(salary.current)
Min.   1st Qu.  Median   Mean  3rd Qu.   Max.
8.780   9.175   9.365   9.478   9.863   10.630
```

(b)
We use the following notation:
p_1: proportion of male employees with current salary values below 9.86
p_2: proportion of female employees with values below 9.86

$H_0: p_1 = p_2$
$H_1: p_1 \neq p_2$ [1]

Proportion calculations:

Current salary for male employees:

```
cmale<- salary.current[gender==0]                             [½]
```

Those less than the upper quartile:

```
qmale<- cmale[cmale <9.86]                                    [1]
```

Number of male employees with current salary lower than the upper quartile:

```
x1<- length(qmale) ; x1                                       [1]
# 26
```

Number of male employees:

```
n1<- length(cmale); n1                                        [1]
# 51
```

Current salary for female employees:

```
cfemale<- salary.current[gender==1]                           [½]
```

Those less than the upper quartile:

```
qfemale<- cfemale[cfemale <9.86]                              [1]
```

Number of female employees with current salary lower than the upper quartile:

```
x2<- length(qfemale) ; x2
# 48
```

[1]

Number of female employees:

```
n2<- length(cfemale); n2
# 49
```

[1]

Perform test:

```
prop.test(c(x1,x2), c(n1 , n2 ), alternative = "two.sided" )
```

[2]

```
2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 26.276, df = 1, p-value = 2.959e-07
alternative hypothesis: two.sided
95 percent confidence interval:
-0.6325920 -0.3069838
sample estimates:
prop 1    prop 2
0.5098039 0.9795918
```

The p-value is less than 0.05, therefore reject H_0. The proportion of males with current salary below 9.86 is significantly different from the proportion of female employees with current salary below 9.86 [1]

(v)

The median, mean and variances of salary.current for each of the job type in job.type can be given as follows:

```
Median<- rep(0,5)
Mean<- rep(0,5)
Var<- rep(0,5
for(i in 1:5){
Median[i]<- median(salary.current[job.type==i] )
Mean[i]<- mean(salary.current[job.type ==i] )
Var[i]<- var(salary.current[job.type ==i] )
}
```

[3]

```
## Median
# 9.180  9.285  9.410 10.160 10.030
```

[1]

```
## Mean
# 9.168571  9.280333  9.440000 10.101333 10.119167
```

[1]

```
## Var
# 0.07471261 0.02906540 0.00320000 0.03128381 0.05988106
```

[1]

(vi) (a)

H_0: Mean starting salaries and the current salaries are equal
H_1: Mean starting salaries and the current salaries are not equal. [½]

```
t.test(salary.start,salary.current, paired = TRUE)          [2]
```

```
#Results                                                     [1]
Paired t-test

data:  salary.start and salary.current
t = -38.506, df = 99, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7190359 -0.6485641
sample estimates:
mean of the differences
-0.6838
```

The p-value is less than 0.05; therefore we reject H_0 and conclude that the current salary is significantly different from the starting salary [1½]

(b)
We define two vectors of current salaries for big- and small-city employees:

```
salary.current_bcity<- salary.current[job.location==0]
salary.current_scity <- salary.current[job.location==1]      [1]
```

H_0: Mean starting salaries for big and small city employees are equal
H_1: Mean starting salaries for big city employee > small city employee [½]

```
t.test(salary.current_bcity, salary.current_scity,alternative=
"greater")                                                   [1½]
```

```
# Results                                                    [½]
 Welch Two Sample t-test
 data:  salary.current_bcity and salary.current_scity
 t = 5.2366, df = 90.239, p-value = 5.303e-07
 alternative hypothesis: true difference in means is greater than 0
 95 percent confidence interval:
 0.2481034       Inf

 sample estimates:
 mean of x mean of y
 9.572297  9.208846
```

We obtain a p-value<0.05, and therefore reject H_0 at 5%. We conclude that the current salary for big-city employees is greater than that for small-city employees [1½]

[Total 37]

This question was answered well in general.

The computations in parts (iv) and (v) can be performed in a number of alternative ways in R, and credit was given as appropriate when alternative R code was presented.

Many candidates failed to provide a fully correct answer in part (iv)(b), where a number of errors were found in the calculations.

In parts (vi)(a), (vi)(b) a number of candidates failed to state the hypotheses of the tests and complete conclusions.

[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

www.actuaries.org.uk