INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2019

Subject CS2A – Risk Modelling and Survival Analysis Core Principles

Introduction

The Examiners' Report is written by the Principal 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
December 2019

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

1. The aim of the Risk Modelling and Survival Analysis subject is to provide a grounding in mathematical and statistical modelling techniques that are of particular relevance to actuarial work, including stochastic processes and survival models and their application.

2. Some of the questions in this paper admit alternative solutions from those presented in this report, or different ways in which the provided answer can be determined. All mathematically correct and valid alternative solutions received credit as appropriate.

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

4. In higher order skills questions, where comments were required, well-reasoned comments that differed from those provided in the solutions also received credit as appropriate.

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

1. Performance was generally satisfactory, with most candidates demonstrating a reasonable understanding and application of core topics in mathematical and statistical modelling techniques.

2. Topics that were not particularly well answered in this paper include Mortality Projection (e.g. Q2) and Machine Learning (e.g. Q5), despite these questions being mostly knowledge or straightforward application questons. Candidates are reminded that it is very important to be familiar with all aspects of the syllabus.

3. It is important that rigourous mathematical notation and derivations are provided by candidates where appropriate. In certain cases, e.g. Q7(ii), Q8(iv) and Q9(iii), the absence of a demonstration that the derived likelihood estimates were maximums led to loss of marks. Additionally, e.g. in Q4(i) and Q9(ii), candidates should ensure that all notation used in quoted formulae are defined clearly.

4. Higher order skills questions were generally answered poorly. Candidates should recognise that these are generally the questions which differentiate those students with a good grasp and understanding of the subject.

5. The comments that follow the questions in the marking schedule below, concentrate on areas where candidates could have improved their performance. Candidates approaching the subject for the first time are advised to concentrate their revision in these areas.

C. Pass Mark

The Combined Pass Mark for the CS2 exam was 58.

Solutions for Subject CS2A – September 2019

1

Mean:
E[Total] = E[A] + E[B] = 2.5 x 1,500 + 3.0 x 1,250 = 7,500 [½]

Variance:
V[Total] = V[A] + V[B] = 2.5 x 2 x $(1,500)^2$ + 3.0 x 2 x $(1,250)^2$ [1]
= 11,250,000 + 9,375,000 = 20,625,000 [½]

[Total 2]

This question was well answered by most candidates. A common mistake was for candidates to forget to sum together the values for policy types A and B. Candidates are reminded of the need to read the question carefully.

2(i)

- The use of an exponential curve is attractive as there is evidence that age-specific mortality has declined exponentially in some past periods. [½]
- The approach is simple to understand and easy to implement. [1]
- However, fitting separate curves at each age risks the projected future mortality rates in any given year not progressing smoothly with age (and even decreasing with age in age ranges where this is implausible) i.e. under-graduated rates [1]
- This problem could be overcome by graduating the projected rates. [½]
- Or by using an alternative method/model in the first place. [½]
- The approach assumes that developments in medical technology, lifestyle, etc. in the future will progress steadily as they have in the past 10 years. [1]
- The appropriateness of this projection method may depend on whether the past history displays an exponential change over time [½]
- Using cohorts to project mortality instead of time period may lead to improvements in the reliability of the projection [½]
- It could be argued that 10 years of historic Life tables may not be sufficient to provide a reliable projection of future mortality [½]

[6, Max 3]

(ii)

EITHER: Lee-Carter model
OR: age, cohort model
OR: age, period, cohort model
OR: penalised splines
OR: decomposition of mortality by cause of death / Explanatory method. [1]

OR:
Adjust rates using projected rates from a similar country [½]

[1½, Max 1]

[Total 4]

Part (i) was not very well answered. Many candidates mistakenly commented on the concept of using a single exponential curve to project mortality across ages rather than projecting across time using a separate curve for each age.

Answers to part (ii) were generally mixed despite this being a straightforward knowledge question.

Candidates are advised to be familiar with all aspects of the syllabus.

3 (i)

Simple random walk. [1]
Discrete state space OR {0, ±1, ±2, ±3, …} [½]
Discrete time domain OR {0, 1, 2, 3, …} [½]

(ii)

EITHER:
At t = 1, the random walk will take the values
+1 with probability p
-1 with probability 1 – p [1]

At t = 2, the random walk will take the values
+2 with probability p^2
0 with probability $2p(1 - p)$ since there are two ways it can reach zero after two transitions
-2 with probability $(1 – p)^2$ [1]

At t = 3, the random walk will take the values
+3 with probability p^3
+1 with probability $3p^2(1 – p)$ since there are three ways it can reach +1 after three transitions
-1 with probability $3p(1 – p)^2$ since there are three ways it can reach -1 after three transitions
-3 with probability $(1 – p)^3$ [1]

OR:

$$\Pr[X_t = t - 2r] = \binom{t}{r} p^{t-r}(1-p)^r$$

, where r = 0, 1, 2, …, t [3]

(iii)

Normal (as N gets large) [1]

[Total 6]

Parts (i) and (ii) were well answered. Some candidates lost marks in part (i) for not including negative integers in the state space when presenting the answer in integer form. Additionally, some candidates lost marks for naming the process a general random walk. A common mistake in part (ii) was for candidates to suggest that there were only two ways to reach +1 and –1 at time 3.

Part (iii) was less well answered. Despite being asked for in the question, the parameters (mean = N(2p – 1), variance = 4Np(1 – p)) were not required for full marks as they were deemed too much to demand for 1 mark. Note that the Binomial distribution only received partial credit as this distribution is restricted to positive integers only.

4(i)

EITHER:
The future development of the process can be predicted from its present state alone, without reference to its past history. [1]

OR:
$P[X_t \in A \mid X_{s_1} = x_1 , X_{s_2} = x_2 , ..., X_{s_n} = x_n , X_s = x] = P[X_t \in A \mid X_s = x]$ [½]

for all times $s_1 < s_2 < ... < s_n < s < t$, all states $x_1, x_2, ..., x_n, x$ in S and all subsets A of S. [½]

OR:
$P[X_t \leq x \mid F_s] = P[X_t \leq x \mid X_s]$ [½]

for all $t \geq s \geq 0$ and where $(F_t)_{t\geq 0}$ is the filtration associated with X_t , $t \geq 0$. [½]

(ii)(a)

I	irreducible	[½]
II	reducible	[½]
III	irreducible	[½]

(b)

I	periodic with period 2	[½]
II	aperiodic	[½]
III	aperiodic	[½]

(iii)

I	{0.5, 0.5}	[1]
II	{1, 0, 0, 0}	[1]
III	Equations are:	

$$\pi_0 = 0.25\pi_0 + 0.5\pi_1 \quad (1)$$

$$\pi_1 = 0.75\pi_0 + 0.5\pi_2 \quad (2)$$

$$\pi_2 = 0.5\pi_1 + 0.75\pi_3 \quad (3)$$

$$\pi_3 = 0.5\pi_2 + 0.25\pi_3 \quad (4)$$

[1]

From (1) we have:

$$\pi_1 = 1.5\pi_0 \quad .$$ [½]

From (2) we then have:

$$\pi_2 = 1.5\pi_0$$

and from (3) we have:

$$\pi_3 = \pi_0$$ [½]

Hence

$$\pi_0 + 1.5\pi_0 + 1.5\pi_0 + \pi_0 = 1$$ [½]

and the stationary probability distribution is
{0.2, 0.3, 0.3, 0.2}. [½]

[Total 9]

.

All parts of this question were very well answered, in particular, parts (i) and (ii).

Candidates are reminded of the need to rigourously state all supporting notation when providing formulae in part (i).

A common mistake in part (iii) was for candidates to assume that a stationary probability distribution for I and/or II did not exist.

5 (i)

The train-validation-test approach uses three data sets as follows: [½]

- A training data set which is the sample of data used to fit the model; [½]
  that is, to train the algorithm to choose the most appropriate hypothesis; [½]

- A validation data set which is the sample of data used to provide an unbiased evaluation of model fit on the training dataset while adjusting the hyper-parameters [½]
  these hyper-parameters are often specified in advance and then adjusted/optimised according to the performance of the model on the validation data; [½]

- A test data set which is the sample of data used to provide an unbiased evaluation of the final model fit on the training data set. [½]
  Under machine learning the results of the modelling exercise are applied to data which was not used to develop the algorithm, [½]
  so the test data should be representative of the data on which the algorithm is to be used. [½]
  A typical split of data is 60% for training, 20% for validation and 20% for testing [½]
  the principle being that enough data must be selected for the validation and testing sets, with the remainder used for the training set. [½]

[5, Max 4]

(ii)

For machine learning to be useful in addressing the problem:

- A pattern should exist given that the model fitting will involve identification of patterns. [1]
  In this case, it is likely that patterns of traffic flow will exist, [½]
  for example, areas where traffic is most dense as well as times of day such as rush hour [½]
- The pattern cannot be practically pinned down mathematically [½]
  as it would be difficult to use a classical mathematical model for traffic. [½]
- Data exist which are relevant to the pattern. [½]
  In this case, data are available on the number of vehicles passing specific locations, which is relevant to modelling traffic volumes. [½]
- Therefore, we can conclude that machine learning would be appropriate for this exercise [½]

[4½, Max 3]

(iii)

- The advantage of having more parameters is that it can improve the accuracy of the model and predictions, [½]
  because a model with, more parameters will fit the data more closely than one with fewer parameters [½]
  For example, the flow of traffic is likely to be affected by a large number of factors, such as time of day, weather, weekday vs weekend [½]
- However, if too many parameters are used there is a risk of over-fitting [½]
  where the estimates from the model will reflect idiosyncratic characteristics of the "training" data set rather than characteristics which apply to the whole data set. [½]
  This may lead to the analyst identifying patterns which do not exist. [½]
  For example, the analyst in this case may have used a training dataset which includes anomalies in traffic flow, [½]
  perhaps due to a vehicle breaking down near one of the sensors which distorted the collection of data of other vehicles had to divert around it [½]
  If too many parameters are used the model can become complex and computationally expensive to run [½]
  Using too many parameters may lead to model stability issues [½]

[5, Max 3]

[Total 10]

Performance in all parts of this question was less satisfactory, in particular part (ii).

In part (i), candidates' answers were often vague and lacking detail despite this being a straightforward knowledge question.

In many cases, answers to part (ii) did not revolve around the three principles set out in Core Reading that define whether Machine Learning is useful for tackling a particular problem.

Appropriate alternative examples in parts (ii) and (iii) received credit.

6 (i)

- The null hypothesis is that national mortality and that of the company's policy holders are the same, [½]
- We perform a chi-squared test. [½]
- If d^e_x are the expected number of deaths at age x assuming the national mortality rate, then the test statistic is

$$\sum_x \left(z_x^{\,2} \right)$$

where

$$z_x = \frac{d_x - d_x^e}{\sqrt{d_x^e}} \quad .$$

The calculations are shown below.

Age x	d_x	d_x^e	$\frac{(d_x - d_x^e)}{\sqrt{d_x^e}}$	$\frac{(d_x - d_x^e)^2}{d_x^e}$	
60	10	9.9225	0.02460	0.00061	
61	11	10.9742	0.00779	0.00006	
62	12	11.8144	0.05400	0.00292	
63	15	13.0900	0.52791	0.27869	
64	12	14.2868	-0.60501	0.36603	
65	5	9.3920	-1.43312	2.05384	
66	5	10.1790	-1.62328	2.63504	
67	6	10.9934	-1.50602	2.26809	
68	8	11.4912	-1.02989	1.06068	
69	8	12.2675	-1.21842	1.48454	[2]
$\sum_x \left(z_x^{\,2} \right) =$				10.15049	[½]

- We compare the test statistic with the critical value of the chi-squared distribution with 10 degrees of freedom, [½]
- as we have 10 ages. [½]
- The critical value at the 5% level is 18.31. [½]
- Since 10.15 < 18.31 [½]
- we do not reject the null hypothesis. [½]

(ii)

- Overall, the mortality of the policyholders reflects the national rates, so the company's pricing policy might not be considered unreasonable, [½]

- The chi-squared test is based on squared deviations and therefore tells us nothing about the direction of any bias… [½]
- …it is clear that, at ages 64 and older there are fewer deaths among the policyholders than expected. [1]
- It may be that competitors have spotted this and reduced their premiums to reflect this. [½]

[2½, Max 2]

(iii)

EITHER:

- The null hypothesis is the same as that stated in part (i). [½]
- To test whether the age pattern of differences is statistically significant, we could use a Grouping of Signs test. [½]
- We have 10 ages, 4 positive signs and 1 positive run. [1]
- According to the table on p. 189 of the Golden Book [½]
- The probability of getting only 1 positive run with 4 positive signs and 10 ages is < 0.05 [½]
- We therefore reject the null hypothesis that the mortality of the policyholders reflects the national mortality rate. [½]

[3½, Max 3]

OR:

Cumulative Deviations Test.

- This should be carried out over sub-sections of the data chosen without reference to the values of the z_xs. [½]
- So consider the data in two halves: ages 60-64 years and ages 65-69 years. [½]
- The null hypothesis is the same as that stated in part (i). [½]

The test statistic is $\dfrac{\sum_x (d_x - d_x^e)}{\sqrt{\sum_x d_x^e}}$

- For ages 60-64 years, this is $\frac{-0.0879}{\sqrt{60.09}} = -0.01134$ [½]
- For ages 65-69 years, this is $\frac{-22.3231}{\sqrt{54.32}} = -3.02874$ [½]
- Since -1.96 < test statistic < +1.96 does not hold for both age ranges, [½]
  we reject the null hypothesis that the mortality of the policyholders reflects the national mortality rate. [½]

[3½, Max 3]

OR:

Serial Correlations Test.

- The null hypothesis is the same as that stated in part (i). [½]
- z_bar_1 = -0.62034
- z_bar_2 = -0.75845 [½]

Age	z_{x}-z_bar_1	z_{x+1}-z_bar_2	(z_{x}-z_bar_1)(z_{x+1}-z_bar_2)
60	0.64494	0.76624	0.49418
61	0.62812	0.81245	0.51032
62	0.67433	1.28636	0.86744
63	1.14825	0.15344	0.17619
64	0.01533	-0.67467	-0.01034
65	-0.81279	-0.86483	0.70292
66	-1.00294	-0.74757	0.74977
67	-0.88568	-0.27144	0.24041
68	-0.40956	-0.45997	0.18838
69	-0.59808	0.75845	-0.45361

Sum = 3.46565

Age	(z_{x}-z_bar_1)^2	(z_{x+1}-z_bar_2)^2
60	0.41595	0.58712
61	0.39454	0.66007
62	0.45473	1.65473
63	1.31848	0.02354
64	0.00023	0.45518
65	0.66062	0.74793
66	1.00590	0.55886
67	0.78443	0.07368
68	0.16774	0.21157
69	0.35770	0.57524

Sum = 5.56031 5.54794

- $r_1 = 3.46565 / \sqrt{(5.56031 * 5.54794)} = 0.62398$

- $(\sqrt{10}) * r_1 = 1.97319$ [1½]

- We compare the test statistic with the critical value of the Normal (0,1) distribution. The critical value at the 5% level is 1.96. [½]
- Since 1.97319 > 1.96 we reject the null hypothesis. [½]

[Max 3]

OR:

Signs test.

- The null hypothesis is the same as that stated in part (i). [½]
- Under the null hypothesis, random variable representing the number of positive deviations is Binomial (10, 0.5). [½]
- There are 10 deviations in total of which 4 are positive. [½]

EITHER:

- The probability of getting four or fewer positive deviations is 0.377 which exceeds 0.025 (two tailed test). [½]
- So there is insufficient evidence to reject the null hypothesis. [½]

OR:

- From the Golden Book, k* = 2, therefore, the test is satisfied if 2 <= P <= 8 where P = number of positive signs [½]
- Since P = 4 there is insignificant evidence to reject the null hypothesis [½]

[Max 2]

OR:

The Individual Standardised Deviations test.

- The Individual Standardised Deviations tests looks for individual large deviations at particular ages. [½]
- The null hypothesis is the same as that stated in part (i). [½]
- Under the null hypothesis, we would expect the standardised individual deviations to be distributed according to Normal (0,1). [½]
- This implies that we should expect only 1 in 20 z to be larger than 1.96 in absolute value. [1]
- Looking at the table of z's, we see that none is larger in absolute value that 1.96. [½]
- This does not provide enough evidence to reject the null hypothesis [½]

[Max 3]

(iv)

- It may be that 65 is retirement age for a substantial proportion of the community leaders. [1]
- Some of the business may be in term assurance for "death-in-service" benefit which expires at age 65 years. [1]
- The leaders who stay on beyond age 65 are in better health than those who retire at that age. [1]
- The lack of data at older ages could be causing more volatility in the results resulting in the life company's rates diverging from the national rate. [½]
- The chi-squared test is based on squared deviations telling us nothing about the direction of any bias and so the null hypothesis was not rejected. However, the test performed in part (iii) identifies bias/clumping at the older ages and so the null hypothesis was rejected [1]

[Max 2]

[Total 13]

Part (i) was well answered although a common mistake was for candidates to fail to provide justification for the number of degrees of freedom to use in the chi-squared distribution. Excessive rounding was also penalised in part (i).

Parts (ii) and (iv) were less well answered. Although appropriate alternative comments received credit here, few candidates tied the results of the tests in parts (i) and (iii) back to the information provided in the question.

Part (iii) was generally well answered. A number of different tests received credit; however, the Signs test only received partial credit as it does not test the key feature of the data.

7 (i)(a)

Y = X if X <= M,
Y = M if X > M [1]

OR:
Y = min(X, M) [1]

(b) $f_Y(y) = fx(y)$ for Y < M, [1]
$P(X > M)$ for Y = M [1]

(ii)

$L = \left(\prod_{i=1}^{10} 0.6cx_i^{-0.4}e^{-cx_i^{0.6}}\right)\left(P(X > 1000)\right)^5$ [1]

$= 0.6^{10}c^{10}\prod_{i=1}^{10}(x_i^{-0.4})e^{-c\sum x_i^{0.6}}(e^{-c*1000^{0.6}})^5$ [1]

$\propto c^{10}e^{-c\sum x_i^{0.6}}e^{-5c*1000^{0.6}}$ [1]

$ln\, L = 10\, ln\, c - c\sum x_i^{0.6} - 5c * 1000^{0.6} + const$ [1]

$\frac{d\, ln\, L}{dc} = \frac{10}{c} - \sum x_i^{0.6} - 5 * 1000^{0.6}$ [1]

$\Rightarrow c = \frac{10}{\sum x_i^{0.6}+5*1000^{0.6}} = \frac{10}{389.474+315.479} = 0.01419$ [1]

$\frac{d^2\, ln\, L}{dc^2} = \frac{-10}{c^2} < 0$ [1]

(iii)

Median is 700. [1]
$F(m) = 1 - e^{-cm^{0.6}} = 0.5$ [1]

$-c * 700^{0.6} = ln\, 0.5$ [1]

$\Rightarrow c = 0.01361$ [1]

[Total 14]

Answers to part (i) were generally less satisfactory. Part (a) was well answered but part (b) was not. Many candidates' answers suggested that Y could exceed M.

Part (ii) was generally well answered, although many candidates failed to deal adequately with the five claims above the rentention limit. Some candidates also lost marks as they did not demonstrate that their likelihood estimate was the maximum.

Part (iii) was well answered but some candidates did not take into account the five claims above retention when calculating the median claim amount.

8 (i) $y_t - \alpha y_{t-1} = \varepsilon_t \sim N(0, \sigma^2)$

$\Rightarrow y_t | y_{t-1} \sim N(\alpha y_{t-1}, \sigma^2)$ [1]

(ii) $L(\alpha, \sigma^2) = p(y_1, \ldots, y_n) = p(y_1)p(y_2|y_1) \ldots p(y_n|y_{n-1}, \ldots, y_1)$

That is,

$L(\alpha, \sigma^2) = p(y_1) \prod_{t=2}^{n} \frac{1}{(\sqrt{2\pi}\sigma)} e^{\frac{-(y_t - \alpha y_{t-1})^2}{2\sigma^2}}$ [1½]

$= p(y_1) \frac{1}{(\sqrt{2\pi}\sigma)^{n-1}} e^{\frac{-\sum_{t=2}^{n}(y_t - \alpha y_{t-1})^2}{2\sigma^2}}$ [½]

(iii) $\log L = -(n-1)\ln(\sqrt{2\pi}\sigma) - \frac{\sum_{t=2}^{n}(y_t - \alpha y_{t-1})^2}{2\sigma^2}$ [1]

(iv)

$\frac{\partial \ln L}{\partial \alpha} = \frac{\sum_{t=2}^{n} 2y_{t-1}(y_t - \alpha y_{t-1})}{2\sigma^2} = 0$ [1]

$\Rightarrow \alpha \sum_{t=2}^{n} 2y_{t-1}^2 = \sum_{t=2}^{n} 2(y_t y_{t-1}) \Rightarrow \alpha = \frac{\sum_{t=2}^{n}(y_t y_{t-1})}{\sum_{t=2}^{n} y_{t-1}^2}$ [1]

$\frac{\partial^2 \ln L}{\partial \alpha^2} = -\frac{\sum_{t=2}^{n} 2y_{t-1}^2}{2\sigma^2} < 0$ so maximum [½]

$\frac{\partial \ln L}{\partial \sigma} = \frac{-(n-1)}{\sigma} + \frac{2\sum_{t=2}^{n}(y_t - \alpha y_{t-1})^2}{2\sigma^3} = 0$ [1]

$$\Rightarrow (n-1)\sigma^2 = \sum_{t=2}^{n}(y_t - \alpha y_{t-1})^2 \Rightarrow \sigma^2 = \frac{1}{n-1}\sum_{t=2}^{n}(y_t - \alpha y_{t-1})^2$$ [1]

$\frac{\partial^2 lnL}{\partial \sigma^2} = \frac{n-1}{\sigma^2} - \frac{3\sum(y_t - \alpha y_{t-1})^2}{\sigma^4} = \frac{n-1}{\sigma^2} - \frac{3(n-1)}{\sigma^2} = \frac{-2(n-1)}{\sigma^2} < 0$ at the turning point. Hence a maximum. [½]

(v) $\alpha = \rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\sum_{t=2}^{n}(y_t - \overline{y})(y_{t-1} - \overline{y})}{\sum_{t=1}^{n}(y_t - \overline{y})^2}$ [1½]

$\sigma^2 = \gamma_0 - \alpha\gamma_1 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})^2 - \alpha\sum_{t=2}^{n}(y_t - \overline{y})(y_{t-1} - \overline{y})$ [1½]

(vi)

The main difference is that using Yule-Walker, we include the sample mean [1]
This may not have a big effect when average y is small. [1]

[Total 14]

In this question, there were a few typos in the formulae provided in the exam paper. The references to "a" should rather have been to "α". Also, in part (ii), the σ in the denominator should have been outside the square root. Full credit was awarded for solutions that were based on either the correct or incorrect formulae.

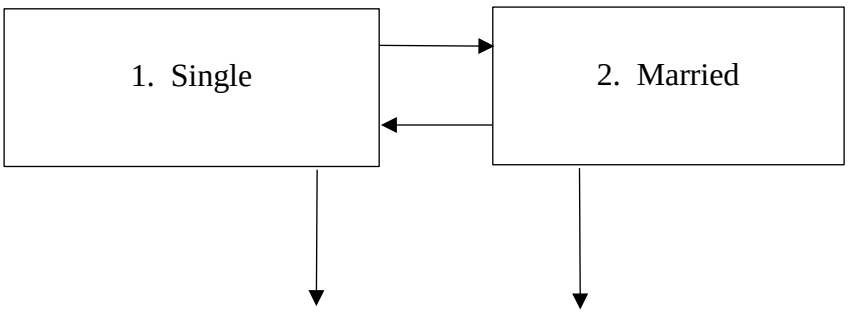
Answers to parts (i) and (ii) were less satisfactory. The typos, however, did not seem to be the cause of the poorer answers to part (ii).

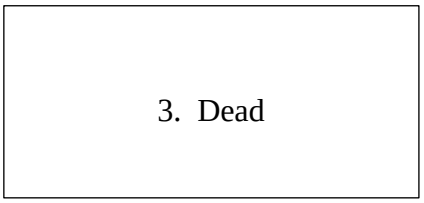
Part (iii) was generally well answered but part (iv) less so. Few candidates attempted to demonstrate that the likelihood estimates derived in part (iv) were maximums. Technically, as we are dealing with more than one variable here, we should be checking whether the Hessian matrix is negative definite. As this is beyond the scope of the syllabus, full credit was awarded for the solutions above or where candidates provided an appropriate explanation for why they hadn't demonstrated that they are maximums.

Answers to part (v) were satisfactory. Full credit was awarded to candidates who based the definitions of α and σ on appropriate derivations of the Yule-Walker equations.

Very few candidates scored marks in part (vi).

9 (i)





[2]

(ii) $$L \propto \exp\left\{\left(-\mu^{12}-\mu^{13}\right)\nu^{1}\right\}\exp\left\{\left(-\mu^{23}-\mu^{21}\right)\nu^{2}\right\}\left(\mu^{12}\right)^{d^{12}}\left(\mu^{21}\right)^{d^{21}}\left(\mu^{13}\right)^{d^{13}}\left(\mu^{23}\right)^{d^{23}}$$

[1½]

Where:

μ^{ij} is the transition intensity from state *i* to state *j* [½]

ν^{i} is the total observed waiting time in state *i* [½]

d^{ij} is the number of transitions from state *i* to state *j* [½]

(iii) Taking the logarithm of the likelihood we get:

$$\log_e(L) = -\mu^{13}\nu^{1} + d^{13}\log_e(\mu^{13}) + \text{terms not involving } \mu^{13}$$ [½]

Differentiate with respect to μ^{13}:

$$\frac{d\ln(L)}{d\mu^{13}} = -\nu^{1} + \frac{d^{13}}{\mu^{13}}$$ [½]

Setting this to zero we obtain: [½]

$$\hat{\mu}^{13} = \frac{d^{13}}{\nu^{1}}$$ [½]

To check it is a maximum differentiate again giving:

$$\frac{d^2\log_e(L)}{(d\mu^{13})^2} = -\frac{d^{13}}{(\mu^{13})^2}$$ which is always negative. [1]

(iv) (a) The maximum likelihood estimate of μ^{13} is 13/11,343 = 0.001146 [1]

(b) The estimated variance is $\left/ \frac{-1}{\frac{d^2\ln(L)}{(d\mu^{13})^2}} \right. = \left/ \frac{13}{11{,}343^2} \right.$ = 1.0104 x 10^{-7}. [1]

(v)

The maximum likelihood estimate of μ^{23} is 30/39,098 = 0.000767 [½]
A suitable test statistic is: [1]

$$\frac{\mu^{13} - \mu^{23}}{\sqrt{\frac{d^{13}}{(v^1)^2} + \frac{d^{23}}{(v^2)^2}}}$$

This is approximately distributed according to the standard Normal distribution under the null hypothesis of no difference in the two rates. [½]

The value of the test statistic in this case is:

$$\frac{0.001146 - 0.000767}{\sqrt{\frac{13}{(11{,}343)^2} + \frac{30}{(39{,}098)^2}}} = \frac{0.000379}{0.00034737} = 1.09043$$

[1]

Since 1.09043 < 1.96, [½]
we do not have enough evidence to reject the null hypothesis of no difference at the 5% level [½]

[Total 14]

Part (i) was very well answered. Some candidates lost marks for forgetting to include a transition rate from the Married to the Single state.

Parts (ii), (iii) and (iv) were well answered. Some candidates lost marks in part (ii) for not defining all terms used and in part (iii) as they did not demonstrate that their likelihood estimate was the maximum.

Part (v) was less well answered. Only partial credit was awarded to candidates who calculated confidence intervals for both maximum likelihood estimates and compared whether the intervals overlapped or not.

10 (i)

Compute the duration and censoring indicators for each case:

Garage	*Duration (months)*	*Censoring indicator (1 = wore out, 0 = censored)*
A	24	1
A	50	1
A	35	1
A	47	0
A	10	1
A	44	0
A	39	0
A	32	1

[1]

B	36	1	
B	62	1	
B	50	1	
B	12	1	
B	45	0	
B	32	1	
B	39	0	
B	35	0	[1]

Then we have for Garage A:

t_j	n_j	d_j	c_j	d_j/n_j	$1 - d_j/n_j$	
10	8	1	0	1/8	7/8	
24	7	1	0	1/7	6/7	
32	6	1	0	1/6	5/6	
35	5	1	3	1/5	4/5	
50	1	1	0	1	0	
[½]	[½]	[½]		[½]		[2]

For Garage B:

t_j	n_j	d_j	c_j	d_j/n_j	$1 - d_j/n_j$	
12	8	1	0	1/8	7/8	
32	7	1	1	1/7	6/7	
36	5	1	2	1/5	4/5	
50	2	1	0	1/2	1/2	
62	1	1	0	1	0	
[½]	[½]	[½]		[½]		[2]

The survival functions are therefore:

Garage A:

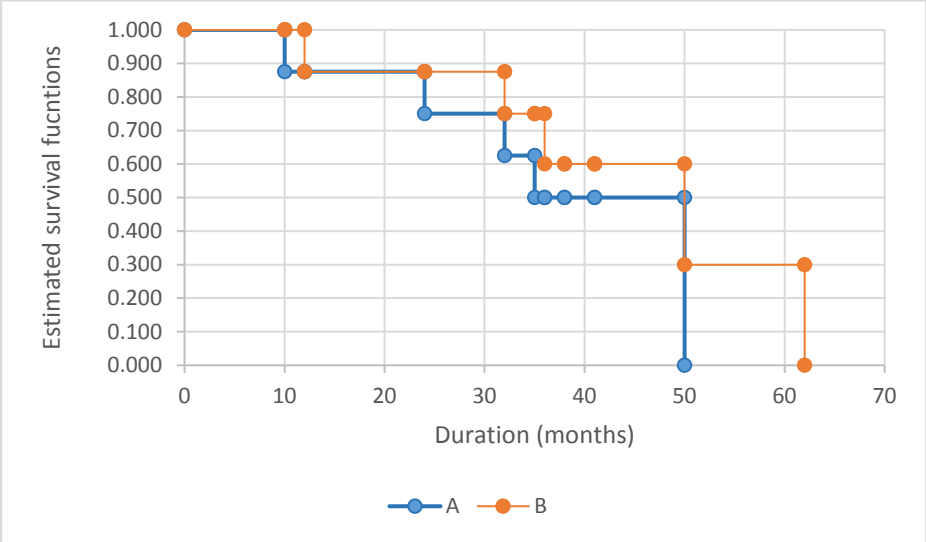
$0 \leq t < 10$	1.000	
$10 \leq t < 24$	0.875	
$24 \leq t < 32$	0.750	
$32 \leq t < 35$	0.625	
$35 \leq t < 50$	0.500	
$t \geq 50$	0.000	
[½]	[½]	[1]

Garage B:

$0 \leq t < 12$	1.000	
$12 \leq t < 32$	0.875	
$32 \leq t < 36$	0.750	
$36 \leq t < 50$	0.600	
$50 \leq t < 62$	0.300	
$t \geq 62$	0.000	
[½]	[½]	[1]

[Total 8]

(ii)



[3]

(iii)

- A proportional hazards (PH) model is potentially a convenient way of measuring the difference in the survival chances of two groups, such as is the case here. [½]
- The proportional hazards model assumes that the hazards for Garages A and B are in a constant proportion at all durations… [½]
- …however, this does not appear to be the case from the graph in part (ii). [½]
- A constant proportion at all durations implies that the survival functions should not cross (ideally, they should diverge gradually as duration increases). [½]
- The evidence for Garages A and B is that the survival functions do not cross, but they do not diverge progressively. [½]
- The PH model would allow controls for confounding variables. [½]
- So it may be appropriate to use a proportional hazards model [½]
- but further tests of the proportionality assumption are likely to be required. [½]

[4, Max 3]

[Total 14]

[Paper Total 100]

Parts (i) and (ii) were well answered. Some candidates lost marks in part (i) because they used an "equals" sign in the final time interval of the survival functions instead of a "greater than or equals to" sign.

Candidates were required to label both axes appropriately to be awarded full marks in part (ii).

Part (iii) was very poorly answered. This is a higher order skills question and so no credit was awarded for quoting Core Reading formulae for proportional hazard models. Appropriate alternative comments received credit here.

END OF EXAMINERS' REPORT