INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2016 (with mark allocations)

Subject CT4 – Models
Core Technical

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.

F Layton
Chair of the Board of Examiners
June 2016

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

1. The aim of the Models subject is to provide a grounding in stochastic processes and survival models and their application.

2. Subject CT4 comprises five main sections:

   (i) a study of the properties of models in general, and their uses for actuaries, including advantages and disadvantages (and a comparison of alternative models of the same processes);

   (ii) stochastic processes, especially Markov chains and Markov jump processes;

   (iii) models of a random variable measuring future lifetime;

   (iv) the calculation of exposed to risk and the application of the principle of correspondence;

   (v) the reasons why mortality (or other decremental) rates are graduated, and a range of statistical tests used both to compare a set of rates with a previous experience and to test the adherence of a graduated set of rates to the original data.

   Throughout the subject the emphasis is on estimation and the practical application of models. Theory is kept to the minimum required in order usefully to apply the models to real problems.

3. Different numerical answers may be obtained to those shown in these solutions depending on whether figures obtained from tables or from calculators are used in the calculations but candidates are not penalised for this. However, candidates may be penalised where excessive rounding has been used or where insufficient working is shown. Credit is given for valid solutions different from those shown below. Partial credit is also given to candidates submitting incomplete solutions with valid intermediate workings.

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

1. The performance of candidates in this diet was disappointing. The examination paper was considered to be of comparable difficulty to those of previous sessions, and a similar Pass Mark was therefore used.

2. One or two questions (or parts of questions) on this examination paper presented simple applications in an unfamiliar way. A substantial number of candidates made little or no attempt at these questions. This suggests that they had learned standard applications by doing examples without really understanding the concepts underlying them, and hence when faced with a test of these concepts which did not use one of the examples they had learned were unable to think through what was required. By contract, those candidates who did understand the concepts scored close to full marks on these sections of the examination paper.

C. Pass Mark

The Pass Mark for this exam was 57%.

Solutions

Q1 Date of birth
OR
exact age at a specified date; [1]

date of entry into observation; and [1]

date of exit from observation. [1]

[TOTAL 3]

Many candidates scored full marks on this question. A common error was to write "date of death" rather than "date of exit", ignoring exits for reasons other than death. The question was about exposed to risk rather than transition rates, so no credit was given for information needed to calculate the numerator. A minority of candidates framed their answers in aggregate terms: these candidates were awarded little or no credit.

Q2 **Advantages**

Systems with long time frames (such as the operation of a pension fund) can be studied in compressed time. [½]

Complex systems with stochastic elements (such as the operation of a life insurance company) can be studied by simulation modelling. [½]

Different future policies or possible actions can be compared to see which best suits the requirements or constraints of a user. [½]

Scenarios which could not be tested in real life can be examined. [½]

In a model of a complex system we can usually get control over the experimental conditions so that we can reduce the variance of the results output from the model without upsetting their mean values. [½]

Disadvantages

Model development requires a considerable investment of resources (time, money or expertise). [½]

In a stochastic model, for any given set of inputs each run gives only estimates of a model's outputs. So to study the outputs for any given set of inputs, several independent runs of the model are needed. [½]

Models can look impressive when run on a computer so that there is a danger that one gets lulled into a false sense of confidence. [½]

Models rely heavily on the data input. If the data quality is poor or lack credibility then the output from the model is likely to be flawed. [½]

There is a danger of using a model as a "black box" from which it is assumed that all results are valid without considering the appropriateness of using that model for the particular data input and the output expected. [½]

It is not possible to include all future events in a model. For example, an unforeseen change in legislation may invalidate the model. [½]

It may be difficult to interpret/communicate some of the outputs of the model. [½]

Models are better at comparing various inputs than optimising outputs. [½]

[MAX 4]

Most candidates were able to make a good attempt at this question. The instruction in the question was "list", so not all the detail under each point mentioned above was required for credit.

Q3 (i) A life alive at time t should be included in the exposure at age x at time t if and only if, were that life to die immediately, he or she would be counted in the death data d_x at age x. [Total 1]

(ii) $P_x(t)$ is the number of policies under observation aged x nearest birthday on 1 January in year t.

To correspond with the claims data, we wish to have policies classified by age last birthday. [1]

Define the number of policies aged x last birthday on 1 January in year t to be $P'_x(t)$. [½]

Assuming that birthdays are evenly distributed over time, [½]

$$P'_x(t) = \frac{1}{2}[P_x(t) + P_{x+1}(t)]\,. \qquad [½]$$

The central exposed to risk is given by

$$E^c_x = \int_0^1 P'_x(t)dt\,. \qquad [½]$$

Assuming that the population varies linearly between census dates, [½]

$$E_x^c \approx \frac{1}{2}[P_x'(t) + P_x'(t+1)]\,. \qquad [½]$$

Substituting for the $P_x'(t)$ in terms of $P_x(t)$ from the equation above gives

$$E_x^c \approx \frac{1}{2}\left[\frac{1}{2}[P_x(t) + P_{x+1}(t)] + \frac{1}{2}[P_x(t+1) + P_{x+1}(t+1)]\right]. \qquad [1]$$

[Total 5]
[TOTAL 6]

This was a straightforward exposed to risk question and was generally well answered. The most common reasons that candidates lost marks in part (ii) were the use of the incorrect age adjustment

$$P_x'(t) = \frac{1}{2}[P_x(t) + P_{x-1}(t)]\,,$$

and a failure to point out where in the argument the assumptions were required.

Q4 (i) The null hypothesis is that the company's schedule and that of the Continuous Mortality Investigation (CMI) are the same. [½]

THEN ALTERNATIVE 1

With 25 ages we can use the Normal approximation.

If p is the number of positive signs, a z-score can be computed as

$$z = \frac{p - 12.5}{\sqrt{6.25}}.$$

Hence

$$p = 12.5 + (\sqrt{6.25})z\,. \qquad [½]$$

We reject the null hypothesis if the number of positive signs is too few OR too many such that $|z| > 1.96$. [½]

This is true if $p > 17.4$ or if $p < 7.6$. [½]

Since we are told by the trainee that one more positive sign would lead to "failure" (i.e. rejection of the null hypothesis) [½]

we must have 17 positive signs. [½]

OR ALTERNATIVE 2

Define k^* as the smallest value of k such that

$$\sum_{j=0}^{k}\binom{25}{j}0.5^{25} \geq 0.025 .$$

We have

$$\sum_{j=0}^{7}\binom{25}{j}0.5^{25} = 0.000000 + 0.000001 + 0.000009 + 0.000069$$
$$+0.000377 + 0.001583 + 0.005278 + 0.014326 = 0.021643$$

and

$$\sum_{j=0}^{8}\binom{25}{j}0.5^{25} = \sum_{j=0}^{7}\binom{25}{j}0.5^{25} + 0.032233 = 0.053876 .$$ [½]

We reject the null hypothesis if the number of positive signs is too few or too many. [½]

$k^* = 8$, so we reject the null hypothesis if we have 7 or fewer positive signs, or if we have 18 or more positive signs. [½]

Since we are told by the trainee that one more positive sign would lead to "failure" (i.e. rejection of the null hypothesis), [½]
we must have 17 positive signs. [½]
[Max 3]

(ii) With 17 positive signs we have 8 negative signs. [½]

The null hypothesis is that the company's schedule and that of the CMI are the same [½]

THEN ALTERNATIVE 1

Using the table on p. 189 of the Golden Book, [½]
the critical value is 3. [½]

The critical value gives the highest number of runs which is incompatible with the null hypothesis at the 95% level. [½]

We reject the null hypothesis with 3 or fewer runs of positive signs. [½]

Since the trainee considered that with one more positive run we would not have rejected the null hypothesis, there must be 3 runs of positive signs. [½]

OR ALTERNATIVE 2

Using the formula we have

$$\text{for 3 positive runs, } \frac{\binom{16}{2}\binom{9}{3}}{\binom{25}{17}} = \frac{(120)(84)}{1,081,575} = 0.0093 \ll 0.05, \qquad [1]$$

$$\text{and for 4 positive runs } \frac{\binom{16}{3}\binom{9}{4}}{\binom{25}{17}} = \frac{(560)(126)}{1,081,575} = 0.065 > 0.05. \qquad [½]$$

At the 95% level we reject the null hypothesis with 3 or fewer runs of positive signs but do not reject it with 4 or more runs. [½]

Since the trainee considered that with one more positive run we would not have rejected the null hypothesis, there must be 3 runs of positive signs. [½]

OR ALTERNATIVE 3

Using the Normal approximation we have

Number of positive runs ~Normal$\left(\frac{17(9)}{25}, \frac{[(17)(8)]^2}{25^3}\right)$, which is Normal(6.12, 1.18). [½]

z-scores for smaller numbers of positive runs are:

5 positive runs: $\frac{5-6.12}{\sqrt{1.18}} = -1.031$,

4 positive runs: $\frac{4-6.12}{\sqrt{1.18}} = -1.952$, and

3 positive runs: $\frac{3-6.12}{\sqrt{1.18}} = -2.872$. [½]

Using a one-tailed test, we are looking for $z = -1.645$ at the 95% level. [½]

We reject the null hypothesis with 4 or fewer runs of positive signs. [½]

Since the trainee considered that with one more positive run we would not have rejected the null hypothesis, there must be 4 runs of positive signs. [½]

[Max 3]

[TOTAL 6]

This question tested a standard application using an unfamiliar context. In part (i) candidates divided into those who used their conceptual understanding to work out what was needed, and who scored close to full marks; and candidates who were unable to make much of an attempt. A common error was to use the "wrong end" of the distribution to produce the answer 7 positive runs. Most candidates who obtained a numerical answer to part (i) were able to use this to answer part (ii). Where candidates had an incorrect answer to part (i) full credit could still be gained for part (ii) if the answer to part (i) was followed through correctly.

Q5 (i) (a) ALTERNATIVE 1

A Poisson process is a counting process in continuous time $\{N_t, t \geq 0\}$, where N_t records the number of occurrences of a type of event within the time interval from 0 to t. [1]

Events occur singly and may occur at any time. [½]

The probability that an event occurs during the short time interval from time t to time $t + h$ is approximately equal to λh for small h, where the parameter λ is the rate of the Poisson process. [½]

OR ALTERNATIVE 2

A Poisson process is an integer valued process in continuous time $\{N_t, t \geq 0\}$, where: [½]

$$\Pr[N_{t+h} - N_t = 1 \mid F_t] = \lambda h + o(h),$$
$$\Pr[N_{t+h} - N_t = 0 \mid F_t] = 1 - \lambda h + o(h),$$
$$\Pr[N_{t+h} - N_t \neq 0,1 \mid F_t] = o(h)$$ [½]

OR ALTERNATIVE 3

A Poisson process with rate λ is a continuous-time integer-valued process $N_t, t \geq 0$, with the following properties: [½]

$N_0 = 0$,
N_t has independent, Poisson distributed stationary increments [1]

$$P[N_t - N_s = n] = \frac{[\lambda(t-s)]^n e^{-\lambda(t-s)}}{n!}, \qquad s < t,\ n = 0, 1, \ldots \qquad [½]$$

OR ALTERNATIVE 4

$\{N_t, t \geq 0\}$ is a Poisson process with rate λ if the holding times T_0, T_1, ... of

$\{N_t, t \geq 0\}$ are independent exponential random variables with parameter λ and

$N(T_0 + T_1 + \ldots + T_{n-1}) = n$.

OR ALTERNATIVE 5

$\{N_t, t \geq 0\}$ is a Poisson process with rate λ if it is a Markov jump process with independent increments and transition rates given by:

$\mu(i,j) = -\lambda$ if $j = i$,
$\mu(i,j) = \lambda$ if $j = i + 1$,
$\mu(i,j) = 0$ otherwise.

(b) Let N_t be a Poisson process, $t \geq 0$ [½]

and let $Y_1, Y_2, \ldots, Y_j, \ldots$, be a sequence of independently and identically distributed random variables. [½]

Then a compound Poisson process is defined by

$$X_t = \sum_{j=1}^{N_t} Y_j, t \geq 0. \qquad [½]$$

[Max 3]

(ii) (a) The number of claims for motorcycle accidents received on an

insurer's telephone claim line

Time inhomogeneous Poisson process [½]

Suitable reason, for example because motorcycle accidents are more likely at certain times of year/of the week and are likely to occur singly. [1]

(b) **The number of breakfast bagels sold by a New York bagel bar**

Time inhomogeneous compound Poisson process [½]

Suitable reason, for example because customers wanting breakfast goods are likely to vary according to the time of day, and if customers arrive following a Poisson process the number sold would follow a compound Poisson process as each customer might buy more than one. [1]

(c) **The number of breakdowns of freezers in a large supermarket**

Time homogeneous Poisson process [½]

Suitable reason, for example because freezers will need to be left on continuously and no reason to expect failures at a particular time of day. Freezers are likely to break down individually. [1]

(d) **The cost of wasted food caused by breakdowns of freezers in a large supermarket**

Time homogeneous compound Poisson process [½]

Suitable reason, for example if the number of failures is a time homogeneous Poisson process consistent with previous answer, the cost of each failure will vary depending on what food is stored in each freezer, how quickly the freezer is fixed, etc. Hence the cost would follow a time homogeneous compound Poisson process. [1]

[Max 6]

[TOTAL 9]

In part (ii) the marks were awarded for any suitable choice provided the explanation supported this. So, for example, in case B credit was given for a time inhomogeneous Poisson process if candidates made the point that this assumed each customer only bought one bagel. The same process could have been selected for more than one example. However, where no reason was given the mark for the process was only awarded for the models suggested above. In case D a common error was to suppose that, because

the amount of food stored in freezers varied seasonally, the process was time inhomogeneous. This is incorrect: variation in the average cost of food stored means that the Y_j which are summed in the compound Poisson process are not identically distributed.

Q6 (i) Re-write the data as shown in the table below (* denotes a person who left without making a purchase).

Customer number	*Duration*
1	8
2	2
3	6
4	6
5	2
6	4
7	10*
8	6
9	5*
10	11
11	4
12	7*

[1]

Treating those who left without making a purchase as censored we create the following table.

t_j	N_j	d_j	c_j	$\frac{d_j}{N_j}$	$1-\frac{d_j}{N_j}$
0	12	0	0		
2	12	2	0	2/12	10/12
4	10	2	1	2/10	8/10
6	7	3	1	3/7	4/7
8	3	1	1	1/3	2/3
11	1	1	0	1	0
[½]	[½]	[½]	[½]	[½]	[½]

[3]

The Kaplan-Meier estimate is thus

t	$S(t)$
$0 \le t < 2$	1
$2 \le t < 4$	0.8333
$4 \le t < 6$	0.6667
$6 \le t < 8$	0.3810
$8 \le t < 11$	0.2540
$11 \le t$	0
[1]	[1]

[2]
[Total 6]

(ii) $S(10) = 0.2540$ [½]

so the daily cost of the scheme will be $0.2540 \times 20{,}000 \times \$2 = \$10{,}159$. [½]
[Total 1]

(iii) The survey data mainly relate to the morning. We assume that the staffing levels of the check-outs relative to customer flow remain the same in the afternoons. [1]

We assume that the introduction of the compensation scheme does not change customers' behaviour (for example discouraging customers from leaving the queue). [1]

The sample size (12) is very small compared to the daily customer base (20,000) which produces a very "steppy" result. We have had to use the value for S(10) which is also the value for S(8). A larger sample size may give a smoother more accurate picture. [1]
[Max 2]
[TOTAL 9]

Many candidates scored highly on part (i). A minority of candidates treated leaving without making a purchase as the decrement of interest, and received partial credit if they applied the method correctly. In part (ii) many candidates did not use their estimate of $S(10)$ from part (i). Answers to part (iii) were encouraging, with a gratifyingly large number of candidates making sensible points. Some credit was given in part (iii) for comments about the assumptions underlying the Kaplan Meier estimate, such as the presence of non-informative censoring or the population being homogeneous.

Q7 (i) Where the probabilities depend only on the length of time interval $t - s$, the process is called time-homogeneous. [½]

If this condition is not met, the process is time-inhomogeneous. [½]

ALTERNATIVE 1

If the probabilities do not depend solely upon the length of the time interval $t-s$, the process is time-inhomogeneous. [1]

OR ALTERNATIVE 2

The transition rates vary with time. [1]

OR ALTERNATIVE 3

The transition rates depend upon the start and end times. [1]

[Max 1]

(ii) A model with time-inhomogeneous rates has more parameters, and there may not be sufficient data available to estimate these parameters. [1]

Also, the solutions to Kolmogorov's equations may not be easy (or even possible) to find analytically. [1]

Time-inhomogeneous processes are computationally harder to simulate. [½]

[Max 2]

(iii) We need $P_{\overline{GG}}(t)$, i.e. probability the process remains in G throughout time 0 to t.

This satisfies

$$\frac{d}{dt} P_{\overline{GG}}(t) = (-0.2 - 0.04t)\ P_{\overline{GG}}(t). \qquad [½]$$

Hence

$$\frac{1}{P_{\overline{GG}}(t)} \frac{d}{dt} P_{\overline{GG}}(t) = (-0.2 - 0.04t);$$

$$\frac{d}{dt} [\ln P_{\overline{GG}}(t)] = (-0.2 - 0.04t). \qquad [½]$$

Integrate both sides:

$$\left|\ln P_{\overline{GG}}(s)\right|_{s=0}^{s=t} = \left|-0.2s - 0.02s^2\right|_{s=0}^{s=t}. \qquad [1]$$

$P_{\overline{GG}}(0) = 1,$ [½]

so $P_{\overline{GG}}(t) = \exp(-0.2t - 0.02t^2).$ [½]

[Total 3]

(iv) Occurs when $P_{\overline{GG}}(t) = 0.5$, so we have [½]

$0.5 = \exp(-0.2t - 0.02t^2),$

$0.02t^2 + 0.2t - 0.69315 = 0,$ [½]

and solving using the quadratic equation formula produces

$t = 2.724$ or $t = -12.724.$ [½]

The answer lies between 0 and 8, so we require $t = 2.724.$ [½]

[Total 2]

(v) $\frac{d}{dt}P_G(t) = (-0.2 - 0.04t)\,P_G(t) + (0.4 - 0.04t)\,P_N(t)$ [½]

But $P_N(t) = 1 - P_G(t),$ [½]

So $\frac{d}{dt}P_G(t) = (-0.2 - 0.04t)\,P_G(t) + (0.4 - 0.04t)\,(1 - P_G(t)),$ [1]

or

$$\frac{d}{dt}P_G(t) = 0.4 - 0.04t - 0.6P_G(t).$$

[Total 2]

[TOTAL 10]

Answers to this question were disappointing, although most candidates made good attempts at parts (i) and (iii).

Q8 (i) Graduation by parametric formula.

Graduation by reference to a standard table.

Graphical graduation.

[Total 2]

(ii) **Parametric formula**

The resultant graduation will be sufficiently smooth provided few parameters are used. [1]

It is a suitable method for producing standard tables. [½]

It can be useful to fit the same formula to several experiences to give insight into the differences between experiences. [½]

Reference to a standard table

It can be used to fit relatively small data sets where a suitable standard table exists. [1]

The graduated rates should be smooth provided that a simple function is used. [1]

The standard table can provide information at extreme ages where data may be scanty. [½]

It can be useful to fit the same table to several experiences with the same link function to give insight into how the experience changes over time. [½]

Graphical graduation

It can be used for scanty data sets where no more sophisticated method is justifiable. [½]

It enables an experienced analyst to allow for known (or likely) features of the data. [½]

It can give a quick initial feel for the rates. [½]

[Max 4]

(iii) To test for the overall goodness of fit use the χ^2 test.

The null hypothesis is that the graduated rates are the same as the true underlying rates in the block of business. [½]

The test statistic $\sum_x z_x^2 \approx \chi_m^2$ where m is the degrees of freedom. [½]

Age	*Exposed to Risk*	*Observed Deaths*	*Graduated Rates*	*Expected Deaths*	z_x	z_x^2
40	24,584	14	0.000625	15.37	–0.34823	0.12126
41	32,587	32	0.000683	22.26	2.06521	4.26508
42	15,784	16	0.000748	11.81	1.22046	1.48953
43	21,336	22	0.000823	17.56	1.05968	1.12291
44	25,874	24	0.000908	23.49	0.10448	0.01092
45	21,544	22	0.001005	21.65	0.07485	0.00560
46	23,967	25	0.001114	26.70	–0.32866	0.10815
47	25,811	30	0.001239	31.98	–0.35010	0.12257
48	26,911	28	0.001378	37.08	–1.49162	2.22492
49	28,445	38	0.001536	43.69	–0.86105	0.74141
50	30,205	45	0.001713	51.74	–0.93717	0.87828
				Total		11.09063

[1½]

The observed test statistic is 11.09. [½]

The number of age groups is 11, [½]

but we lose an unknown number of degrees of freedom for the choice of the standard table, say 2, [½]

and a further two for the parameters in the link function. [½]

So $m = 7$ say (8 or 6 also acceptable). [½]

The critical value of the χ^2 distribution with 7 degrees of freedom at the 95% significance level is 14.07 (6 d.f.12.59, 8 d.f. 15.51). [½]

Since 11.09 < 14.07 (or 12.59, or 15.51) [½]

We have insufficient evidence to reject the null hypothesis. [½]

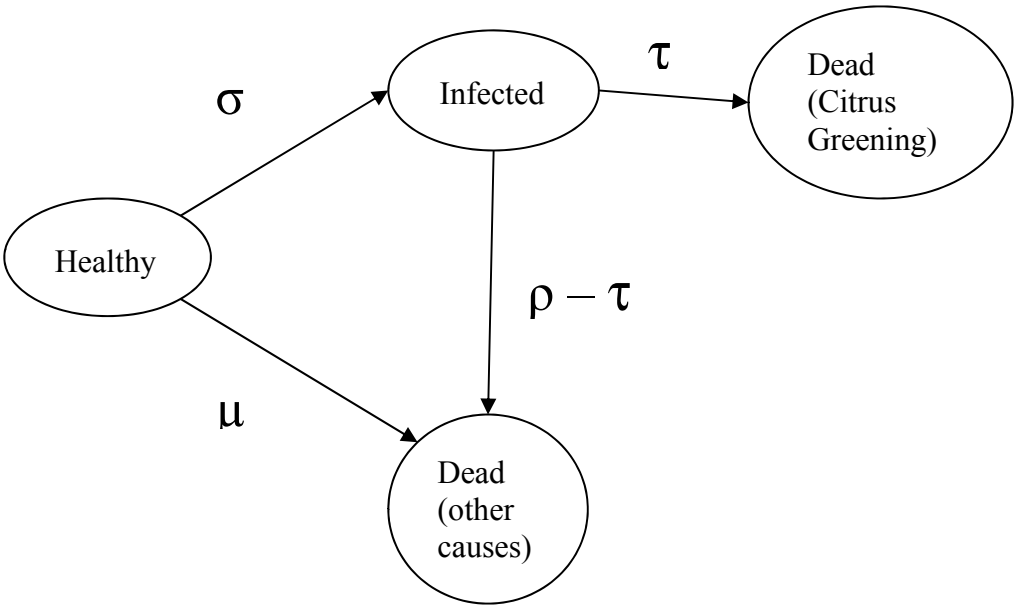
[Max 6]

[TOTAL 12]

Almost all candidates knew three methods of graduation. Part (ii) was poorly answered, with a substantial minority of candidates simply describing the three methods rather than their advantages. In part (iii) a common error was to consider that the link function only involved one parameter, where there were two (0.94 and 0.0001). However, the weakest element of candidates' answers was the description of the null hypothesis, with many candidates

writing incorrect formulations, such as "the crude rates were equal to the graduated rates", or that "the graduated rates were the same as the rates in the standard table".

Q9 (i)



[Total 2]

(ii) Forward equations are:

$$\frac{d}{dt}P(t) = P(t)A\,, \qquad [1]$$

where A is the generator matrix:

$$A = \begin{pmatrix} -\sigma-\mu & \sigma & 0 & \mu \\ 0 & -\rho & \tau & \rho-\tau \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad [1½]$$

with the order of states being
{Healthy, Infected, Dead (caused by Citrus Greening),
Dead (other causes)}. [½]

[Total 3]

(iii) ALTERNATIVE 1

To estimate τ we re-parameterise so that a new parameter $\xi = \rho - \tau$ is the death rate from other causes of infected trees. [½]

The likelihood of the data can then be written

$$L \propto \exp\left[(-\mu-\sigma)v^H\right]\exp\left[(-\xi-\tau)v^I\right](\sigma)^{d^{HI}}(\mu)^{d^{HDOC}}(\tau)^{d^{IDCG}}(\xi)^{d^{IDOC}} \quad [2]$$

where v^H and v^I are the waiting times in the Healthy and Infected states respectively,

d^{HI} is the number of transitions from Healthy to Infected,

d^{HDOC} is the number of transitions from Healthy to Dead from Other Causes

d^{IDCG} is the number of transitions from Infected to Dead from Citrus Greening

and d^{IDOC} is the number of transitions from Infected to Dead from Other Causes. [½]

OR ALTERNATIVE 2

$$L \propto \exp\left[(-\mu-\sigma)v^H\right]\exp\left[-\tau-(\rho-\tau)v^I\right](\sigma)^{d^{HI}}(\mu)^{d^{HDOC}}(\tau)^{d^{IDCG}}(\rho-\tau)^{d^{IDOC}}, \quad [2]$$

which equals

$$L \propto \exp\left[(-\mu-\sigma)v^H\right]\exp\left[-\rho v^I\right](\sigma)^{d^{HI}}(\mu)^{d^{HDOC}}(\tau)^{d^{IDCG}}(\rho-\tau)^{d^{IDOC}}, \quad [½]$$

where v^H and v^I are the waiting times in the Healthy and Infected states respectively,

d^{HI} is the number of transitions from Healthy to Infected,

d^{HDOC} is the number of transitions from Healthy to Dead from Other Causes

d^{IDCG} is the number of transitions from Infected to Dead from Citrus Greening

and d^{IDOC} is the number of transitions from Infected to Dead from Other Causes.

[½]

OR ALTERNATIVE 3

Since we have 40 deaths in total, 10 of healthy trees and 30 from Citrus Greening, then no infected tree dies from a cause other than Citrus Greening.

[½]

Hence $\rho = \tau$ and the likelihood of the data can be written

$$L \propto \exp\left[1200(-\mu-\sigma)\right]\exp\left[-600\tau\right](\sigma)^{d^{HI}}(\mu)^{10}(\tau)^{30}, \qquad [2]$$

where

d^{HI} is the number of transitions from Healthy to Infected. [½]

[Max 3]

(iv) Taking logarithms of the likelihood we have:

$\log_e L = -\tau\nu^I + d^{IDCG}\log_e(\tau)$ + terms not dependent on τ. [½]

Partially differentiating with respect to τ gives:

$$\frac{d(\log_e L)}{d\tau} = -\nu^I + \frac{d^{IDCG}}{\tau}. \qquad [½]$$

Setting the derivative to zero [½]

we obtain the maximum likelihood estimator:

$$\hat{\tau} = \frac{d^{IDCG}}{\nu^I}. \qquad [½]$$

The second derivative of the log likelihood is

$$\frac{d^2(\log_e L)}{(d\tau)^2} = -\frac{d^{IDCG}}{(\tau)^2}, \qquad [½]$$

which is negative, so this is a maximum. [½]

OR ALTERNATIVE 2

Taking logarithms of the likelihood we have:

$\log_e L = d^{IDCG} \log_e(\tau) + d^{IDOC} \log_e(\rho - \tau)$ + terms not dependent on τ. [½]

Partially differentiating with respect to τ gives:

$$\frac{d(\log_e L)}{d\tau} = \frac{d^{IDCG}}{\tau} - \frac{d^{IDOC}}{\rho - \tau}.$$ [½]

Setting the derivative to zero [½]

we obtain the maximum likelihood estimator:

$$\hat{\tau} = \frac{\rho d^{IDCG}}{d^{IDOC} + d^{IDCG}}.$$ [½]

The second derivative of the log likelihood:

$$\frac{d^2(\log_e L)}{(d\tau)^2} = -\frac{d^{IDCG}}{(\tau)^2} - \frac{d^{IDOC}}{(\rho - \tau)^2},$$ [½]

is negative, therefore this is a maximum. [½]

[Max 3]

(v) 30/600 = 0.05 (per tree-month). [Total 1]

[TOTAL 12]

Several candidates tried to write the Kolmogorov equations in part (ii) in component form. Credit was given for this if the resulting equations were correct. ALTERNATIVE 1 in part (iv) follows from ALTERNATIVES 1 and 3 in part (iii). ALTERNATIVE 2 in part (iv) follows from ALTERNATIVE 2 in part (iii). To obtain a numerical estimate of τ from ALTERNATIVE 2 in part (iv) it was necessary also to differentiate the logarithm of the likelihood with respect to ρ, set this derivative to zero, and solve the resulting simultaneous equations. A few candidates did this, but it was not required for full credit. Credit was given in part (v) for the correct numerical answer even if this did not follow from the answers to previous sections.

Q10 (i) To provide cover for at least five years before she changes provider, Yolanda must renew her policy at least four times. [1]

The probability of renewing four times is $0.4^4 = 0.0256$ (or 16/625). [1]
[Total 2]

(ii) The company covering the house on 12 March 2015 will be that securing Zachary's business at the second renewal. [½]

The second order transition matrix is:

$$\begin{pmatrix} 0.5 & 0.2 & 0.2 & 0.1 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0.3 & 0.2 & 0.4 & 0.1 \\ 0 & 0.2 & 0.2 & 0.6 \end{pmatrix} * \begin{pmatrix} 0.5 & 0.2 & 0.2 & 0.1 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0.3 & 0.2 & 0.4 & 0.1 \\ 0 & 0.2 & 0.2 & 0.6 \end{pmatrix} = \begin{pmatrix} 0.35 & 0.28 & 0.22 & 0.15 \\ 0.25 & 0.44 & 0.16 & 0.15 \\ 0.31 & 0.28 & 0.26 & 0.15 \\ 0.1 & 0.28 & 0.22 & 0.4 \end{pmatrix}$$

[1]

So the probability of being with Company *A* is 0.35,

and hence the probability of not being with Company *A* is 0.65. [½]
[Total 2]

(iii) The long run probabilities satisfy

$\pi P = \pi$ [½]

$0.5\pi_A + 0.2\pi_B + 0.3\pi_C = \pi_A$ (1)

$0.2\pi_A + 0.6\pi_B + 0.2\pi_C + 0.2\pi_D = \pi_B$ (2)

$0.2\pi_A + 0.1\pi_B + 0.4\pi_C + 0.2\pi_D = \pi_C$ (3)

$0.1\pi_A + 0.1\pi_B + 0.1\pi_C + 0.6\pi_D = \pi_D$ (4) [1]

and $\pi_A + \pi_B + \pi_C + \pi_D = 1$. (5) [½]

(4) gives (using (5))

$0.1\pi_A + 0.1\pi_B + 0.1\pi_C + 0.1\pi_D = 0.5\pi_D = 0.1$,

so

$\pi_D = \frac{1}{5}$.

(3)–(2) gives

$0.5\pi_B + 0.2\pi_C = \pi_C$

so $\pi_B = \frac{8}{5}\pi_C$.

(1) gives

$\pi_A = \frac{31}{25}\ \pi_C$

so $\left(\frac{31}{25}+\frac{40}{25}+1\right)\pi_C + 0.2 = 1.$ [1]

Hence $\pi_A = 31/120$, $\pi_B = 1/3$, $\pi_C = 5/24$, and $\pi_D = 1/5$.

So the long run probabilities are 0.2583, 0.3333, 0.2083 and 0.2 [½]

for companies *A*, *B*, *C* and *D* respectively. [½]

[Total 4]

(iv) The matrix would be:

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.6 & 0.1 \\ 0.4 & 0.2 & 0.4 \end{pmatrix} \end{array}$$

[Total 2]

(v) There may be reasons customers of Company *D* do not want to use Company *A*. [½]

Observe that currently the rate of customers going from Company *D* to Company *A* is zero. [½]

Adddda might merge its pricing system. This would change the relative pricing of an individual's cover from the different companies. To the extent that pricing is a driver of the likelihood of customers moving this might change the probabilities. [1]

To the extent that customer service is a driver, it is not clear what the customer service of *Adddda* would be relative to Company *A* or Company *D*. This might change the probabilities. [1]

Reduction in competition might encourage a new entrant. [½]

It might be a valid assumption that customer behaviour continues unaltered after the merger. [½]

[Max 2]

[TOTAL 12]

Answers to this question were disappointing, particularly parts (i) and (ii). In part (i), a very large number of candidates did not spot that five years' continuous cover with the same provides only requires the decision not to change to be made four times.

Q11 (i) $h(t, z_i) = h_0(t)\exp\left(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4\right)$, where [½]

$h(t, z_i)$ is the hazard at time t; [½]

$h_0(t)$ is the baseline hazard; [½]

$\beta_1 \ldots \beta_4$ are regression parameters; [½]

z_1 is a covariate which takes the value 1 if the client is Male, 0 otherwise;

z_2 is a covariate which takes the value 1 if energy consumption is high, 0 otherwise;

z_3 is a covariate which takes the value 1 if the area of residence is City Centre, 0 otherwise; and

z_4 is a covariate which takes the value 1 if the area of residence is Rural, 0 otherwise. [1]

[Max 3]

(ii) The baseline hazard refers to a female with low energy consumption who lives in a city but not in the city centre. [Total 1]

(iii) The 95% confidence interval for β is $\beta \pm 1.96\sqrt{\text{Var}(\beta)}$ so the intervals are:

Male	–0.4900, –0.0100
Female	0
High consumption	0.1447, 0.4953
Low consumption	0
City Centre	–0.0247, 0.4047
City (not centre)	0
Rural	–0.4886, –0.2114

[½ for each correct interval] [Max 2]

(iv) The parameter associated with males is –0.25 so for two otherwise identical clients, the transfer rates for males is exp (–0.25) = 0.7788
OR
The hazard ratio between the transfer rates for women and men is 1/exp(–0.25) =1.284. [1]

Therefore women do seem to transfer more than men (or men less than women). [½]

The 95% confidence interval for the parameter does not span zero
OR
the z-score for the parameter is $0.25/\sqrt{0.015} = 2.04$, and this is greater than 1.96. [1]

So at the 95% confidence level we can state that women do switch providers more frequently than men. [½]
[Total 3]

(v) ALTERNATIVE 1

For the rural male, the probability that he has transferred is 0.7, the sum of the parameters is –0.25 + 0 – 0.35 = –0.6 and the hazard is $h_0(t)\exp(-0.6)$. [½]

So the probability that the contract is still in force is

$$0.3 = \exp\left\{-\int_0^2 h_0(t)\exp(-0.6)dt\right\} = \exp\left\{-0.5488\int_0^2 h_0(t)dt\right\}. \quad [½]$$

$$\text{So} \quad \int_0^2 h_0(t)dt = \frac{\ln 0.3}{0.5488} \quad [½]$$

For the City Centre male, the sum of the parameters is 0.26. [½]

$$\text{So we want} \quad \exp\left\{-\int_0^2 h_0(t)\exp(0.26)dt\right\} = \exp\left\{-1.2969 * \frac{\ln 0.3}{0.5488}\right\} \quad [½]$$

= 0.058124. [½]

OR ALTERNATIVE 2

$$\left\{0.3^{e^{0.6}}\right\}^{e^{0.26}} = 0.3^{e^{0.86}} \quad [2½]$$

= 0.058124, [½]

which is the probability that he is still with the company.

[Max 3]

(vi) For each pair of covariates z_i and z_j: [½]

fit a model with the original covariates plus the interaction between the pair as an extra covariate.

OR

fit a model with the original covariates and a term $z_i * z_j$. [1]

If the log-likelihood for each of the models are L_{original} and $L_{\text{with interaction}}$, [½]

then the test statistic is $-2(L_{\text{original}} - L_{\text{with interaction}})$. [1]

The null hypothesis is that the parameter for the interaction term is zero. [½]

The test statistic has a chi-squared distribution with one degree of freedom. [½]

If the test statistic is greater than 3.84 (at the 5% level of significance)

OR

If the 95% confidence interval around the interaction parameter does not include zero, [½]

we can reject the null hypothesis and conclude that the interaction term is needed. [½]

[Max 5]

[TOTAL 17]

Most candidates successfully wrote down the equation in part (i) and defined the covariates. Most candidates also correctly identified the characteristics of the person whose hazard was equal to the baseline in part (ii). Common errors in part (iii) were failure to multiply by 1.96 or to take the square root of the variance. Parts (iv) and (v) were disappointingly answered. In part (vi) several candidates knew that a likelihood ratio test was required but were rather vague about the details. These candidates were given limited credit.

END OF EXAMINERS' REPORT