# EXAMINATION

21 April 2010 (am)

## Subject CT4 — Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 12 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** List four factors often used to subdivide life insurance mortality statistics. [2]

**2** Write down integral equations for the mean and variance of the complete future lifetime at age $x$, $T_x$. [2]

**3** For each of the following processes:

counting process;
general random walk;
compound Poisson process;
Poisson process;
Markov jump chain.

(a) State whether the state space is discrete, continuous or can be either.

(b) State whether the time set is discrete, continuous, or can be either.

[5]

**4** A Markov Chain with state space {A, B, C} has the following properties:

- it is irreducible
- it is periodic
- the probability of moving from A to B equals the probability of moving from A to C

(i) Show that these properties uniquely define the process. [4]

(ii) Sketch a transition diagram for the process. [1]
[Total 5]

**5** Ten years ago, a confectionery manufacturer launched a new product, the Scrummy Bar. The product has been successful, with a rapid increase in consumption since the product was first sold. In order to plan future investment in production capacity, the manufacturer wishes to forecast the future demand for Scrummy Bars. It has data on age-specific consumption rates for the past ten years, together with projections of the population by age over the next twenty years. It proposes the following modelling strategy:

- extrapolate past age-specific consumption rates to forecast age-specific consumption rates for the next 20 years
- apply the forecast age-specific consumption rates to the projected population by age to obtain estimated total consumption of the product by age for each of the next 20 years
- sum the results to obtain the total demand for each year

Describe the advantages and disadvantages of this strategy. [5]

**6** An oil company has discovered a vast deposit of oil in an equatorial swamp. The area is extremely unhealthy and inhabited by venomous spiders. There is an antidote to bites from these spiders but it is expensive. The antidote acts instantly but does not provide future immunity. The company commissions a study to estimate the rate of being bitten by the spiders among its employees, in order to determine the amount of antidote to provide.

Employees of the company are posted to the swamp for six month tours of duty starting on 1 January, 1 April, 1 July or 1 October. The first employees to be posted arrived on 1 January 2008. The swamp is so inaccessible that no employees are allowed to leave before their six month tours of duty are completed.

Accidental deaths are common in this dangerous location.

The table below gives some data from the study.

| *Quarter beginning* | *Number of new arrivals at start of quarter* | *Number of accidental deaths during quarter* | *Number of spider bites during quarter* |
|---|---|---|---|
| 1 January 2008 | 90 | 10 | 15 |
| 1 April 2008 | 80 | 8 | 25 |
| 1 July 2008 | 114 | 10 | 30 |
| 1 October 2008 | 126 | 13 | 40 |

(i) Estimate the quarterly rate of being bitten by a spider for each quarter of 2008, stating any assumptions you make. [7]

(ii) Suggest reasons why the assumptions you made in (i) might not be valid. [1]
[Total 8]

**7** A government has introduced a two-tier driving test system. Once someone applies for a provisional licence they are considered a Learner driver. Learner drivers who score 90% or more on the primary examination (which can be taken at any time) become Qualified. Those who score between 50% and 90% are obliged to sit a secondary examination and are given driving status Restricted. Those who score 50% or below on the primary examination remain as Learners. Restricted drivers who pass the secondary examination become Qualified, but those who fail revert back to Learner status and are obliged to start again.

(i) Sketch a diagram showing the possible transitions between the states. [2]

(ii) Write down the likelihood of the data, assuming transition rates between states are constant over time, clearly defining all terms you use. [3]

Figures over the first year of the new system based on those who applied for a provisional licence during that time in one area showed the following:

| | |
|---|---|
| Person-months in Learner State | 1,161 |
| Person-months in Restricted State | 1,940 |
| Number of transitions from Learner to Restricted | 382 |
| Number of transitions from Restricted to Learner | 230 |
| Number of transitions from Restricted to Qualified | 110 |
| Number of transitions from Learner to Qualified | 217 |

(iii) (a) Derive the maximum likelihood estimator of the transition rate from Restricted to Learner.

(b) Estimate the constant transition rate from Restricted to Learner.

[3]
[Total 8]

**8** A certain profession admits new members to the status of student. Students may qualify as fellows of the profession by virtue of passing a series of examinations. Normally student members sit the examinations whilst working for an employer. There are two sessions of the examinations each year.

An employer provides study support to student members of the profession. It wishes to assess the cost of providing this study support and therefore wishes to know the average time it can expect to take for its students to qualify.

The employer has maintained records for 23 of its students who all sat their first examination in the first session of 2003. The students' progress has been recorded up to and including the last session of 2009. The following data records the number of sessions which had been held before the specified event occurred for a student in this cohort:

| | |
|---|---|
| Qualified | 6, 8, 8, 9, 9, 9, 11, 11, 13, 13, 13 |
| Stopped studying | 4, 5, 8, 11, 14 |

The remaining seven students were still studying for the examinations at the end of 2009.

(i) Determine the median number of sessions taken to qualify for those students who qualified during the period of observation. [2]

(ii) Calculate the Kaplan-Meier estimate of the survival function, $S(t)$, for the "hazard" of qualifying, where $t$ is the number of sessions of examinations since 1 January 2003. [5]

(iii) Hence estimate the median number of sessions to qualify for the students of this employer. [2]

(iv) Explain the difference between the results in (i) and (iii) above. [2]

[Total 11]

**9** (i) Write down the hazard function for the Cox proportional hazards model defining all the terms that you use. [2]

A farmer is concerned that he is losing a lot of his birds to a predator, so he decides to build a new enclosure using taller fencing. This fencing is expensive and he cannot afford to build a large enough area for all his birds. He therefore decides to put half his birds in the new enclosure and leave the others in the existing enclosure. He is convinced that the new enclosure is an improvement, but has asked an actuarial student to determine whether the new enclosure will result in an increase in the life expectancy of his birds. The student has fitted a Cox proportional hazards model to data on the duration until a bird is killed by a predator and calculated the following figures relating to the regression parameters:

| | | *Parameter estimate* | *Variance* |
|---|---|---|---|
| *Bird* | Chicken | 0 | 0 |
| | Duck | –0.210 | 0.002 |
| | Goose | 0.075 | 0.004 |
| *Enclosure* | New | 0.125 | 0.0015 |
| | Old | 0 | 0 |
| *Sex* | Male | 0.2 | 0.0026 |
| | Female | 0 | 0 |

(ii) State the features of the bird to which the baseline hazard applies. [1]

(iii) For each regression parameter:

(a) Define the associated covariate.

(b) Calculate the 95% confidence interval based on the standard error. [3]

(iv) Comment on the farmer’s belief that the new enclosure will result in an increase in his birds’ life expectancy. [2]

(v) Calculate, using this model, the probability that a female duck in the new enclosure has been killed by a predator at the end of six months, given that the probability that a male goose in the old enclosure has been killed at the end of the same period is 0.1 (all other decrements can be ignored). [4]

[Total 12]

**10** An airline runs a frequent flyer scheme with four classes of member: in ascending order Ordinary, Bronze, Silver and Gold. Members receive benefits according to their class. Members who book two or more flights in a given calendar year move up one class for the following year (or remain Gold members), members who book exactly one flight in a given calendar year stay at the same class, and members who book no flights in a given calendar year move down one class (or remain Ordinary members).

Let the proportions of members booking 0, 1 and 2+ flights in a given year be $p_0$, $p_1$ and $p_{2+}$ respectively.

(i) (a) Explain how this scheme can be modelled as a Markov chain.

(b) Explain why there must be a unique stationary distribution for the proportion of members in each class.

[3]

(ii) Write down the transition matrix of the process. [1]

The airline's research has shown that in any given year, 40% of members book no flights, 40% book exactly one flight, and 20% book two or more flights.

(iii) Calculate the stationary probability distribution. [5]

The cost of running the scheme per member per year is as follows:

| | |
|---|---|
| Ordinary members | £0 |
| Bronze members | £10 |
| Silver members | £20 |
| Gold members | £30 |

The airline makes a profit of £10 per passenger for every flight before taking into account costs associated with the frequent flyer scheme.

(iv) Assess whether the airline makes a profit on the members of the scheme. [4]

[Total 13]

**11** A reinsurance policy provides cover in respect of a single occurrence of a specified catastrophic event. If such an event occurs, future cover is suspended. However if a reinstatement premium is paid within one time period of occurrence of the event then the insurance coverage is reinstated. If a second specified event occurs it is not permitted to reinstate the cover and the policy will lapse.

The transition rate for the hazard of the specified event is a constant 0.1. Whilst policies are eligible for reinstatement, the transition rate for resumption of cover through paying a reinstatement premium is 0.05.

(i) Explain whether a time homogeneous or time inhomogeneous model would be more appropriate for modelling this situation. [2]

(ii) (a) Explain why a model with state space {Cover In Force, Suspended, Lapsed} does not possess the Markov property.

(b) Suggest, giving reasons, additional state(s) such that the expanded system would possess the Markov property. [3]

(iii) Sketch a transition diagram for the expanded system. [2]

(iv) Derive the probability that a policy remains in the Cover In Force state continuously from time 0 to time $t$. [2]

(v) Derive the probability that a policy is in the Suspended state at time $t > 1$ if it is in state Cover In Force at time 0. [5]

[Total 14]

**12** (i) State three different methods of graduating raw mortality data and for each method give an example of a situation when the method would be appropriate. [3]

A life insurance company last priced its whole of life contract 30 years ago using a standard mortality table. The company wishes to establish whether recent mortality experience in the portfolio of business is in line with the pricing basis. These are the data:

| *Recent Experience* | | | *Extract from the standard table used for pricing the product* | |
|---|---|---|---|---|
| *Age last birthday* | *Exposed to Risk during 2009* | *Deaths during 2009* | *x* | *Number of survivors to age x* |
| 50 | 2,381 | 16 | 50 | 32,669 |
| 51 | 3,177 | 21 | 51 | 32,513 |
| 52 | 3,460 | 22 | 52 | 32,338 |
| 53 | 1,955 | 15 | 53 | 32,143 |
| 54 | 3,122 | 24 | 54 | 31,926 |
| 55 | 3,485 | 29 | 55 | 31,685 |
| 56 | 2,781 | 26 | 56 | 31,417 |
| 57 | 3,150 | 31 | 57 | 31,121 |
| 58 | 3,651 | 39 | 58 | 30,795 |
| 59 | 3,991 | 48 | 59 | 30,435 |
| | | | 60 | 30,039 |

(ii) Test the goodness of fit of these data with the pricing basis and comment on your results. [8]

(iii) (a) State, with reasons, one further test which you would deem appropriate to perform on these data.

(b) Carry out that test.

[4]
[Total 15]

## END OF PAPER

# EXAMINERS' REPORT

April 2010 Examinations

# Subject CT4 — Models
# Core Technical

**Introduction**

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

R D Muckart
Chairman of the Board of Examiners

July 2010

**Comments**

*Comments on solutions presented to individual questions for the April 2010 examination paper are given below. In general, those using this report should be aware that in the case of non-numerical answers full credit could often be obtained for rather less than is given in the solutions which follow. The solutions are meant as a guide to the various points which could have been made and considered relevant.*

*Questions without comments in this section were generally well answered, and no specific issues were identified.*

**Q3** *A common error was to confuse a Markov Jump Chain with a Markov Jump Process. A Markov Jump Chain has a discrete time set, whereas the corresponding Markov Jump Process has a continuous time set.*

**Q4** *This was poorly answered. In part (i), many candidates merely gave definitions of the terms "periodic" and "irreducible", rather than applying them to the question. In part (ii), many candidates simply drew the three states with arrows denoting all possible transitions between them.*

**Q5** *Answers to this question were disappointing. Many candidates simply wrote down general lists of the advantages and disadvantages of models, without reference to the problem and the modelling strategy described in the question. Such attempts were given little credit.*

**Q6** *This was a fairly difficult exposed-to-risk question and many candidates found it challenging. A common error was to include only the first quarter of each employee's tour of duty in the exposed-to-risk. Many candidates assumed that all accidental deaths happened at the end of each quarter. This seems unrealistic and was penalised, though credit was given for computations of the exposed-to-risk that were correct given this assumption. In part (ii), a large number of candidates made no sensible attempt to analyse their own assumptions made in part (i).*

**Q8** *Parts (iii) and (iv) of this troubled most candidates. Only a minority realised that, since the Kaplan-Meier estimator is a step function, the point at which S(t) attains the value must lie on one of the "risers" of the steps and therefore be at one of the event durations. In part (iv), many candidates realised that the median estimated in part (iii) included the candidates who had not qualified by the last session of 2009, whereas the median in part (i) did not, but were unable to argue coherently that this meant that the median in part (i) was biased, and under-estimated the true median.*

**Q9** *In part (iv), a disturbingly large number of candidates (the majority) wrote that since the parameter was positive, the life expectancy must have increased. In fact the opposite is the case. The positive parameter increases the hazard, which leads to a greater risk of death and hence a decline in the birds' life expectancy in the new enclosure. In part (v) a common error was to assume that 0.1 was the probability of survival, rather than the probability of being killed.*

**Q10** *In part (iv) a common error was to assume that every member makes exactly one flight. This produces a profit per member of £2.67 compared with the true profit of at least £0.67, and more if some members make more than two flights per year.*

**Q11** *This difficult question was a challenge for almost all candidates. In part (iv) many candidates simply wrote down the probability rather than deriving it. Credit was given for attempts to part (v) which made use of an integrating factor.*

**Q12** *In part (iii), many candidates chose the Signs Test. Since from the answer to part (ii) there are five consecutive negative signs followed by five consecutive positive signs it is clear by inspection that the experience will "pass" the Signs Test and so carrying it out is not appropriate. Hence no credit was given for the Signs Test in part (iii)(a). It is much more sensible to conduct a Grouping of Signs Test. As has been the case in previous examinations, the numerical aspects of the tests were generally well performed, but the descriptions of the tests and explanations of what was being done and why were less consistent.*

**1** *Sex*
Age
Type of policy
Smoker/non-smoker status
Level of underwriting OR lifestyle/participation in dangerous sports
Duration in force
Sales channel
Policy size
Occupation of policyholder
Known impairments
Post code OR region/county/country OR address

*Marks were given for up to four factors from the list above.*

**2** $E[T_x] = \dot{e}_x = \int_0^{\omega-x} {}_t p_x dt$ OR $E[T_x] = \dot{e}_x = \int_0^{\omega-x} t\, {}_t p_x \mu_{x+t} dt$

$$\text{Var}[T_x] = \left\{\int_0^{\omega-x} t^2\, {}_t p_x \mu_{x+t} dt\right\} - \dot{e}_x{}^2$$

*The upper limits to the integrals can also be anything above ω-x, for example ω or ∞, since any age above ω-x just adds zero to the summation.*

**3**

| | *State Space* | *Time Set* |
|---|---|---|
| Counting Process | Discrete | Discrete or Continuous |
| General Random Walk | Discrete or Continuous | Discrete |
| Compound Poisson Process | Discrete or Continuous | Continuous |
| Poisson Process | Discrete | Continuous |
| Markov Jump Chain | Discrete | Discrete |

**4** (i) As periodic and irreducible then all states are periodic, hence probability of staying in any state is zero.

By law of total probability, $P_{AA} + P_{AB} + P_{AC} = 1$.

But $P_{AB} = P_{AC}$ and $P_{AA} = 0$ so $P_{AB} = P_{AC} = 0.5$.

To be irreducible at least one of $P_{BA}$ or $P_{CA}$ must be greater than zero.

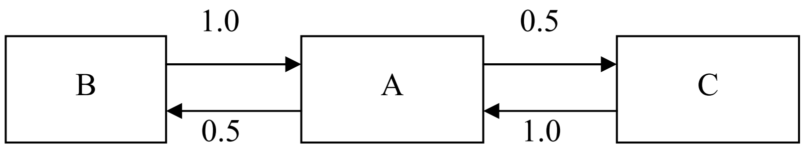
If $P_{BA} > 0$ then to be periodic must have $P_{CB} = 0$,

and to be irreducible $P_{CA} > 0$,

and if $P_{CA} > 0$ then to be periodic must have $P_{BC} = 0$, and to be irreducible $P_{BA} > 0$.

So must have $P_{BC} = P_{CB} = 0$ and $P_{BA} = P_{CA} = 1$.

(ii)




**5** **Advantages**

The model is simple to understand and to communicate.

The model takes account of one major source of variation in consumption rates, specifically age.

The model is easy and cheap to implement.

The past data on consumption rates by age are likely to be fairly accurate.

The model can be adapted easily to different projected populations OR takes into account future changes in the population.

**Disadvantages**

Past trends in consumption by age may not be a good guide to future trends.

Extrapolation of past age-specific consumption rates may be complex or difficult and can be done in different ways.

Consumption of chocolate may be affected by the state of the economy, e.g. whether there is a recession.

Factors other than age may be important in determining consumption, e.g. expenditure on advertising.

Consumption may be sensitive to pricing, which may change in the future.

A rapid increase in consumption rates is unlikely to be sustained for a long period as there is likely to be an upper limit to the amount of Scrummy Bars a person can eat.

The projections of the future population by age may not be accurate, as they depend on future fertility, mortality and migration rates.

The proposed strategy does not include any testing of the sensitivity of total demand to changes in the projected population, or variations in future consumption trends from that used in the model.

Unforeseen events such as competitors launching new products, or the nation becoming increasingly health-aware, may affect future consumption.

The consumption of Scrummy Bars may vary with cohort rather than age, and the model does not capture cohort effects.

*Not all the points listed above were required for full credit. Other advantages, for example those related to business prospects, were also given credit.*

**6** (i) A central exposed to risk for each quarter in person-quarters can be constructed as follows.

In the first quarter there are 90 employees in the first three months of their six-month tour of duty. Of these 10 will die during the quarter, and these contribute 0.5 each to the exposed to risk.

Therefore the total exposed to risk for the first quarter is $80 + (10 \times 0.5) = 85$ person-quarters.

This assumes that accidental deaths occur on average half way through the quarter in which they were reported. OR that accidental deaths are uniformly distributed across quarters.

In the second quarter there are 80 new employees in the first three months of their six-month tour, and 80 (90 minus the 10 who have died) employees in the second three months of their six-month tour. Of these 8 die during the quarter, and these contribute 0.5 each to the exposed to risk.

Therefore the total exposed to risk for the second quarter is $152 + (8 \times 0.5) = 156$ person-quarters

In the third quarter there are 114 new employees in the first three months of their six-month tour, and 76 (the 80 who were new on 1 April 2009 minus half of the 8 who died in the second quarter) employees in the second three months of their six-month tour. Of these 10 die during the quarter, and these contribute 0.5 months each to the exposed to risk.

This assumes that accidental deaths are equally likely for employees in the first quarter of their tour of duty, and those in the second quarter of their tour of duty.

Therefore the total exposed to risk for the third quarter is $180 + (10 \times 0.5) = 185$ person-quarters

Finally, in the fourth quarter there are 126 new employees in the first three months of their six-month tour, and 108 (the 114 who were new on 1 April 2009 minus a proportion equal to $114/(114 + 76) = 0.6$ of the 10 who died in the third quarter) employees in the second three months of their six-month tour.

Of these 13 died during the quarter, and these contribute 0.5 quarters each to the exposed to risk.

Therefore the total exposed to risk for the fourth quarter is $221 + (13 \times 0.5) = 227.5$ person-quarters.

We assume there are no deaths apart from accidental deaths.

These calculations are summarised in the table below.

| *Quarter beginning* | *Employees in first quarter of tour* | *Employees in second quarter of tour* | *Less 0.5 × accidental deaths* | *Central exposed to risk in quarters* |
|---|---|---|---|---|
| 1 January | 90 | 0 | 5 | 85 |
| 1 April | 80 | 80 | 4 | 156 |
| 1 July | 114 | 76 | 5 | 185 |
| 1 October | 126 | 108 | 6.5 | 227.5 |

The quarterly rates of being bitten are therefore as follows:

| *Quarter beginning* | *Spider bites* | *Exposed to risk* | *Rate of being bitten* | |
|---|---|---|---|---|
| 1 January | 15 | 85 | 15/85 | = 0.176 |
| 1 April | 25 | 156 | 25/156 | = 0.160 |
| 1 July | 30 | 185 | 30/185 | = 0.162 |
| 1 October | 40 | 227.5 | 40/227.5 | = 0.176 |

We assume that all spider bites are treated.

(ii) The assumption that there are no deaths apart from accidental deaths is unlikely to be true, and probably the company would have data on these which could be included in the calculations.

Accidental deaths may be more likely among employees in their first quarter than their second, as those in their second quarter have more experience.

Accidental deaths may be more likely at the beginning of a quarter, when there are newly arrived employees.

The experience of the quarter beginning 1 January may be different from that of other quarters because that is the first quarter that any employees are stationed in the swamp, and they may not know about the spiders when they arrive. In subsequent quarters they may be able to adjust their arrangements to reduce the possibility of being bitten.

*Several alternatives to part (i) were also given credit. For example assuming spider bites are all fatal produces the following solution to part (i):*

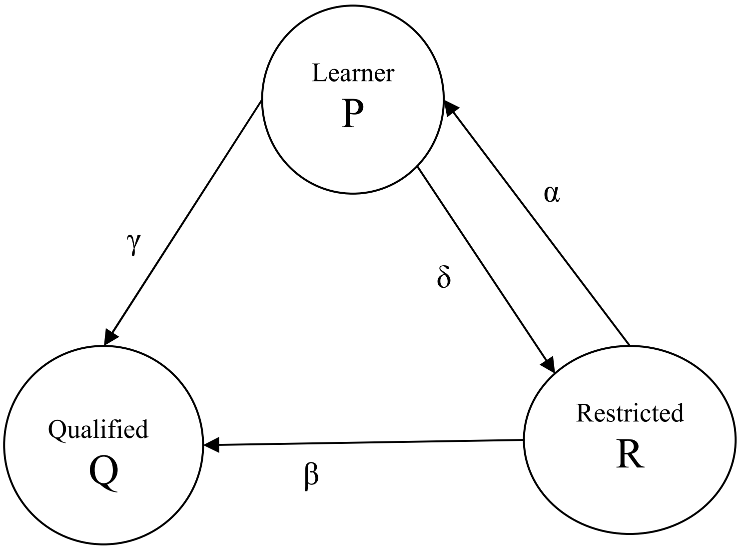
| *Quarter beginning* | *Employees in first quarter of tour* | *Employees in second quarter of tour* | *Less 0.5 × total deaths* | *Central exposed to risk in quarters* |
|---|---|---|---|---|
| *1 January* | *90* | *0* | *12.5* | *77.5* |
| *1 April* | *80* | *65* | *16.5* | *128.5* |
| *1 July* | *114* | *62* | *20* | *156.0* |
| *1 October* | *126* | *88* | *26.5* | *187.5* |

*The quarterly rates of being bitten are therefore as follows:*

| *Quarter beginning* | *Spider bites* | *Exposed to risk* | *Rate of being bitten* |
|---|---|---|---|
| *1 January* | *15* | *77.5* | *15/77.5 = 0.194* |
| *1 April* | *25* | *128.5* | *25/128.5 = 0.195* |
| *1 July* | *30* | *156* | *30/156 = 0.192* |
| *1 October* | *40* | *187.5* | *40/187.5 = 0.213* |

*In part (ii) credit was only given if the points made related to one of the assumptions stated in the answer to part (i).*

**7** (i)




(ii) Let α be the transition rate R to P
β be the transition rate R to Q
γ be the transition rate P to Q
δ be the transition rate P to R

Let $P$ be the time spent in Learner state
$R$ be the time spent in Restricted state

Let $a$ be the number of transitions from Restricted to Learner
$b$ be the number of transitions from Restricted to Qualified
$c$ be the number of transitions from Learner to Qualified
$d$ be the number of transitions from Learner to Restricted

$$L \propto \exp\{(-\delta-\gamma)P\}\exp\{(-\alpha-\beta)R\}\alpha^a\beta^b\gamma^c\delta^d$$

(iii) Take the logarithm of the likelihood

$$\log_e L = k - P(\delta+\gamma) - R(\alpha+\beta) + a\ln\alpha + b\ln\beta + c\ln\gamma + d\ln\delta$$

Differentiate with respect to α

$$\frac{d\log_e L}{d\alpha} = -R + \frac{a}{\alpha}$$

Set equal to zero to get estimator:

$$\frac{a}{\alpha} = R \qquad \hat{\alpha} = {}^{a}\!/_{R}.$$

Differentiate a second time:

$$\frac{d \ln L}{d^2\alpha} = -\frac{a}{\alpha^2}\,.$$

which is always negative, so that we have a maximum.

Thus $\hat{\alpha} = 230/1940 = 0.1186$

**8** (i) 11 students qualified during the period of observation, so the median is the number of sessions taken to qualify by the sixth student to qualify.

This is 9 sessions.

(ii) Define *t* as the number of sessions which have taken place since 1 Jan 2003.

Stopped studying implies recorded *after* the session number reported.

| $t_j$ | $N_j$ | $D_j$ | $C_j$ | $\frac{D_j}{N_j}$ | $1-\frac{D_j}{N_j}$ |
|---|---|---|---|---|---|
| 0 | 23 | 0 | 2 | – | 1 |
| 6 | 21 | 1 | 0 | 1/21 | 20/21 |
| 8 | 20 | 2 | 1 | 2/20 | 18/20 |
| 9 | 17 | 3 | 0 | 3/17 | 14/17 |
| 11 | 14 | 2 | 1 | 2/14 | 12/14 |
| 13 | 11 | 3 | 0 | 3/11 | 8/11 |

The Kaplan-Meier estimate is given by product of $1-\frac{D_j}{N_j}$

Then the Kaplan-Meier estimate of the survival function is

| $t$ | $\hat{S}(t)$ |
|---|---|
| $0 \le t < 6$ | 1 |
| $6 \le t < 8$ | 0.9524 |
| $8 \le t < 9$ | 0.8571 |
| $9 \le t < 11$ | 0.7059 |
| $11 \le t < 13$ | 0.6050 |
| $13 \le t < 14$ | 0.4400 |

(iii) The median time to qualify as estimated by the Kaplan-Meier estimate is the first time at which $S(t)$ is below 0.5.

Therefore the estimate is 13 sessions.

(iv) The estimate based on students qualifying during the period is a biased estimate because it does not contain information about students still studying at the end of the period, or about those who dropped out (stopped studying without qualifying).

The students still studying at the end of 2009 have (by definition) a longer period to qualification than those who qualified in the period.

Hence the Kaplan-Meier estimate is higher than the median using only students who qualified during the period.

*In part (i) the question said "determine" so some explanation of where the answer comes from was required for full credit. In part (ii) the question said "calculate" so the correct S(t) and associated ranges of t scored full marks.*

**9** (i) $h(z, t) = h_0(t) \exp(\beta z_i^T)$

$h(z,t)$ is the hazard at time $t$ (or just $h(t)$ is OK)
$h_0(t)$ is the baseline hazard

$z_i$ are covariates
$\beta$ is a vector of regression parameters

(ii) The baseline hazard refers to a female chicken in the old enclosure

(iii) The 95 per cent confidence interval for a parameter β is given by the formula

$$\beta \pm 1.96(\text{SE}[\beta]) = \beta \pm 1.96\sqrt{Var(\beta)},$$

where SE[β] is the standard error of the parameter β.

Thus, for the covariate $z_1$ =1 if Duck 0 otherwise, we have

95 per cent confidence interval =
$$-0.210 \pm 1.96\sqrt{0.002} = -0.210 \pm 0.088 = \{-0.298, -0.122\}$$

| | *95% C.I.* |
|---|---|
| $z_1$ = 1 if Duck 0 otherwise | $\beta_1$ = (–0.298, -0.122) |
| $z_2$ = 1 if Goose 0 otherwise | $\beta_2$ = (–0.049, 0.199) |
| $z_3$ = 1 if New enclosure 0 otherwise | $\beta_3$ = (0.049, 0.201) |
| $z_4$ = 1 if Male 0 otherwise | $\beta_4$ = (0.100, 0.300) |

(iv) The parameter for the new enclosure is 0.125 so the ratio of the hazard for two otherwise identical birds is exp(0.125) = 1.133.

So the hazard appears to have got worse.

The 95% confidence interval is entirely positive OR does not include zero

so at the 95% level the deterioration is statistically significant.

(v) ALTERNATIVE 1

Hazard for a Male, Goose in the Old enclosure is
$h_0(t)$ exp (0.2 + 0.075 + 0) = $h_0(t)$ exp (0.275)

Therefore the probability of still being alive in 6 months is

$$S_{\text{Goose}} = \exp\left[-\int_0^6 h_0(t)\exp(0.275)\,dt\right]$$

$$= \exp\left[-1.31653\int_0^6 h_0(t)\,dt\right]$$

This is equal to 0.9 so

$$\frac{\ln 0.9}{1.31653} = -\int_0^6 h_0(t)\,dt$$

$$\int_0^6 h_0(t)\,dt = 0.080028951$$

Hazard of a Female, Duck in the New enclosure is
$h_0(t)\exp(0-0.210+0.125) = h_0(t)\exp(-0.085)$

So, the probability she is alive after 6 months is

$$S_{\text{Duck}} = \exp\left[-\int_0^6 h_0(t)\exp(-0.085)\,dt\right]$$

$$= \exp\{-0.080028951(0.918512284)\}$$

$$= \exp\{-0.073507574\}$$

$$= 0.929129$$

So the probability she's dead is 0.07087

ALTERNATIVE 2

Hazard for a Male, Goose in the Old enclosure is
$h_0(t)$ exp $(0.2 + 0.075 + 0) = h_0(t)$ exp $(0.275)$

Therefore the probability of still being alive in 6 months is

$$S_{\text{Goose}} = \exp\left[-\int_0^6 h_0(t)\exp(0.275)\,dt\right]$$

Similarly, the probability of still being alive in 6 months for
A Female Duck in the New enclosure is

$$S_{\text{Duck}} = \exp\left[-\int_0^6 h_0(t)\exp(-0.085)\,dt\right].$$

Therefore we can write

$$\frac{S_{\text{Goose}}}{S_{\text{Duck}}} = \frac{\exp\left[-\int_0^6 h_0(t)\exp(0.275)\,dt\right]}{\exp\left[-\int_0^6 h_0(t)\exp(-0.085)\,dt\right]},$$

whence

$$\frac{\log_e S_{\text{Goose}}}{\log_e S_{\text{Duck}}} = \frac{-\int_0^6 h_0(t)\exp(0.275)\,dt}{-\int_0^6 h_0(t)\exp(-0.085)\,dt} = \frac{\exp(0.275)}{\exp(-0.085)}.$$

Hence

$$\log_e S_{\text{Duck}} = \frac{\log_e S_{\text{Goose}}[\exp(-0.085)]}{\exp(0.275)}.$$

Since $S_{\text{Goose}} = 0.9$, therefore

$$\log_e S_{\text{Duck}} = \frac{\log_e 0.9[\exp(-0.085)]}{\exp(0.275)} = -0.07351$$

So $S_{\text{Duck}} = 0.929129$

So the probability she's dead is 0.07087

**10** (i) (a) The state space is discrete (with four states: $O$ – ordinary passenger, $B$ – bronze member, $S$ – silver member and $G$ – gold member)

The probability that a passenger has a particular membership status next year depends only on their membership status in the current year (i.e. the status in previous years is not relevant).

Therefore the process is Markov.

(b) The state space is finite and therefore there is at least one stationary probability distribution.

Since any state can be reached from any other state, the Markov chain is irreducible.

Therefore the stationary probability distribution is unique.

(ii) The transition matrix $P$ is:

$$\begin{array}{c} O \\ B \\ S \\ G \end{array} \begin{pmatrix} p_0+p_1 & p_{2+} & 0 & 0 \\ p_0 & p_1 & p_{2+} & 0 \\ 0 & p_0 & p_1 & p_{2+} \\ 0 & 0 & p_0 & p_1+p_{2+} \end{pmatrix}$$

where the probability that a passenger books $i$ flights in a year is $p_i$.

(iii) Let the probability that a passenger is in state $j$ according to the stationary distribution be $\pi_j$ ($j = O, B, S, G$).

The $\pi_j$ are given by the general formula

$\pi = \pi P$.

With $p_0 = 0.4$, $p_1 = 0.4$ and $p_{2+} = 0.2$, we therefore have the equations

$$\pi_O = 0.8\pi_O + 0.4\pi_B \quad (1)$$
$$\pi_B = 0.2\pi_O + 0.4\pi_B + 0.4\pi_S \quad (2)$$
$$\pi_S = 0.2\pi_B + 0.4\pi_S + 0.4\pi_G \quad (3)$$
$$\pi_G = 0.2\pi_S + 0.6\pi_G \quad (4)$$

We also know that

$\pi_O + \pi_B + \pi_S + \pi_G = 1$.

Using equation (1) we have

$$0.2\pi_O = 0.4\pi_B$$

so that

$$\pi_O = 2\pi_B .$$

Substituting in equation (2) this yields

$$\pi_B = 0.2(2\pi_B) + 0.4\pi_B + 0.4\pi_S ,$$

so that

$$0.2\pi_B = 0.4\pi_S$$

and hence

$$\pi_S = 0.5\pi_B .$$

Finally, substituting in equation (3) yields

$$0.5\pi_B = 0.2\pi_B + 0.4(0.5\pi_B) + 0.4\pi_G ,$$

so that

$$0.1\pi_B = 0.4\pi_G$$

and hence

$$\pi_G = 0.25\pi_B .$$

We therefore have

$$2\pi_B + \pi_B + 0.5\pi_B + 0.25\pi_B = 1 ,$$

whence

$$\pi_B = \frac{1}{3.75} = \frac{4}{15} = 0.2667 ,$$

and the stationary distribution is

$$\pi_O = \frac{8}{15} = 0.5333$$

$$\pi_B = \frac{4}{15} = 0.2667$$

$$\pi_S = \frac{2}{15} = 0.1333$$

$$\pi_G = \frac{1}{15} = 0.0667$$

(iv) EITHER

The expected cost of the scheme per member per year is
$(0 \times 0.5333) + (£10 \times 0.2667) + (£20 \times 0.1333) + (£30 \times 0.0667) = £7.33$

For the scheme to be worth running, therefore, the average profit per member per year must exceed £7.33.

The profit per member is 0 for the 40% who book no flights, £10 for the 40% who book one flight, and £10$m$ for the 20% who book two or more flights, where $m$ is the average number of flights booked by those in the latter category.

For there to be a profit, we must have

$(0.4 \times 0) + (0.4 \times £10) + (0.2 \times £10m) > 7.33$

or

$4 + 2m > 7.33$

$2m > 3.33$

$m > 1.67$

This must be the case since $m$ cannot be less than 2.

Therefore the airline makes a profit on the members of the scheme.

OR

Assuming that the distribution of the number of flights taken is the same for all membership statuses, then for an Ordinary member the expected profit is

$(0.4 \times 0) + (0.4 \times 10) + (0.2 \times 20) = £8$

Similarly for the other classes of member the expected profit is

Bronze: $(0.4 \times -10) + (0.4 \times 0) + (0.2 \times 10) = £\,-2$
Silver: $(0.4 \times -20) + (0.4 \times -10) + (0.2 \times 0) = £\,-12$
Gold: $(0.4 \times -30) + (0.4 \times -20) + (0.2 \times -10) = £\,-22$

In any one year, the proportions of members in each category are given by the stationary distribution,

so the expected profit per member is

$$\frac{8}{15}(£8) + \frac{4}{15}(£-2) + \frac{2}{15}(£-12) + \frac{1}{15}(£-22) = £0.667$$

This assumes no member makes more than 2 flights per year, so is a minimum estimate of the profit.

This minimum estimate is positive, so the airline makes a profit.

**11** (i) A time inhomogeneous model should be used.

Because transition probabilities out of the "Suspended" state between times $s$ and $t$ may depend not only on the time difference $t - s$ but on the <u>the duration</u> $s$ the policy has been in that state (e.g. the probability of remaining in the suspended state for $t = 0.75$ and $s = 0.25$ is $\exp(-0.025)$), but the probability for $t = 1.25$ and $s = 0.75$ is 0.

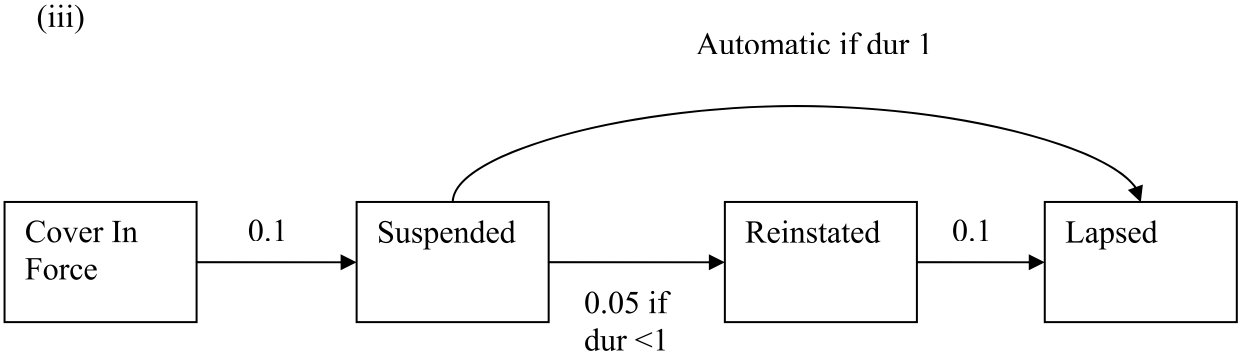
(ii) (a) A model with this state space would not satisfy the Markov property because a policy can only be reinstated once,

so if in state Cover in Force we would need to know if the policy has previously been Suspended.

(b) A Markov model could be obtained by expanding the state space to {Cover In Force, Suspended, Reinstated, Lapsed}.

In this case the future transitions will depend only on the state currently occupied and duration, irrespective of previous states.

(iii)



(iv) Labelling states as *C*, *S*, *R* and *L*.

$P_{CC}(0,t) = P_{\overline{CC}}(0,t)$ as no return to this state

$\frac{d}{dt} P_{\overline{CC}}(0,t) = -0.1 * P_{\overline{CC}}(0,t)$

$\frac{1}{P_{\overline{CC}}(0,t)} \frac{d}{dt} P_{\overline{CC}}(0,t) = \frac{d}{dt}(\ln P_{\overline{CC}}(0,t)) = -0.1$

$\ln(P_{\overline{CC}}(0,t)) = -0.1t + \text{Constant}$

$P_{CC}(0,t) = \exp(-0.1t)$ with const = 0 as $P_{CC}(0,0) = 1$

(v) To be in *S* at time *t*, must have remained in state *C* until some time *w*, then transitioned to *S* at time *w*, then remained in state *S* until *t* time.

(or express in terms of conditioning)

Probabilities are $P_{\overline{CC}}(0,w)$, 0.1*dw*, and $P_{\overline{SS}}(w,t)$ respectively.

Integrating over the possible values of *w*:

$$P_{CS}(0,t) = \int_{t-1}^{t} P_{CC}(0,w) * 0.1 * P_{SS}(w,t)dw$$

As probability of remaining in *S* if $t - w > 1$ is zero.

If $t - w < 1$

$P_{SS}(w,t) = \exp(-0.05(t-w))$

By natural extension from (iv).

Substituting

$$P_{CS}(0,t) = \int_{t-1}^{t} \exp(-0.1w) * 0.1 * \exp(-0.05(t-w))dw$$

$$P_{CS}(0,t) = 0.1\exp(-0.05t)\int_{t-1}^{t} \exp(-0.05w)dw$$

$$P_{CS}(0,t) = -2\exp(-0.05t)\left|\exp(-0.05w)\right|_{t-1}^{t}$$

$P_{CS}(0,t) = -2\exp(-0.05t)(\exp(-0.05t) - \exp(-0.05t).\exp(0.05))$

$= 2(\exp(0.05) - 1)\exp(-0.1t)$

OR

$0.1025\exp(-0.1t)$

ALTERNATIVELY

This assumes that can remain in state 'Suspended' for more than 1 time period (after which permanently suspended)

To be in $S$ at time $t$, must have remained in state $C$ until some time $w$, then transitioned to $S$ at time $w$, then remained in state $S$ until $t$ time.

*(or express in terms of conditioning)*

Probabilities are $P_{\overline{CC}}(0,w)$, 0.1dw, and $P_{\overline{SS}}(w,t)$ respectively.

Integrating over the possible values of $w$:

$$P_{CS}(0,t) = \int_0^t P_{CC}(0,w) * 0.1 * P_{SS}(w,t)dw$$

As transition probability out of state S if $t - w > 1$ is zero.

If $t - w < 1$

$P_{SS}(w,t) = \exp(-0.05(t - w))$

By natural extension from part (iv).

Splitting the integral into the parts for $t - w > 1$ and $t - w < 1$

$$P_{CS}(0,t) = \int_{t-1}^{t} \exp(-0.1w) * 0.1 * \exp(-0.05(t-w))dw + \int_0^{t-1} \exp(-0.1w) * 0.1 * \exp(-0.05((w+1) - w))dw$$

$$P_{CS}(0,t) = 0.1\exp(-0.05t)\int_{t-1}^{t} \exp(-0.05w)dw + 0.1\exp(-0.05)\int_0^{t-1} \exp(-0.1w)dw$$

$$P_{CS}(0,t) = -2\exp(-0.05t)\left|\exp(-0.05w)\right|_{t-1}^{t} - \exp(-0.05)\left|\exp(-0.1w)\right|_0^{t-1}$$

$$P_{CS}(0,t) = -2\exp(-0.05t)(\exp(-0.05t) - \exp(-0.05t).\exp(0.05)) + \exp(0.05) - \exp(0.05).\exp(-0.1t)$$

$$= (\exp(0.05) - 2)\exp(-0.1t) + \exp(-0.05)$$

*In part (iii) the label on the arrow going directly from "Suspended" to "Lapsed" is not needed, provided that the label on the arrow going from the "Suspended" to "Reinstated" indicates that the rate of 0.05 only applies if the duration is less than 1. If the label on the arrow going from "Suspended" to "Reinstated" does not indicate this, then we need an indication that movement from "Suspended" to "Lapsed" is automatic if duration = 1*

**12** (i) By reference to a standard table – appropriate if data are scanty or a table of similar lives exists.

Graphical graduation – appropriate if a "quick and dirty" result needed OR for scanty data where no other method is appropriate

By parametric formula, if the experience is large.

(ii) Standard table data

| *Age x* | *Number of survivors* | $p_x$ | $q_x$ |
|---|---|---|---|
| 50 | 32,669 | 0.99522 | 0.00478 |
| 51 | 32,513 | 0.99462 | 0.00538 |
| 52 | 32,338 | 0.99397 | 0.00603 |
| 53 | 32,143 | 0.99325 | 0.00675 |
| 54 | 31,926 | 0.99245 | 0.00755 |
| 55 | 31,685 | 0.99154 | 0.00846 |
| 56 | 31,417 | 0.99058 | 0.00942 |
| 57 | 31,121 | 0.98952 | 0.01048 |
| 58 | 30,795 | 0.98831 | 0.01169 |
| 59 | 30,435 | 0.98699 | 0.01301 |
| 60 | 30,039 | | |

Calculations:

| *Age last* | *Exposed to risk* | *Expected deaths (E)* | *Observed Deaths (O)* | *O-E* | $(O-E)^2/E$ | $\frac{(O-E)^2}{E(1-q)}$ |
|---|---|---|---|---|---|---|
| 50 | 2,381 | 11.3697 | 16 | 4.6303 | 1.8857 | 1.8948 |
| 51 | 3,177 | 17.1001 | 21 | 3.8999 | 0.8894 | 0.8942 |
| 52 | 3,460 | 20.8640 | 22 | 1.1360 | 0.0619 | 0.0622 |
| 53 | 1,955 | 13.1984 | 15 | 1.8016 | 0.2459 | 0.2476 |
| 54 | 3,122 | 23.5671 | 24 | 0.4329 | 0.0080 | 0.0080 |
| 55 | 3,485 | 29.4770 | 29 | –0.4770 | 0.0077 | 0.0078 |
| 56 | 2,781 | 26.2016 | 26 | –0.2016 | 0.0016 | 0.0016 |
| 57 | 3,150 | 32.9970 | 31 | –1.9970 | 0.1209 | 0.1221 |
| 58 | 3,651 | 42.6810 | 39 | –3.6810 | 0.3175 | 0.3212 |
| 59 | 3,991 | 51.9282 | 48 | –3.9282 | 0.2972 | 0.3011 |
| | | | | Total | 3.8356 | 3.8606 |

The null hypothesis is that the data come from a population where the mortality is that represented by the standard table.

The test statistic $\sum \frac{(O-E)^2}{E}$ is distributed $\chi^2$.

There are 10 age groups.

No degrees of freedom lost for choice of table, parameters or constraints on data.

So we use 10 degrees of freedom.

This is a one-tailed test.

The upper 5% point of the $\chi^2$ with 10 degrees of freedom is 18.31.

The observed test statistic is 3.84.

Since 3.84 < 18.31.

We have insufficient evidence to reject the null hypothesis.

(iii) ALTERNATIVE 1

(a) The data easily pass the chi squared test, but there does seem to be a gradual drift of $(O-E)$ figures from strongly positive to strongly negative. I would do a grouping of signs test to see if the data display runs or "clumps" of deviations of the same sign.

(b) $G$ = Number of groups of positive $z$s = 1

$m$ = number of deviations = 10
$n_1$ = number of positive deviations = 5
$n_2$ = number of negative deviations = 5

THEN EITHER

We want $k^*$ the largest $k$ such that

$$\sum_{t=1}^{k} \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}} < 0.05$$

The test fails at the 5% level if $G \leq k^*$.

From the Gold Book $k^* = 1$, so we reject the null hypothesis.

OR

For $t = 1$

$$\binom{n_1-1}{t-1}=\binom{4}{0}=1 \quad \text{and} \quad \binom{n_2+1}{t}=\binom{6}{1}=6 \quad \text{and} \quad \binom{m}{n_1}=\binom{10}{5}=252$$

So Pr[$t = 1$] if the null hypothesis is true is 6/252 = 0.0238, which is less than 5% so we reject the null hypothesis.

ALTERNATIVE 2

(a) The data easily pass the chi squared test, but there does seem to be a gradual drift of ($O - E$) figures from strongly positive to strongly negative. I would do a serial correlation test to see if the data displays runs or clumps" of deviations of the same sign.

(b) The calculations are shown in the table below

| $x$ | $z_x$ | $z_{x+1}$ | $A = z_x - \bar{z}$ | $B = z_{x+1} - \bar{z}$ | $AB$ | $A^2$ | $B^2$ |
|---|---|---|---|---|---|---|---|
| 50 | 1.373 | 0.943 | 0.908 | 0.570 | 0.517 | 0.824 | 0.325 |
| 51 | 0.943 | 0.249 | 0.478 | –0.125 | –0.060 | 0.228 | 0.016 |
| 52 | 0.249 | 0.496 | –0.217 | 0.123 | –0.027 | 0.047 | 0.015 |
| 53 | 0.496 | 0.089 | 0.031 | –0.284 | –0.009 | 0.001 | 0.081 |
| 54 | 0.089 | 0.088 | –0.376 | –0.286 | 0.107 | 0.142 | 0.082 |
| 55 | 0.088 | 0.039 | –0.378 | –0.334 | 0.126 | 0.143 | 0.112 |
| 56 | 0.039 | 0.348 | –0.426 | –0.026 | 0.011 | 0.181 | 0.001 |
| 57 | 0.348 | 0.563 | –0.118 | 0.190 | –0.022 | 0.014 | 0.036 |
| 58 | 0.536 | 0.545 | 0.098 | 0.172 | 0.017 | 0.010 | 0.029 |
| 59 | 0.545 | | | | | | |
| $\bar{z}$ | 0.465 | 0.373 | | Sum | 0.661 | 1.589 | 0.695 |

$0.661/(1.589*0.695)^{0.5} = 0.629$

Test $0.629\ (9^{0.5}) = 1.887$ against Normal (0,1), and, since

$0.629\ (9^{0.5}) = 1.887 > 1.645$, we reject the null hypothesis.

ALTERNATIVE 3

(a) Do the signs test to detect overall bias.

(b) Under the null hypothesis, the number of positive signs amongst the $z_x$s is distributed Binomial (10, ½ ).

We observe 5 positive signs.

The probability of obtaining 5 or more positive signs is 0.623

OR

The probability of obtaining exactly 5 positive signs is 0.246

Since this is greater than 0.025 (two-tailed test), we cannot reject the null hypothesis.

*Note that because this test is not really appropriate in a case where there are five negative and five positive deviations, no marks were awarded for part (a) to candidates who chose the Signs Test unless earlier errors meant that the number of negative and positive signs were unequal.*

ALTERNATIVE 4

(a) Do the cumulative deviations test to detect overall bias.

(b) The test statistic is $\dfrac{\sum_x \left(\theta_x - E_x \overset{o}{q}_x\right)}{\sqrt{\sum_x E_x \overset{o}{q}_x}} \sim \text{Normal}(0,1)$

| *Age x* | $\theta_x$ | $E_x \overset{o}{q}_x$ | $\theta_x - E_x \overset{o}{q}_x$ |
|---|---|---|---|
| 50 | 16 | 11.37 | 4.63 |
| 51 | 21 | 17.10 | 3.90 |
| 52 | 22 | 20.86 | 1.14 |
| 53 | 15 | 13.20 | 1.80 |
| 54 | 24 | 23.57 | 0.43 |
| 55 | 29 | 29.48 | –0.48 |
| 56 | 26 | 26.20 | –0.20 |
| 57 | 31 | 33.00 | –2.00 |
| 58 | 39 | 42.68 | –3.68 |
| 59 | 48 | 51.93 | –3.93 |
| | $\sum$ | 269.38 | 1.62 |

So the value of the test statistic is $\dfrac{1.62}{\sqrt{269.38}} = 0.09846$.

Using a 5% level of significance,
we see that $-1.96 < 0.09846 < 1.96$.

We do not reject the null hypothesis.

ALTERNATIVE 5

(a) To check for outliers we do the individual standardised deviations test.

(b) If the standard table rates were the true rates underlying the observed rates

we would expect the individual deviations to be distributed Normal (0,1)

and therefore only 1 in 20 $z_x$ s should have absolute magnitudes greater than 1.96

OR

none should lie outside the range (–3, +3)

OR

or diagram showing split of deviations actual versus expected.

Looking at the $z_x$ s we see that the largest individual deviation is 1.373.

Since this is less in absolute magnitude than 1.96 we cannot reject the null hypothesis.

*In part (ii) credit was only given for the null hypothesis if the wording used by the candidate indicates that (s)he understands that it is the mortality <u>underlying</u> the observed data that is not significantly different from that in the standard table, or that the standard table "represents" the mortality in the observed data. The null hypothesis is not that the mortality in the observed data is <u>the same as</u> that in the standard table – as it will normally not be.*

## END OF EXAMINERS' REPORT

# EXAMINATION

1 October 2010 (am)

## Subject CT4 — Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 12 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** Following a review of the results of a stochastic model run, an actuary requests that a parameter is changed. The change is not expected to alter the results significantly, but results on the final basis are required in order to complete a report. Unfortunately the actuarial student who produced the original model run is away on study leave, and so the revised run is assigned to a different student.

When the revised results are produced, they are significantly different from the original results.

Discuss possible reasons why the results are different. [3]

**2** Compare the characteristics of deterministic and stochastic models, by considering the relationship between inputs and outputs. [4]

**3** The government of a small island state intends to set up a model to analyse the mortality of the island's population over the past 50 years.

Describe the process that would be followed to carry out the analysis. [6]

**4** A large pension scheme conducts an investigation into the mortality of its younger male pensioners. The crude mortality rates are graduated using a standard table by subtracting a constant from the rates given in the table.

A trainee has been asked to test the goodness-of-fit of the proposed graduation using a chi-squared test. The trainee's workings are reproduced below:

"Test $H_0$: good fit against $H_1$: bad fit.

| *Age* | *Actual Deaths* | *Expected Deaths* | *(Actual Deaths – Expected Deaths)$^2$ /Actual Deaths* |
|---|---|---|---|
| 60 | 8 | 8.23 | 0.00661 |
| 61 | 8 | 10.01 | 0.50501 |
| 62 | 10 | 10.52 | 0.02704 |
| 63 | 12 | 14.80 | 0.65333 |
| 64 | 14 | 14.21 | 0.00315 |
| 65 | 13 | 17.37 | 1.46899 |
| Test Statistic | | | 2.66413 |

Age range is 65–60 = 5 years so 5 degrees of freedom.

Two-tailed test so take 2 * 2.66413 = 5.32826 and compare against tabulated value of chi-square distribution with 5 degrees of freedom at 2.5% level, which is 12.833.

So we accept the null hypothesis."

Identify the errors in the trainee's workings, without performing any detailed calculations. [6]

**5** (i) Write down a formula for ${}_tq_x$ ($0 \le t \le 1$) under each of the following assumptions:

(a) uniform distribution of deaths
(b) constant force of mortality
(c) the Balducci assumption

[2]

(ii) Calculate ${}_{0.5}p_{60}$ to six decimal places under each assumption given $q_{60} = 0.05$. [2]

(iii) Comment on the relative magnitude of your answers to part (ii). [2]
[Total 6]

**6** (i) Outline the circumstances under which graphical graduation of crude mortality rates might be useful. [1]

(ii) List the steps involved in graphical graduation. [5]
[Total 6]

**7** Two neighbouring small countries have for many years taken annual censuses of their populations on 1 January in which each inhabitant must give his or her age. Country *A* uses an "age last birthday" definition of age, whereas Country *B* uses an "age nearest birthday" definition. Each country has also operated a system in which deaths are recorded on an "age nearest birthday at date of death" basis.

On 30 June 2009 Country *A* invaded Country *B* and the two countries became one state. The new government wishes to estimate a single set of age-specific death rates, $\mu_x$, for the new unified state using the census data taken in the years before the invasion.

Derive a formula which the new government may use to estimate $\mu_x$ in terms of the recorded number of deaths in each country, and the population of each country recorded as being aged $x$ in the censuses. State any assumptions you make. [8]

**8** Rocky Bay is a small seaside town in the north of Europe. In a leaflet advertising the town, the tourist office has claimed that "in August, Rocky Bay has a Mediterranean climate". An actuarial student spent August 2009 on holiday in Rocky Bay with his family, and became sceptical of this claim. When he returned home, he thought it might be interesting to examine the claim by applying some of the methods he had learned while studying for the Core Technical subjects. For each of the 31 days in August 2009 he collected data recorded by various meteorological offices on the maximum temperature in Rocky Bay and the mean of the maximum temperatures reported on the same day at a range of places in the Mediterranean region.

The data are shown below, where, for each of the days in August, "+" means that Rocky Bay had the higher maximum temperature and "–" means that the Mediterranean average was higher.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | - | - | - | - | - | - | - | - | - | - | - | + | + | + | + | - | - | - | - |

| 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
|---|---|---|---|---|---|---|---|---|---|---|
| - | - | - | - | - | - | - | - | + | + | + |

(i) Carry out a statistical test to examine the tourist office's claim. [5]

(ii) Suggest reasons why the test might not be an appropriate way to examine the tourist office's claim. [2]

[Total 7]

**9** A researcher is reviewing a study published in a medical journal into survival after a certain major operation. The journal only gives the following summary information:

- the study followed 16 patients from the point of surgery
- the patients were studied until the earliest of five years after the operation, the end of the study or the withdrawal of the patient from the study
- the Nelson-Aalen estimate, $S(t)$, of the survival function was as follows:

| *Duration since operation t (years)* | *S(t)* |
|---|---|
| $0 \le t < 1$ | 1 |
| $1 \le t < 3$ | 0.9355 |
| $3 \le t < 4$ | 0.7122 |
| $4 \le t < 5$ | 0.6285 |

(i) Describe the types of censoring which are present in the study. [2]

(ii) Calculate the number of deaths which occurred, classified by duration since the operation. [6]

(iii) Calculate the number of patients who were censored. [1]

[Total 9]

**10** A study is undertaken of marriage patterns for women in a country where bigamy is not permitted. A sample of women is interviewed and asked about the start and end dates of all their marriages and where the marriages had ended, whether this was due to death or divorce (all other reasons can be ignored). The investigators are interested in estimating the rate of first marriage for all women and the rate of re-marriage among widows.

(i) Draw a diagram illustrating a multiple-state model which the investigators could use to make their estimates, using the four states: "Never married", "Married", "Widowed" and "Divorced". [1]

(ii) Derive from first principles the Kolmogorov differential equation for first marriages. [5]

(iii) Write down the likelihood of the data in terms of the waiting times in each state, the numbers of transitions of each type, and the transition intensities, assuming the transition intensities are constant. [3]

(iv) Derive the maximum likelihood estimator of the rate of first marriage. [2]
[Total 11]

**11** At a certain airport, taxis for the city centre depart from a single terminus. The taxis are all of the same make and model, and each can seat four passengers (not including the driver). The terminus is arranged so that empty taxis queue in a single line, and passengers must join the front taxi in the line. As soon as it is full, each taxi departs. A strict environmental law forbids any taxi from departing unless it is full. Taxis are so numerous that there is always at least one taxi waiting in line.

Customers arrive at the terminus according to a Poisson process with a rate $\beta$ per minute.

(i) Explain how that the number of passengers waiting in the front taxi can be modelled as a Markov jump process. [2]

(ii) Write down, for this process:

(a) the generator matrix
(b) Kolmogorov's forward equations in component form

[4]

(iii) Calculate the expected time a passenger arriving at the terminus will have to wait until his or her taxi departs. [4]

 **PLEASE TURN OVER**

The four-passenger taxis were highly polluting, and the government instituted a "scrappage" scheme whereby taxi drivers were given a subsidy to replace their old four-passenger taxis with new "greener" models. Two such models were on the market, one of which had a capacity of three passengers and the other of which had a capacity of five passengers (again, not including the driver in each case). Half the taxis were replaced with three-passenger models, and half with five-passenger models.

Assume that, after the replacement, three-passenger and five-passenger models arrive randomly at the terminus.

(iv) Write down the transition matrix of the Markov jump chain describing the number of passengers in the front taxi after the vehicle replacement. [2]

(v) Calculate the expected waiting time for a passenger arriving at the terminus after the vehicle scrappage scheme and compare this with your answer to part (iii). [3]

[Total 15]

**12** A pet shop has four glass tanks in which snakes for sale are held. The shop can stock at most four snakes at any one time because:

- if more than one snake were held in the same tank, the snakes would attempt to eat each other and
- having snakes loose in the shop would not be popular with the neighbours

The number of snakes sold by the shop each day is a random variable with the following distribution:

| *Number of Snakes Potentially Sold in Day (if stock is sufficient)* | *Probability* |
|---|---|
| None | 0.4 |
| One | 0.4 |
| Two | 0.2 |

If the shop has no snakes in stock at the end of a day, the owner contacts his snake supplier to order four more snakes. The snakes are delivered the following morning before the shop opens. The snake supplier makes a charge of $C$ for the delivery.

(i) Write down the transition matrix for the number of snakes in stock when the shop opens in a morning, given the number in stock when the shop opened the previous day. [2]

(ii) Calculate the stationary distribution for the number of snakes in stock when the shop opens, using your transition matrix in part (i). [4]

(iii) Calculate the expected long term average number of restocking orders placed by the shop owner per trading day. [2]

If a customer arrives intending to purchase a snake, and there is none in stock, the sale is lost to a rival pet shop.

(iv) Calculate the expected long term number of sales lost per trading day. [2]

The owner is unhappy about losing these sales as there is a profit on each sale of $P$. He therefore considers changing his restocking approach to place an order before he has run out of snakes. The charge for the delivery remains at $C$ irrespective of how many snakes are delivered.

(v) Evaluate the expected number of restocking orders, and number of lost sales per trading day, if the owner decides to restock if there are fewer than two snakes remaining in stock at the end of the day. [5]

(vi) Explain why restocking when two or more snakes remain in stock cannot optimise the shop's profits. [2]

The pet shop owner wishes to maximise the profit he makes on snakes.

(vii) Derive a condition in terms of $C$ and $P$ under which the owner should change from only restocking where there are no snakes in stock, to restocking when there are fewer than two snakes in stock. [2]

[Total 19]

## END OF PAPER

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS' REPORT

September 2010 examinations

## Subject CT4 — Models
## Core Technical

**Introduction**

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

T J Birse
Chairman of the Board of Examiners

December 2010

## Question 1

One or both of the runs (the original or the new) may have been incorrect as, for example, the second trainee may not have been fully aware of the set-up (for example he or she may not have followed the procedure correctly, or may have used different assumptions)

The difference between the two runs may not have only been the parameter change, for example the two runs may have used different random seeds, or the second run may have had fewer simulations.

The expectation that the model was not sensitive to this parameter could have been incorrect.

*Other valid points were given credit, for example that some parameters might be linked to live data, which will necessarily have changed; or that there may have been other amendments to the data in the meantime. However, the maximum number of marks attainable on this question was 3.*

## Question 2

A deterministic model is a model which does not contain any random components.

The output is determined once the fixed inputs and the relationships between inputs and outputs have been defined.

A stochastic model is one that recognises the random nature of the input components.

The inputs to a stochastic model are random variables, and hence for any given values of the inputs the outputs are an estimate of the characteristics of the model.

Several independent iterations of the model are required for each set of inputs to study their implications.

The output of a stochastic model gives the distribution of relevant results for a distribution of scenarios.

A deterministic model can be seen as a special case of a stochastic model.

The output of a stochastic model can be reproduced if the same random seed is used.

The output of a deterministic model is only a snap shot or an estimate of the characteristics of the model for a given set of inputs.

*Full marks could be obtained for rather less than is written above. The maximum number of marks attainable was 4 even if all the above points were made.*

## Question 3

Define the objectives of the model – what aspects of mortality are to be analysed (e.g. average mortality rates, split male/female, analysis of trends over 50 years).

Plan the model.

Establish what data are available – collect data

Evaluate the accuracy of data and the consistency of the data over time (e.g. there may have been changes to the way deaths and census data were recorded over a 50-year period)

Try to identify the main features of the mortality, and measure them.

Involve experts – e.g. there may be a national census office or Government department who can advise.

Decide between simulation package or general purpose language, or use of spreadsheet package.

Set up computer program and input data.

Debug program.

Test the output for reasonableness – is the model faithful to the actual mortality.experience of the island over the required time frame?

Check the sensitivity of model to small changes to input parameters.

Analyse the model output.

Communicate and document the results.

*This question was generally well answered, though many candidates simply reproduced the list in the Core Reading, Unit 1, pages 2 and 3, without any reference to the specific problem in the question – the analysis of mortality. These candidates did not gain full credit. A minority of candidates interpreted this question as being about a mortality investigation, making reference to the estimation of mortality rates and their subsequent graduation. Credit was given to such candidates.*

## Question 4

The null hypothesis is poorly expressed – should be "underlying rates are the graduated rates" or similar.

The test statistic is incorrect – the denominator should be expected deaths.

Cannot comment on figures in table as no access to workings.

Number of ages is 6 not 5.

However fewer than 6 degrees of freedom is appropriate because should deduct 1 for estimated parameter and some for choice of standard table

This is a one-tailed test not two-tailed.

Even if it were two-tailed, multiplying test statistic by 2 is inappropriate.

The trainee has not stated the level of significance to which he or she is working (presumably 5 per cent)

Does not explain that the reason for conclusion is 12.833 > 5.32826.

The null hypothesis should never be "accepted" rather it is "not rejected".

The trainee has not stated his or her conclusion in terms of the null hypothesis

All the graduated rates are above the crude rates so although the graduation has been accepted it is suspect.

*This question was reasonably well answered.*

## Question 5

(i) (a) ${}_tq_x = t \times q_x$

(b) EITHER ${}_tq_x = 1 - e^{-\mu t}$ OR ${}_tq_x = 1-(1-q_x)^t$

(c) $${}_tq_x = \frac{tq_x}{1-(1-t)q_x}$$

(ii) (a) ${}_{½}q_{60} = 0.025$

therefore ${}_{½}p_{60} = 0.975000$

(b) $1 - e^{-\mu} = 0.05$

so $-\mu = \ln 0.95$ and $\mu = 0.051293$

${}_{½}p_{60} = e^{-0.5\mu} = 0.974679$

(c) $${}_{\frac{1}{2}}q_{60} = \frac{\frac{1}{2}(0.05)}{1-\frac{1}{2}(0.05)}$$

$= 0.025641025$

so ${}_{½}p_{60} = 0.974359$

(iii) The Balducci assumption has the smallest value, and the uniform distribution of deaths (UDD) the largest value

<u>This is because</u> the UDD implies an increasing force of mortality over the year of age, whereas the Balducci assumption implies a decreasing force and a constant force is clearly constant.

The higher the force of mortality in the second half of the year of age relative to its magnitude in the first half of the year of age, the higher the probability of survival to age 60.5 years

The difference between the three values of ${}_{0.5}q_{60}$ is very small in this case.

*Most candidates answered the parts relating to the uniform distribution of deaths and the constant force of mortality correctly. Far fewer correctly worked out the formula for ${}_tq_x$ under the Balducci assumption. Instead, many candidates simply wrote down ${}_{1-t}q_{x+t} = (1-t)q_x$ in answer to (i)(c), which was not given credit, as it is not a formula for ${}_tq_x$ and hence is not answering the question set. However, credit was given to such candidates in (ii)(c) if they calculated the correct numerical value for ${}_{½}p_{60}$. Some candidates did not calculate the quantities in (ii) to six decimal places, and this was penalised.*

## Question 6

(i) Graphical graduation might be used when EITHER a quick visual impression OR a rough estimate is all that is required,

This is useful when the data are scanty and
EITHER
there is very little prior knowledge about the class of lives being analysed so that a suitable standard table cannot be found
OR
the experience of a professional person can be called upon

(ii) Plot the crude data,

preferably on a logarithmic scale.

If data are scanty, group ages together,

choosing evenly spaced groups and making sure there are a reasonable number of deaths (e.g. at least 5) in each group.

Plot approximate confidence limits or error bars around the plotted crude rates.

Draw the curve as smoothly as possible, trying to capture the overall shape of the crude rates.

Test the graduation for goodness-of-fit and
EITHER test for smoothness OR examine third differences

If the graduation fails the test, re-draw the curve.

"Hand polishing" individual ages may be necessary to ensure adequate smoothness.

*Many answers to this question were very sketchy and missed several of the points listed above. In (i) simply saying "when data are scanty" was not sufficient for credit, as graduation with reference to a standard table can also be used with scanty data sets provided a suitable standard table can be found. In (ii) credit was given for additional points, including noting that the curve can go outside the 95% confidence intervals at one out of every 20 or so ages, and mentioning that the analyst might want to look at obvious outliers before drawing the curve, as these may indicate data errors. A maximum of 5 marks was available for (ii).*

## Question 7

We adjust the exposed to risk to correspond to the deaths data.

Deaths are recorded on an "age nearest birthday" basis. Let the number of deaths to persons aged $x$ in countries $A$ and $B$ respectively in year $t$ be $\theta^A_{x,t}$ and $\theta^B_{x,t}$.

This means that the estimated rate $\mu_x$ will apply to exact age $x$, no further adjustment being required.

Let the populations recorded in the censuses of the two countries as being aged $x$ in the census on 1 January in year $t$ be $P^A_{x,t}$ and $P^B_{x,t}$.

A central exposed to risk for each country for year $t$ which corresponds to the deaths data is

$$E^{cA}_{x,t} = \int_{s=0}^{s=1} P*^{A}_{x,t+s}\, ds \quad \text{and} \quad E^{cB}_{x,t} = \int_{s=0}^{s=1} P*^{B}_{x,t+s}\, ds\,,$$

where $P*^{A}_{x,t+s}$ and $P*^{B}_{x,t+s}$ are the populations aged $x$ nearest birthday in countries $A$ and $B$ at time $t + s$.

This central exposed to risk can be approximated by

$$E^{cA}_{x,t} = \frac{1}{2}(P*^{A}_{x,t} + P*^{A}_{x,t+1})$$

and

$$E^{cB}_{x,t} = \frac{1}{2}(P*^{B}_{x,t} + P*^{B}_{x,t+1})\,,$$

assuming the population varies linearly between census dates.

But in country $A$ the census does not collect $P*^{A}_{x,t}$, but $P^{A}_{x,t}$, the population aged $x$ last birthday.

Assuming birthdays are evenly distributed across the calendar year, however, we can write

$$P*^{A}_{x,t} = \frac{1}{2}(P^{A}_{x,t} + P^{A}_{x-1,t})\,.$$

We also know that $P*^{B}_{x,t} = P^{B}_{x,t}$.

Therefore an exposed to risk for the two countries combined which corresponds to the deaths data is

$$E^{cA}_{x,t} + E^{cB}_{x,t} = \frac{1}{2}(P*^{A}_{x,t} + P*^{A}_{x,t+1}) + \frac{1}{2}(P*^{B}_{x,t} + P*^{B}_{x,t+1})$$

$$= \frac{1}{2}\left(\frac{1}{2}(P^{A}_{x,t} + P^{A}_{x-1,t}) + \frac{1}{2}(P^{A}_{x,t+1} + P^{A}_{x-1,t+1})\right) + \frac{1}{2}(P^{B}_{x,t} + P^{B}_{x,t+1})\,,$$

and hence the combined age specific death rate can be estimated as

$$\mu_x = \frac{\theta^{A}_{x,t} + \theta^{B}_{x,t}}{\frac{1}{4}P^{A}_{x,t} + \frac{1}{4}P^{A}_{x-1,t} + \frac{1}{4}P^{A}_{x,t+1} + \frac{1}{4}P^{A}_{x-1,t+1} + \frac{1}{2}P^{B}_{x,t} + \frac{1}{2}P^{B}_{x,t+1}}\,.$$

*The solution above assumes that the estimates of $\mu_x$ are to be made using a single calendar year t. Additional credit was given to candidates who stated that the appropriate time over which the estimates are to be made should be defined at the outset, and that if this period is longer than one year the deaths and exposed to risk for all relevant calendar years should be summed and the total deaths divided by the total exposed to risk (subject to a maximum of 8 marks being available). The Examiners were looking for understanding of the process that must be gone through in order to obtain the required estimates. Answers consisting mainly of "disembodied" statements, without a coherent argument received limited credit.*

## Question 8

(i) The null hypothesis, $H_0$, is that the climate – or the underlying (long-run average) temperature – in Rocky Bay in August is the same as that in the Mediterranean.

EITHER

**Signs Test**

Let $P$ be the number of days for which the maximum temperature in Rocky Bay is greater than that expected in the Mediterranean.

Under $H_0$, $P \sim \text{Binomial}(31, 0.5)$.

THEN EITHER NORMAL APPROXIMATION

Using the Normal approximation as we have more than 20 days,

$$P \sim \text{Normal}\left(\frac{31}{2}, \frac{31}{4}\right)$$

In the observations $P = 7$,

The value of the test statistic is therefore

$$Z = \frac{7 - 15.5}{\sqrt{7.75}} = -3.05$$

Since $|Z| > 1.96$ we reject $H_0$ at the 5 per cent level of significance

OR EXACT CALCULATION

We have $P = 7$

The probability of obtaining 7 or fewer positive signs is

$$\binom{31}{7}0.5^{31}+\binom{31}{6}0.5^{31}+\ldots+\binom{31}{0}0.5^{31}$$

which is 0.00122 + 0.00034 + 0.00008 + ... + 0.00000 = 0.00166

since this is less than 0.025 (two-tailed test)

we reject $H_0$ at the 5 per cent level of significance

and conclude that the climate of Rocky Bay is not the same as that in the Mediterranean.
OR

**Grouping of Signs Test**

Let $P$ be the number of days for which the maximum temperature in Rocky Bay is greater than that expected in the Mediterranean.

Let $Q$ (= 31 – $P$) be the number of days for which the maximum temperature in Rocky Bay is less than that expected in the Mediterranean.

To test the null hypothesis, we need to calculate the maximum number

of positive runs, $g$, for which $\sum_{t=1}^{g}\frac{\binom{P-1}{t-1}\binom{Q+1}{t}}{\binom{P+Q}{P}}<0.05$.

since $P$ = 7 and $Q$ = 24,

THEN EITHER

using the table on p. 189 of the *Formulae and Tables for Examinations*,

we find that $g$ = 3.

OR

using the normal approximation we have

$G$ ~ Normal(5.64, 0.95),

so, using a one-tailed test, the critical value at the 5% level is 5.645 – 1.645*√0.947 = 4.04.

Since we only have 2 positive runs in the data we reject $H_0$ at the 5 per cent level of significance and conclude that the climate of Rocky Bay is not the same as that in the Mediterranean.

(ii) Runs of consecutive days with the same sign are likely since the weather tends to be determined by atmospheric conditions lasting more than one day.

The Mediterranean averages are averages for the month of August 2009, not long-run averages.

August 2009 might have been an unusually hot month in the Mediterranean region.

Maximum temperature is not the only measure of climate, also consider mean temperature, hours of sunshine, windiness, etc.

Choice of locations used for Mediterranean data could be important.

Also tests just look at whether one is higher or lower – the difference in each case could be negligible (e.g. 25.001 degrees vs 25.002 degrees)

A non-standard measurement method might have been used in Rocky Bay, which confounds the comparison.

*For the signs test the continuity correction was not required, but if done has to be correct. Candidates were given credit for a one-sided signs test in (i) provided that they set the null hypothesis up correctly – i.e. that the average maximum temperature in Rocky Bay in August is **no lower** than that in the Mediterranean. In (ii) other sensible comments were given credit, and the maximum score of 2 marks could be obtained for making four sensible points – not all the points listed above were required.*

## Question 9

(i) Type I (right censoring) of patients who survive to duration 5 years.

Random censoring of patients who withdraw from the study.

(ii) Since $S(t) = \exp(-\Lambda_t)$ where $\Lambda_t = \sum_{t_j \le t} \left( \frac{d_j}{n_j} \right)$

$\Lambda_t = -\ln [S(t)]$

So

| *Duration since operation t (years)* | *S(t)* | $\Lambda_t$ |
|---|---|---|
| $0 \le t < 1$ | 1 | 0 |
| $1 \le t < 3$ | 0.9355 | 0.0667 |
| $3 \le t < 4$ | 0.7122 | 0.3394 |
| $4 \le t < 5$ | 0.6285 | 0.4644 |

Let $d_j$ and $n_j$ be the number of deaths and the number in the risk set at the $j$th point at which events occur.

Consider $t = 1$

$$\frac{d_1}{n_1} = 0.0667$$

Since there can be no more than 16 patients at risk at $t = 1$, the only possible combination is $d_1 = 1$ and $n_1 = 15$

Consider $t = 3$ (the second point at which events occur)

$$\frac{d_2}{n_2} = 0.3394 - 0.0667 = 0.2727$$

Recognising this as 3/11, and that there are at most 14 patients at risk, this implies that $d_2 = 3$ and $n_2 = 11$.

Consider $t = 4$ (the third point at which events occur)

$$\frac{d_3}{n_3} = 0.4644\text{-}0.0667 - 0.2727 = 0.125$$

Recognising this as 1/8, and that there are at most 11 patients at risk, this implies that $d_3 = 1$ and $n_3 = 8$.

So the answer is:

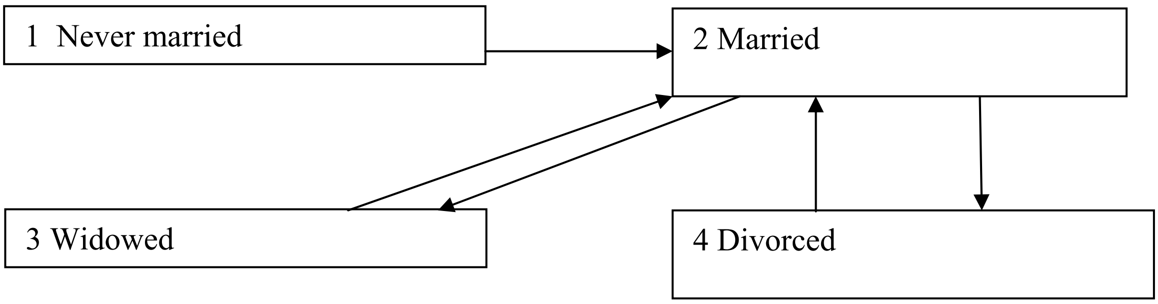
1 death at duration 1 year
3 deaths at duration 3 years
1 death at duration 4 years

(iii) Patients either die or are censored. As the total number of patients is 16 and 5 die the number censored is 16 – 5 = 11.

*This was the best answered question on the examination paper. In (i) "right" censoring was awarded credit, as was some explanation of whether the censoring was informative or non-informative. In (ii) a common error was to state the durations as ranges (i.e. 1 death at durations between 1 and 3 years, 3 deaths at durations between 3 and 4 years, and 1 death at durations over 4 years). This reveals a misunderstanding of the estimator, and was penalised by the loss of 1 mark. Candidates who calculated an incorrect number of deaths in (ii) were given credit for (iii) if their answer to (iii) was consistent with their answer to (ii).*

## Question 10

(i)




(ii) Using the numbering of the states above, let the probability that a women who is in state *i* at time *x* will be in state *j* at time *x* + *t* be ${}_t p_x^{ij}$.

Using the Markov property, and conditioning on the state occupied at time *x* + *t*,

and noting that for first marriages return from the widowed or divorced state is not possible, we can write

$${}_{t+dt} p_x^{12} = {}_t p_x^{11} \, {}_{dt} p_{x+t}^{12} + {}_t p_x^{12} \, {}_{dt} p_{x+t}^{22}$$

Using the law of total probability, ${}_{dt} p_{x+t}^{22} = 1 - {}_{dt} p_{x+t}^{23} - {}_{dt} p_{x+t}^{24}$,

so that

$${}_{t+dt} p_x^{12} = {}_t p_x^{11} \, {}_{dt} p_{x+t}^{12} + {}_t p_x^{12} (1 - {}_{dt} p_{x+t}^{23} - {}_{dt} p_{x+t}^{24})$$

Let the transition rate from state *i* to state *j* at time *x*+*t* be $\mu_{x+t}^{ij}$.

Assume that ${}_{dt} p_{x+t}^{ij} = \mu_{x+t}^{ij} dt + o(dt)$, $i \neq j$

where $\lim_{dt \to 0^+} \frac{o(dt)}{dt} = 0$.

Substituting for the ${}_t p_x^{ij}$ in the equation above produces

$${}_{t+dt} p_x^{12} = {}_t p_x^{11} \mu_{x+t}^{12} dt + {}_t p_x^{12} (1 - \mu_{x+t}^{23} dt - \mu_{x+t}^{24} dt) + o(dt)$$

Therefore

$${}_{t+dt} p_x^{12} - {}_t p_x^{12} = {}_t p_x^{11} \mu_{x+t}^{12} dt - {}_t p_x^{12} \mu_{x+t}^{23} dt - {}_t p_x^{12} \mu_{x+t}^{24} dt + o(dt)$$

and, taking limits, we have

$$\lim_{dt \to 0^+} \frac{{}_{t+dt} p_x^{12} - {}_t p_x^{12}}{dt} = {}_t p_x^{11} \mu_{x+t}^{12} - {}_t p_x^{12} \mu_{x+t}^{23} - {}_t p_x^{12} \mu_{x+t}^{24}$$

So

$$\frac{d}{dt} {}_t p_x^{12} = {}_t p_x^{11} \mu_{x+t}^{12} - {}_t p_x^{12} \mu_{x+t}^{23} - {}_t p_x^{12} \mu_{x+t}^{24}$$

(iii) Let the waiting time in state $i$ be $v_i$,

and the number of transitions from state $i$ to state $j$ be $d_{ij}$

and the transition intensity from state $i$ to state $j$ *is* $\mu_{ij}$

Then the likelihood, $L$, may be written

$$L = K \exp[-v_1\mu_{12} - v_2(\mu_{23} + \mu_{24}) - v_3\mu_{32} - v_4\mu_{42}] \mu_{12}^{d_{12}} \mu_{23}^{d_{23}} \mu_{24}^{d_{24}} \mu_{32}^{d_{32}} \mu_{42}^{d_{42}}.$$

(iv) The logarithm of the likelihood is

$$\log_e L = \log_e K - v_1\mu_{12} - v_2(\mu_{23} + \mu_{24}) - v_3\mu_{32} - v_4\mu_{42}$$
$$+ d_{12} \log_e \mu_{12} + d_{23} \log_e \mu_{23} + d_{24} \log_e \mu_{24} + d_{32} \log_e \mu_{32} + d_{42} \log_e \mu_{42}$$

Differentiating with respect to $\mu_{12}$ gives

$$\frac{\partial L}{\partial \mu_{12}} = -v_1 + \frac{d_{12}}{\mu_{12}}.$$

Setting this equal to 0 and solving for $\mu_{12}$ gives

$$\hat{\mu}_{12} = \frac{d_{12}}{v_1}.$$

This is a maximum because $\dfrac{\partial^2 L}{\partial \mu_{12}^2} = -\dfrac{d_{12}}{\mu_{12}^2}$ which is negative.

*Answers to (i), (iii) and (iv) were generally good. In (i) the arrows from Widowed to Married and Divorced to Married were not required for full marks, as the question is about first marriages. Answers to (ii) were more disappointing, with many candidates omitting steps in the argument. In (ii), some candidates included the extra terms* $+{}_t p_x^{13} \mu_{x+t}^{32} + {}_t p_x^{14} \mu_{x+t}^{42}$. *Since it is just about possible to interpret the question in a way such that these should be included, this was not heavily penalised.*

## Question 11

(i) A Markov jump process is a continuous-time Markov process with a discrete state space.

For a process to be Markov, the future development of the process must depend only on its current state.

This is the case here, as the future of the process depends only on the number of passengers currently in the front taxi.

The number of passengers in the front taxi also has a discrete state space {0, 1, 2, 3}. (Note that immediately a fourth passenger arrives the taxi will depart so the front taxi in the queue will never have four passengers in it.)

(ii) (a) The generator matrix $A$ is

$$\begin{pmatrix} -\beta & \beta & 0 & 0 \\ 0 & -\beta & \beta & 0 \\ 0 & 0 & -\beta & \beta \\ \beta & 0 & 0 & -\beta \end{pmatrix}$$

(b) Kolmogorov's forward equations can be written in compact form as

$$\frac{d}{dt} P(t) = P(t)A,$$

Which are, for $j = 0$

$$\frac{d}{dt} p_{i0}(t) = \beta p_{i3}(t) - \beta p_{i0}(t)$$

and, for $j$ = 1,2,3

$$\frac{d}{dt} p_{ij}(t) = \beta p_{i,j-1}(t) - \beta p_{ij}(t) .$$

(iii) Since the waiting times under a Poisson process are exponential the expected waiting time between the arrival of passengers at the terminus is $\frac{1}{\beta}$ minutes.

Successive waiting times are independent, therefore the expected waiting time for a passenger arriving at the terminus is

$$E[t] = \sum_{i=0}^{3} p_i \frac{3-i}{\beta},$$

where $p_i$ is the probability that the front taxi has exactly $i$ previous passengers waiting in it when the passenger arrives.

Since the $p_i$s are all equal for $i$ = 0, 1, 2, 3

$$E[t] = 0.25\left(\frac{3}{\beta} + \frac{2}{\beta} + \frac{1}{\beta} + \frac{0}{\beta}\right) = \frac{3}{2\beta} \text{ minutes.}$$

(iv) The transition matrix, $P$, is

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

(v) The expected waiting time if the front taxi is a three-passenger model is

$$E[t \mid 3 - \text{passenger model}] = \sum_{i=0}^{2} p_i \frac{2-i}{\beta} = \frac{1}{3}\left(\frac{2}{\beta} + \frac{1}{\beta} + \frac{0}{\beta}\right) = \frac{1}{\beta}$$

The expected waiting time if the front taxi is a five-passenger model is

$$E[t \mid 5 - \text{passenger model}] = \sum_{i=0}^{4} p_i \frac{4-i}{\beta} = \frac{1}{5}\left(\frac{4}{\beta} + \frac{3}{\beta} + \frac{2}{\beta} + \frac{1}{\beta} + \frac{0}{\beta}\right) = \frac{2}{\beta}.$$

But 5-passenger models must expect to wait $\frac{5}{3}$ times as long at the front of the queue than do 3-passenger models.

So when a passenger arrives at the terminus, $\frac{5}{8}$ of the time the taxi at the front of the queue will be a five-passenger model and only $\frac{3}{8}$ of the time will is be a three-passenger model.

So the overall expected waiting time in minutes is

$$\frac{3}{8}(E[t\,|\,3-\text{passenger model}])+\frac{5}{8}(E[t\,|\,5-\text{passenger model}])=\frac{13}{8\beta}.$$

As this is longer than $\frac{3}{2\beta}$, the service provided to the passengers has deteriorated.

*Many candidates struggled with this question. A common error in (ii) was to draw a matrix with five states rather than four, failing to recognise that taxis with four passengers in do not wait at the front of the queue, but depart as soon as the fourth passenger arrives. Most candidates who attempted (iv) wrote down a generator matrix, whereas the question asked for a transition matrix.*

## Question 12

(i)

| *Start previous day* | *Start morning* 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0.4 | 0 | 0 | 0.6 |
| 2 | 0.4 | 0.4 | 0 | 0.2 |
| 3 | 0.2 | 0.4 | 0.4 | 0 |
| 4 | 0 | 0.2 | 0.4 | 0.4 |

(ii) If stationary distribution is $\underline{\pi} = \left(\pi_1 \quad \pi_2 \quad \pi_3 \quad \pi_4\right)$

Then $\underline{\pi}A = \underline{\pi}$ where A is the matrix in (i)

$0.4\pi_1 + 0.4\pi_2 + 0.2\pi_3 = \pi_1$ (a)
$0.4\pi_2 + 0.4\pi_3 + 0.2\pi_4 = \pi_2$ (b)
$0.4\pi_3 + 0.4\pi_4 = \pi_3$ (c)
$0.6\pi_1 + 0.2\pi_2 + 0.4\pi_4 = \pi_4$ (d)

From (c) $\pi_3 = 0.666666\pi_4$
From (b) $\pi_2 = 0.7778\pi_4$
From (a) $\pi_1 = 0.7407\pi_4$

$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1 = (0.7407 + 0.7778 + 0.6666 + 1)\pi_4$

Implies $\pi_1 = 0.2325$, $\pi_2 = 0.2442$, $\pi_3 = 0.2093$, $\pi_4 = 0.31395$

OR $\pi_1 = \dfrac{10}{43}$, $\pi_2 = \dfrac{21}{86}$, $\pi_3 = \dfrac{9}{43}$, $\pi_4 = \dfrac{27}{86}$.

(iii) Probability of restocking is 0.6 if in $\pi_1$ and 0.2 if in $\pi_2$

So long term rate = 0.6 * 0.2325 + 0.2 * 0.2442 = 0.1884 per trading day

(iv) Probability of losing a sale is 0.2 if in $\pi_1$

So expected lost sales per day = 0.2 * 0.2325 = 0.0465

(v) If restock when fewer than two in stock then transition matrix changes to:

| *Start previous day* | *Start morning* 2 | 3 | 4 |
|---|---|---|---|
| *2* | *0.4* | *0* | *0.6* |
| 3 | 0.4 | 0.4 | 0.2 |
| 4 | 0.2 | 0.4 | 0.4 |

Label stationary distribution $\underline{\lambda}$. Then

$0.4\lambda_2 + 0.4\lambda_3 + 0.2\lambda_4 = \lambda_2$ (b1)

$0.4\lambda_3 + 0.4\lambda_4 = \lambda_3$ (c1)

$0.6\lambda_2 + 0.2\lambda_3 + 0.4\lambda_4 = \lambda_4$ (d1)

From (c1) $\lambda_3 = 0.666666\lambda_4$

From (b1) $\lambda_2 = 0.7778\lambda_4$

$\lambda_2 = 0.3182$ OR 7/22

$\lambda_3 = 0.2727$ OR 3/11

$$\lambda_4 = \frac{1}{(1+2/3+7/9)} = 0.4091 \quad \text{OR} \quad 9/22$$

As no more than two snakes sell per day,

there are no lost sales.

Probability of restocking 0.6 if in $\lambda_2$ and 0.2 in $\lambda_3$ = 0.2455

(vi) Restocking at two or more snakes would not result in fewer lost sales than restocking at 1.

Because the probability of selling more than 2 snakes is zero.

It would, however, result in more restocking charges than restocking at 1.

Therefore it must result in lower profits than restocking at 1 so is not optimal.

(vii) Costs if restock at 0

$0.1884C + 0.0465P$

Costs if restock at 1

$0.24546C$

So should change restocking approach if

$0.24546C < 0.1884C + 0.0465P$

$C < 0.8148P$

*In this question, many candidates answered (i) and (ii), but made no further progress. Candidates who wrote down the wrong matrix in (i) but evaluated the stationary distribution correctly for the matrix they had written down were given full credit for (ii), and gained credit in (iii) and (iv) if these parts were answered correctly given the matrix which had been written down in (i). A common error was to write down a five-state model in (i). Few candidates attempted the later sections of this question.*

## END OF EXAMINERS' REPORT

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINATION

15 April 2011 (am)

## Subject CT4 — Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 12 questions, beginning your answer to each question on a separate sheet.*

5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** Give three advantages of the two-state model over the Binomial model for estimating transition intensities where exact dates of entry into and exit from observation are known. [3]

**2** Distinguish between the conditions under which a Markov chain:

(a) has at least one stationary distribution.
(b) has a unique stationary distribution.
(c) converges to a unique stationary distribution. [3]

**3** Describe the ways in which the design of a model used to project over only a short time frame may differ from one used to project over fifty years. [4]

**4** Children at a school are given weekly grade sheets, in which their effort is graded in four levels: 1 "Poor", 2 "Satisfactory", 3 "Good" and 4 "Excellent". Subject to a maximum level of Excellent and a minimum level of Poor, between each week and the next, a child has:

- a 20 per cent chance of moving up one level.
- a 20 per cent chance of moving down one level.
- a 10 per cent chance of moving up two levels.
- a 10 per cent chance of moving down two levels.

Moving up or down three levels in a single week is not possible.

(i) Write down the transition matrix of this process. [2]

Children are graded on Friday afternoon in each week. On Friday of the first week of the school year, as there is little evidence on which to base an assessment, all children are graded "Satisfactory".

(ii) Calculate the probability distribution of the process after the grading on Friday of the third week of the school year. [3]
[Total 5]

**5** (i) Explain why a mortality experience would need to be graduated. [3]

An actuary has conducted investigations into the mortality of the following classes of lives:

(a) the female members of a medium-sized pension scheme

(b) the male population of a large industrial country

(c) the population of a particular species of reptile in the zoological collections of the southern hemisphere

The actuary wishes to graduate the crude rates.

(ii) State an appropriate method of graduation for each of the three classes of lives and, for each class, briefly explain your choice. [3]
[Total 6]

**6** A study of the mortality of a certain species of insect reveals that for the first 30 days of life, the insects are subject to a constant force of mortality of 0.05. After 30 days, the force of mortality increases according to the formula:

$$\mu_{30+x} = 0.05\exp(0.01x),$$

where $x$ is the number of days after day 30.

(i) Calculate the probability that a newly born insect will survive for at least 10 days. [1]

(ii) Calculate the probability that an insect aged 10 days will survive for at least a further 30 days. [3]

(iii) Calculate the age in days by which 90 per cent of insects are expected to have died. [4]
[Total 8]

**7** (i) Define a counting process. [2]

For each of the following processes:

- simple random walk
- compound Poisson
- Markov chain

(ii) (a) State whether each of the state space and the time set is discrete, continuous or can be either.

(b) Give an example of an application which may be useful to a shopkeeper selling dried fruit and nuts loose.

[6]
[Total 8]

**8** (i) Explain the difference between the central and the initial exposed to risk, in the context of mortality investigations. [2]

An investigation studied the mortality of infants aged under 1 year. The following table gives details of 10 lives involved in the investigation. Infants with no date of death given were still alive on their first birthday.

| *Life* | *Date of birth* | *Date of death* |
|---|---|---|
| 1 | 1 August 2008 | - |
| 2 | 1 September 2008 | - |
| 3 | 1 December 2008 | 1 February 2009 |
| 4 | 1 January 2009 | - |
| 5 | 1 February 2009 | - |
| 6 | 1 March 2009 | 1 December 2009 |
| 7 | 1 June 2009 | - |
| 8 | 1 July 2009 | - |
| 9 | 1 September 2009 | - |
| 10 | 1 November 2009 | 1 December 2009 |

(ii) Calculate the maximum likelihood estimate of the force of mortality, using a two-state model and assuming that the force is constant. [3]

(iii) Hence estimate the infant mortality rate, $q_0$. [1]

(iv) Estimate the infant mortality rate, $q_0$, using the initial exposed to risk. [1]

(v) Explain the difference between the two estimates. [2]

[Total 9]

**9** (i) Define a Markov jump process. [2]

A study of a tropical disease used a three-state Markov process model with states:

1. Not suffering from the disease
2. Suffering from the disease
3. Dead

The disease can be fatal, but most sufferers recover. Let ${}_tp_x^{ij}$ be the probability that a person in state *i* at age *x* is in state *j* at age *x*+*t*. Let $\mu_{x+t}^{ij}$ be the transition intensity from state *i* to state *j* at age *x*+*t*.

(ii) Show from first principles that:

$$\frac{d}{dt}\,{}_tp_x^{13} = {}_tp_x^{11}\,\mu_{x+t}^{13} + {}_tp_x^{12}\,\mu_{x+t}^{23}.$$ [4]

The study revealed that sufferers who contract the disease a second or subsequent time are more likely to die, and less likely to recover, than first-time sufferers.

(iii) Draw a diagram showing the states and possible transitions of a model which allows for this effect yet retains the Markov property. [3]

[Total 9]

**10** At Miracle Cure hospital a pioneering new surgery was tested to replace human lungs with synthetic implants. Operations were carried out throughout June 2010. Patients who underwent the surgery were monitored daily until the end of August 2010, or until they died or left hospital if sooner. The results are shown below. Where no date is given, the patient was alive and still in hospital at the end of August.

| *Patient* | *Date of surgery* | *Date of leaving observation* | *Reason for leaving observation* |
|---|---|---|---|
| A | June 1 | June 3 | Died |
| B | June 3 | July 2 | Left Hospital |
| C | June 5 | | |
| D | June 8 | | |
| E | June 9 | July 11 | Died |
| F | June 12 | | |
| G | June 16 | June 21 | Died |
| H | June 17 | Aug 12 | Left Hospital |
| I | June 22 | | |
| J | June 24 | June 29 | Died |
| K | June 25 | Aug 20 | Died |
| L | June 26 | | |
| M | June 29 | Aug 6 | Left Hospital |
| N | June 30 | | |

(i) Explain whether each of the following types of censoring is present and for those present explain where they occur:

- right censoring
- left censoring
- informative censoring

[3]

(ii) Calculate the Kaplan-Meier estimate of the survival function for these patients, stating all assumptions that you make. [6]

(iii) Sketch, on a suitably labelled graph, the Kaplan-Meier estimate of the survival function. [2]

(iv) Estimate the probability that a patient will die within four weeks of surgery. [1]

[Total 12]

**11** An historian has investigated the force of mortality from tuberculosis in a particular town in a developed country in the 1860s using a sample of records from a cemetery. He wishes to test whether the underlying mortality from tuberculosis in the town is the same as the national force of mortality from this cause of death, as reported in death registration data. The data are shown in the table below.

| *Age-group* | *Deaths in sample* | *Central exposed to risk in sample* | *National force of mortality* |
|---|---|---|---|
| 5–14 | 13 | 3,685 | 0.0051 |
| 15–24 | 47 | 2,540 | 0.0199 |
| 25–34 | 52 | 1,938 | 0.0309 |
| 35–44 | 50 | 1,687 | 0.0316 |
| 45–54 | 33 | 1,386 | 0.0286 |
| 55–64 | 23 | 1,018 | 0.0230 |
| 65–74 | 13 | 663 | 0.0202 |
| 75–84 | 3 | 260 | 0.0070 |

(i) Carry out an overall test of the null hypothesis that the underlying mortality from tuberculosis in the town is the same as the national force of mortality, and state your conclusion. [6]

(ii) (a) Identify two differences between the experience of the sample and the national experience which the test you performed in (i) might not detect.

(b) Carry out a test for each of the differences in (ii)(a). [7]

(iii) Comment on the results from all the tests carried out in (i) and (ii). [1]

[Total 14]

**12** Farmer Giles makes hay each year and he makes far more than he could possibly store and use himself, but he does not always sell it all. He has decided to offer incentives for people to buy large quantities so it does not sit in his field deteriorating. He has devised the following “discount” scheme.

He has a Base price, *B* of £8 per bale. Then he has three levels of discount: Good price, *G*, is a 10% discount, Loyalty price, *L* is a 20% discount and Super price, *S*, is a 25% discount on the Base price.

- Customers who increase their order compared with last year move to one higher discount level, or remain at level *S*.
- Customers who maintain their order from last year stay at the same discount level.
- Customers who reduce their order from last year drop one level of discount or remain at level *B* provided that they maintained or increased their order the previous year.

- Customers who reduce their order from last year drop two levels of discount if they also reduced their order last year, subject to remaining at the lowest level *B*.

(i) Explain why a process with the state space of {*B*, *G*, *L*, *S*} does not display the Markov property. [2]

(ii) (a) Define any additional state(s) required to model the system with the Markov property.

(b) Construct a transition graph of this Markov process clearly labelling all the states. [3]

Farmer Giles thinks that each year customers have a 60% likelihood of increasing their order and a 30% likelihood of reducing it, irrespective of the discount level they are currently in.

(iii) (a) Write down the transition matrix for the Markov process.

(b) Calculate the stationary distribution.

(c) Hence calculate the long run average price he will get for each bale of hay.

[8]

(iv) Calculate the probability that a customer who is currently paying the Loyalty price, *L*, will be paying *L* in two years' time. [3]

(v) Suggest reasons why the assumptions Farmer Giles has made about his customers' behaviour may not be valid. [3]

[Total 19]

## END OF PAPER

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS' REPORT

April 2011 examinations

## Subject CT4 — Models
## Core Technical

**Introduction**

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

T J Birse
Chairman of the Board of Examiners

July 2011

## Question 1

We can calculate the maximum likelihood estimate (MLE) of the transition intensities directly using the two-state model, whereas the Binomial model requires additional assumptions.

The variance of the Binomial estimate is greater than that of the estimate from the two-state model (though the difference is tiny unless the transition intensities are large).

The MLE in the two-state model is consistent and unbiased, whereas the Binomial estimate is only consistent and unbiased if lives are observed for exactly one year, which is rarely the case.

The two-state model is easily extended to encompass increments and additional decrements, whereas the Binomial model is not.

The two-state model uses the exact times of the transitions, whereas the Binomial model only uses the number of transitions.

*This question was poorly answered by many candidates, despite being straightforward bookwork. Many candidates commented that the two-state model and the Binomial model make different assumptions about the shape of the force of mortality within the year of age. This was only be given credit if candidates also explained why the multiple state model's assumption is BETTER than the Binomial model's assumption (which it might be, for example, at younger ages).*

*Full marks could be obtained for giving three reasons. It was not necessary to give all the points listed above in order to obtain full marks.*

## Question 2

(a) A Markov chain with a finite state space has at least one stationary probability distribution.

(b) An irreducible Markov chain with a finite state space has a unique stationary probability distribution.

(c) A Markov chain with a finite state space which is irreducible, and which is also aperiodic converges to a unique stationary probability distribution.

*Many candidates scored full marks on this question. The question asked candidates to "distinguish". Therefore for full credit it is important that candidates did, indeed, understand and make the relevant distinction.*

## Question 3

Individual variables may behave differently, for example a model over 50 years may be more sensitive to differences in the input values of certain variables than one over the short term.

A variable which has an ignorable effect in the short term may have a non-ignorable effect over 50 years.

Over the short term, it may be reasonable to assume the values of some variables to be constant or to vary linearly, whereas this would not be reasonable over 50 years. For example, growth which is exponential may appear linear if studied over a short time frame.

The interaction between variables in the short-term may be different from that over the long-term.

Higher order relationships between variables may be ignored for simplicity if modelling over a short time frame.

The time units used in the model might be shorter for a model projecting over a short time frame, so that the total number of time units used in each model is roughly the same.

Over 50 years, regulatory changes and other "shock" events are more likely to occur, and the model design may need to consider the circumstances in which the results or conclusions may be materially impacted (e.g. in the short term the tax basis may be known, but in the long run it is likely to change).

*The marks on this question were the lowest on any question. The question was a "higher skills" question and so required candidates to think about the context. Little credit was given to candidates who uncritically reproduced sections of the Core Reading. In particular, the question is about model DESIGN, so the points made should relate to the design of the model.*

## Question 4

(i) $$\begin{pmatrix} 0.7 & 0.2 & 0.1 & 0 \\ 0.3 & 0.4 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0 & 0.1 & 0.2 & 0.7 \end{pmatrix}$$

(ii) If the probability distribution in the first week is $\Pi$, and the transition matrix is *M*, then the probability distribution at the end of the third week is

$$\Pi M^2 = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.7 & 0.2 & 0.1 & 0 \\ 0.3 & 0.4 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0 & 0.1 & 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} 0.7 & 0.2 & 0.1 & 0 \\ 0.3 & 0.4 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0 & 0.1 & 0.2 & 0.7 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.56 & 0.24 & 0.15 & 0.05 \\ 0.35 & 0.27 & 0.21 & 0.17 \\ 0.17 & 0.21 & 0.27 & 0.35 \\ 0.05 & 0.15 & 0.24 & 0.56 \end{pmatrix}$$

so that there is a probability of

35% that a child will be graded Poor',
27% that a child will be graded Satisfactory,
21% that a child will be graded Good and
17% that a child will be graded Excellent..

*There were two common errors on this question. The first was to assume that if a child could not move up or down two levels, he or she would not move at all. The phrase in the question "[s]ubject to a maximum level of Excellent and a minimum level of Poor" was intended to indicate that children could not move beyond these limits in either direction, but would move as far as they could. Thus a child at level "Good", who had a 20% chance of moving up one level and a 10% chance of moving up two levels, would have a 30% chance of moving to level Excellent, as the 10% who would have moved up two levels will only be able to move up one level. The second error was to use $\Pi M^3$ in part (ii). Candidates who made the first error were penalised in part (i) but could gain full credit for part (ii) if they followed through correctly.*

## Question 5

(i) We believe that mortality varies smoothly with age (and evidence from large experiences supports this belief).

Therefore the crude estimate of mortality at any age carries information about mortality at adjacent ages.

By smoothing the experience, <u>we can make use of data at adjacent ages</u> to improve the estimates at each age.

This reduces sampling (or random) errors.

The mortality experience may be used in financial calculations.

Irregularities, jumps and anomalies in financial quantities (such as premiums for life insurance contracts) are hard to justify to customers.

(ii) (a) Female members of a medium-sized pension scheme.

With reference to a standard table, because there are many extant tables dealing with female pensioners.

(b) Male population of a large industrial country.

By parametric formula, because the experience is large.
OR
because the graduated rates may form a new standard table for the country.

(c) Population of a particular species of reptile in the zoological collections of the southern hemisphere.

Graphical, because no suitable standard table is likely to exist and the experience is small.

*This question was well answered. In part (i)(c) BOTH elements of the reason were needed for credit (i.e. that no suitable table is likely to exist AND the experience is small).*

## Question 6

(i) The probability that an insect will survive for 10 days, ${}_{10}p_0$, is given by the formula

$$ {}_{10}p_0 = \exp\left(-\int_0^{10} \mu_x dx\right). $$

Since the force of mortality is constant up to age 30 days at a value of 0.05,

$$ {}_{10}p_0 = \exp\left(-\int_0^{10} 0.05dx\right) = \exp\left(-\left[0.05x\right]_0^{10}\right) = \exp(-0.5) = 0.6065. $$

(ii) The probability that an insect 10 days old will survive for a further 30 days (that is to exact age 40 days) is given by

$$ {}_{30}p_{10} = \exp\left(-\int_{10}^{40} \mu_x dx\right). $$

Since ${}_{30}p_{10} = {}_{20}p_{10} \cdot {}_{10}p_{30}$, this is equal to

$$ \exp\left(-\int_{10}^{30} 0.05dx\right)\exp\left(-\int_0^{10} 0.05\exp(0.01x)dx\right) $$

$$ = \exp\left(-\left[0.05x\right]_{10}^{30}\right)\exp\left(-\left[\frac{0.05}{0.01}\exp(0.01x)\right]_0^{10}\right) $$

$$ = e^{-(1.5-0.5)}e^{-(5\exp(0.1)-5\exp(0))} $$

$$ = e^{-1}e^{-0.5258} = 0.3679 \times 0.5911 = 0.2174. $$

(iii) If the required age is $30+a$, then we have

$$ {}_{30+a}p_0 = {}_{30}p_0 \cdot {}_a p_{30} = 0.1. $$

Now

$$ {}_{30}p_0 = \exp\left[-\int_0^{30} 0.05dx\right] = \exp(-1.5) = 0.2231. $$

So $\quad {}_a p_{30} = \dfrac{0.1}{0.2231} = 0.4483.$

Using the result from part (ii), we have

$$ {}_a p_{30} = \exp\left(-\left[\frac{0.05}{0.01}e^{0.01x}\right]_0^a\right) = \exp\left(-\left[\frac{0.05}{0.01}e^{0.01a} - \frac{0.05}{0.01}\right]\right) = \exp(5 - 5e^{0.01a}) $$

Therefore

$e^{5(1-\exp(0.01a))} = 0.4483$,

whence

$\log_e 0.4483 = 5(1 - e^{0.01a})$,

so that

$$ \begin{aligned} 1 - e^{0.01a} &= -0.1605 \\ e^{0.01a} &= 1.1605 \\ 0.01a &= 0.1488 \\ a &= 14.88 \end{aligned} $$

Therefore the required age is 14.88+30 = 44.88 days.

*Most candidates answered part (i) of this question correctly. Part (ii) was less well answered, and only a minority of candidates managed to obtain the correct answer to part (iii). A common error was to use the limits 40 and 30 when integrating 0.05exp(0.01x).*

## Question 7

(i) It is a <u>stochastic process</u> in <u>discrete or continuous time</u>.

The state space is all the natural numbers {0, 1, 2, ... }

The value of the process $X(t)$ is a non-decreasing (OR an increasing) function of time $t$
OR
the value of the process goes up one at a time.

(ii) (a)

| *Process* | *State space* | *Time set* |
|---|---|---|
| Simple random walk | Discrete | Discrete |
| Compound Poisson process | Either | Continuous |
| Markov Chain | Discrete | Discrete |

(b)

| *Process* | *Application* |
|---|---|
| Simple random walk | The number of customers in the shop each time the door is opened |
| Compound Poisson process | The weight of almonds remaining in stock at any time in the day.<br>OR<br>value of goods sold at any time during the day |
| Markov chain | The number of customers owning loyalty cards at the end of each week. |

*In part (i), it was not sufficient just to say "discrete state space". The fact the state space is all natural numbers should be indicated for credit. In part (ii)(B), some candidates only gave one example IN TOTAL, whereas the question asked for an example FOR EACH PROCESS. In part (ii)(b) other examples were given credit. The criterion used to award credit were whether the example COULD be modelled using the relevant process and how USEFUL to the shopkeeper such a model might be!*

## Question 8

(i) EITHER

The central exposed to risk at age $x, E_x^c$ , is the waiting time in a multiple-state or Poisson model.

The initial exposed to risk is equal to the central exposed to risk plus the time elapsing between the date of death and the end of the rate interval for those who are observed to die during the rate interval.

OR

If the age at entry of life $i$ is $x + a_i$ , and the age at exit is $x+b_i$ for lives which do not die, and $x+t_i$ for lives who die, then the central exposed to risk is equal to
$\sum_i [(x_i + b_i) - (x_i - a_i)] = \sum_i (b_i - a_i)$ for lives who do not die, and
$\sum_i [(x_i + t_i) - (x_i - a_i)] = \sum_i (t_i - a_i)$ for lives who die.

The initial exposed to risk is given by the central exposed to risk plus a quantity equal to $\sum_i (1 - t_i)$ for the lives who die.

If the rate interval is the year of age between exact ages $x$ and $x$+1, and if deaths are approximately uniformly distributed across the year of age, the initial exposed to risk is approximately equal to $E_x^c + 0.5d_x$, where $d_x$ is the number of deaths between exact ages $x$ and $x$+1.

The central exposed to risk estimates $\mu_x$ whereas the initial exposed-to risk estimates $q_x$.

(ii) The maximum likelihood estimate of the force of mortality in the two-state model is deaths divided by the central exposed to risk.

The central exposed to risk is calculated as shown in the table below.

| *Life* | *Entry into observation* | *Exit from observation* | *Months exposed to risk* |
|---|---|---|---|
| 1 | 1 August 2008 | 1 August 2009 | 12 |
| 2 | 1 September 2008 | 1 September 2009 | 12 |
| 3 | 1 December 2008 | 1 February 2009 | 2 |
| 4 | 1 January 2009 | 1 January 2010 | 12 |
| 5 | 1 February 2009 | 1 February 2010 | 12 |
| 6 | 1 March 2009 | 1 December 2009 | 9 |
| 7 | 1 June 2009 | 1 June 2010 | 12 |
| 8 | 1 July 2009 | 1 July 2010 | 12 |
| 9 | 1 September 2009 | 1 September 2010 | 12 |
| 10 | 1 November 2009 | 1 December 2009 | 1 |

The total number of months exposed to risk is therefore

12 + 12 + 2 + 12 + 12 + 9 + 12 + 12 + 12 + 1 = 96

which is 8 years

There were 3 deaths.

Therefore the maximum likelihood estimate of the force of mortality is $\frac{3}{8} = 0.375$.

(iii) If the force of mortality is $\mu_0$, then

$$q_0 = 1 - \exp(-\mu_0) = 1 - \exp(-0.375) = 0.3127.$$

EITHER ALTERNATIVE 1

(iv) The initial exposed to risk, $E_0$ is approximately equal to $E_0^c + 0.5d_0$, where $E_0^c$ is the central exposed to risk and $d_0$ is the number of deaths.

Therefore we have

$$q_0* = \frac{d_0}{E_0^c + 0.5d_0} = \frac{3}{8+0.5(3)} = \frac{3}{9.5} = 0.3158.$$

(v) $q_0$ * is calculated assuming a uniform distribution of deaths over the year of age between birth and exact age 1 year, whereas $q_0$ assumes a constant force of mortality between exact ages 0 and 1.

These assumptions are different, implying a different distribution of deaths over the first year of life.

OR ALTERNATIVE 2

(iv) As the only way of leaving observation is through death, the initial exposed to risk is 10 and $q_0* = \frac{3}{10} = 0.3$.

(v) $q_0$ * is calculated using the exact initial exposed to risk, making no assumptions about the shape of the force of mortality during the interval,

OR

In the calculation of $q_0$ * lives could die at any time during the year of age, so they are treated as being exposed to risk for the entire year, whereas $q_0$ assumes a constant force of mortality between exact ages 0 and 1, which implies an assumption about the distribution of deaths over this interval.

*In part (i) full credit could be obtained for rather less than is written in the solution above. Credit can be given for any clear algebraic expressions in terms of the entry age $x+a_i$, the age at death, $x+t_i$ and the age at exit if the life did not die, $x+b_i$, which made clear the difference between the central and initial exposeds to risk.*

*In part (v) the wording did not have to be precise. The Examiners were looking for some understanding of the idea that different assumptions are made about the shape of the force of mortality over the rate interval.*

# Question 9

(i) A Markov jump process is a continuous time, discrete state process

THEN EITHER

in which, given the present state of the process, additional knowledge of the past is irrelevant for the calculation of the probability distribution of future values of the process.

OR

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

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 $S$.

(ii) Using the Markov property, and conditioning on the state occupied at age $x + t$, we have

$$_{t+dt}p_x^{13} = {}_t p_x^{11}\,{}_{dt}p_{x+t}^{13} + {}_t p_x^{12}\,{}_{dt}p_{x+t}^{23} + {}_t p_x^{13}\,{}_{dt}p_{x+t}^{33}$$

Assume that $_{dt}p_{x+t}^{ij} = \mu_{x+t}^{ij}dt + o(dt)$, $i \neq j$

where $\lim\limits_{dt \to 0^+} \dfrac{o(dt)}{dt} = 0$.

Substituting for the $_{dt}p_{x+t}^{ij}$ in the equation above, and noting that $_{dt}p_{x+t}^{33} = 1$ since return from the state "Dead" is impossible, produces

$$_{t+dt}p_x^{13} = {}_t p_x^{11}\mu_{x+t}^{13}dt + {}_t p_x^{12}\mu_{x+t}^{23}dt + {}_t p_x^{13} + o(dt)$$

so that

$$_{t+dt}p_x^{13} - {}_t p_x^{13} = {}_t p_x^{11}\,\mu_{x+t}^{13}dt + {}_t p_x^{12}\mu_{x+t}^{23}dt + o(dt)$$

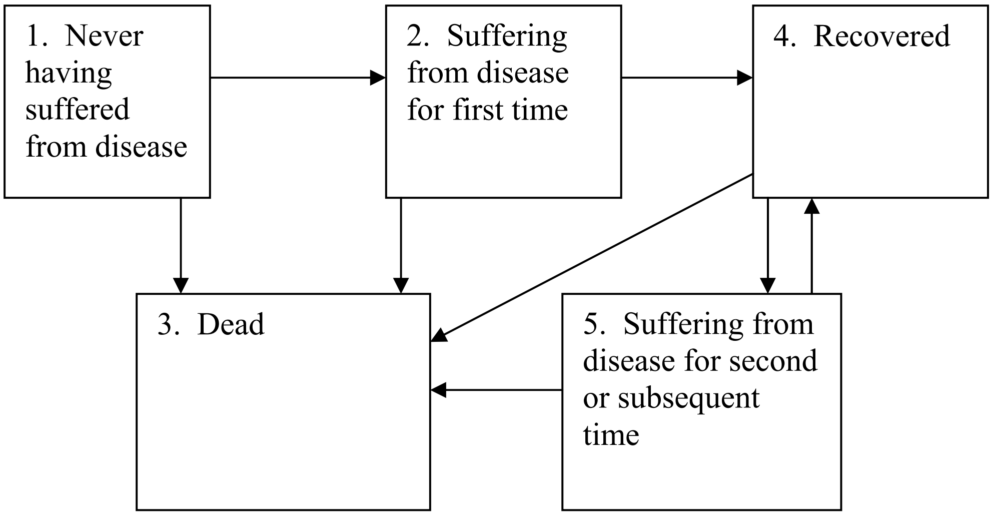
and, taking limits, we have

$$\lim_{dt \to 0^+} \frac{_{t+dt}p_x^{13} - {}_t p_x^{13}}{dt} = {}_t p_x^{11}\,\mu_{x+t}^{13} + {}_t p_x^{12}\mu_{x+t}^{23}$$

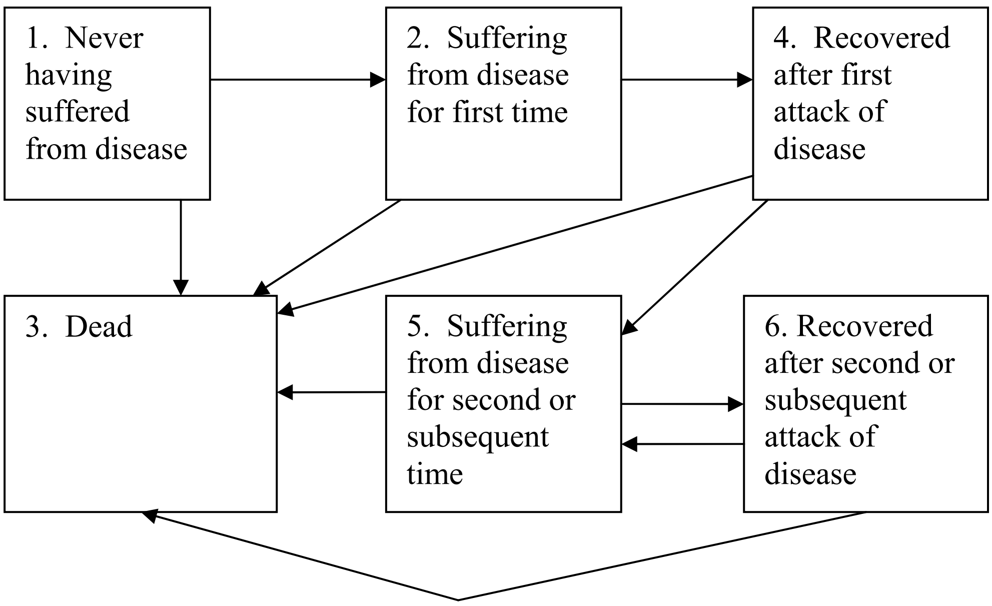
So

$$\frac{d}{dt}\,{}_t p_x^{13} = {}_t p_x^{11}\,\mu_{x+t}^{13} + {}_t p_x^{12}\mu_{x+t}^{23}.$$

(iii) EITHER




OR




*This question was fairly well answered, though many candidates omitted the initial point in part (ii), that we need the Markov property to be able to condition on the state occupied at x+t. A common error in part (iii) was a four-state solution with the states "1 – Never having suffered from the disease", "2 – Suffering from the disease", "3 – Dead", and "4 – Recovered". This is not correct, as the probability of moving from the state "Suffering from the disease" to the state "Dead" depends on whether the person is suffering from the disease for the first time or the second or subsequent time.*

*In part (iii) both alternatives were accepted. The second allows for the possibility that the effect of contracting the disease for the second time in raising the risk of death persists even after the patient recovers from the second or subsequent attack.*

## Question 10

(i) Right censoring is present
for those still alive and in hospital at the end of August
OR
for those who left hospital while still alive

Left censoring is not present

The censoring is likely to be informative, since those leaving hospital are likely to be in much better health than those who remain. (The idea of going home to die when you have had a lung transplant is a little tenuous.)

(ii) The durations and outcomes are shown in the table below.

| *Patient* | *Died/Censored* | *Duration* |
|---|---|---|
| A | Died | 2 |
| G | Died | 5 |
| J | Died | 5 |
| B | Censored | 29 |
| E | Died | 32 |
| M | Censored | 38 |
| H | Censored | 56 |
| K | Died | 56 |
| N | Censored | 62 |
| L | Censored | 66 |
| I | Censored | 70 |
| F | Censored | 80 |
| D | Censored | 84 |
| C | Censored | 87 |

EITHER ALTERNATIVE 1

Assuming that at duration 56 the death occurred before the life was censored, the Kaplan-Meier estimate is as follows:

| $t_j$ | $n_j$ | $d_j$ | $c_j$ | $\lambda_j = \dfrac{d_j}{n_j}$ |
|---|---|---|---|---|
| 0 | 14 | 0 | 0 | 0 |
| 2 | 14 | 1 | 0 | 1/14 |
| 5 | 13 | 2 | 1 | 2/13 |
| 32 | 10 | 1 | 1 | 1/10 |
| 56 | 8 | 1 | 7 | 1/8 |
| +½ | +½ | +½ | +½ | |

The Kaplan-Meier estimate at duration $t$ is given by the product of $1-\dfrac{d_j}{n_j}$ over

durations up to and including $t$. Thus the Kaplan-Meier estimate of the survival function is

| $t$ | $\hat{S}(t)$ | |
|---|---|---|
| $0 \le t < 2$ | 1.0000 | |
| $2 \le t < 5$ | 0.9286 | OR 13/14 |
| $5 \le t < 32$ | 0.7857 | OR 11/14 |
| $32 \le t < 56$ | 0.7071 | OR 99/140 |
| $56 \le t < 92$ | 0.6188 | OR 99/160 |

OR ALTERNATIVE 2

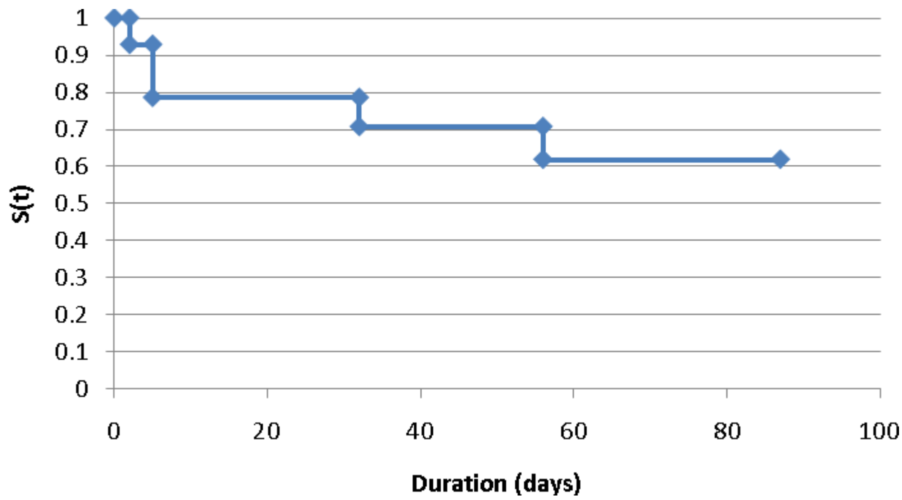
Assuming that at duration 56 the death occurred after the life was censored, the Kaplan-Meier estimate is as follows:

| $t_j$ | $n_j$ | $d_j$ | $c_j$ | $\lambda_j = \dfrac{d_j}{n_j}$ |
|---|---|---|---|---|
| 0 | 14 | 0 | 0 | 0 |
| 2 | 14 | 1 | 0 | 1/14 |
| 5 | 13 | 2 | 1 | 2/13 |
| 32 | 10 | 1 | 2 | 1/10 |
| 56 | 7 | 1 | 6 | 1/7 |

The Kaplan-Meier estimate at duration $t$ is given by the product of $1-\dfrac{d_j}{n_j}$ over durations up to and including $t$. Thus the Kaplan-Meier estimate of the survival function is

| $t$ | $\hat{S}(t)$ | | |
|---|---|---|---|
| $0 \le t < 2$ | 1.0000 | | |
| $2 \le t < 5$ | 0.9286 | OR | 13/14 |
| $5 \le t < 32$ | 0.7857 | OR | 11/14 |
| $32 \le t < 56$ | 0.7071 | OR | 99/140 |
| $56 \le t < 92$ | 0.6061 | OR | 297/490 |

(iii)




(iv) The probability of death within 4 weeks is $1 - S(28) = 0.2143$.

*In part (i) candidates could receive credit for saying that left censoring was present IF they gave a valid reason (which typically involved the imprecise measurement of the times of surgery or of events – the left censoring arising as a special case of interval censoring).*
*In part (ii) each error was only penalised once. Correct calculations which carried forward earlier errors were given full credit. However, candidates who did not list the durations they were using, but then presented incorrect estimates of the survival function, were more heavily penalised, as it was not clear how many errors they had made.*

*In part (ii) candidates who assume that the death at duration 56 takes place after the censoring at the same duration (ALTERNATIVE 2) were required to state this assumption for full credit. For ALTERNATIVE 1, the assumption that the death at duration 56 takes place before the censoring does not need to be stated for full credit, as it is the convention when calculating Kaplan-Meier estimates.*

*In part (iii) the plotted function should be consistent with the answer to part (ii). If the answer to part (ii) was incorrect but the incorrect answer to part (ii) was correctly plotted in part (iii), full credit could be awarded to part (iii).*

## Question 11

(i) The chi-squared test is a suitable overall test.

Let $\mu_x$ be the force of mortality in age-group $x$ in the sample.
Let $\mu_x^s$ be the force of mortality in age group $x$ in the national population.
Let $E_x^c$ be the central exposed to risk in the sample.

Then if $z_x = \dfrac{E_x^c \mu_x - E_x^c \mu_x^s}{\sqrt{E_x^c \mu_x^s}}$

the test statistic is $\sum_x {z_x}^2 \sim \chi^2{}_m$,

THEN EITHER

where $m$ is the number of age groups, which in this case is 8.

The calculations are shown below.

| *Age-group* | *Expected deaths* | $z_x$ | $z_x^2$ |
|---|---|---|---|
| 5–14 | 18.7935 | –1.3364 | 1.7860 |
| 15–24 | 50.5460 | –0.4988 | 0.2488 |
| 25–34 | 59.8842 | –1.0188 | 1.0380 |
| 35–44 | 53.3092 | –0.4532 | 0.2054 |
| 45–54 | 39.6396 | –1.0546 | 1.1121 |
| 55–64 | 23.4140 | –0.0856 | 0.0073 |
| 65–74 | 13.3926 | –0.1073 | 0.0115 |
| 75–84 | 1.8200 | 0.8747 | 0.7651 |

Therefore the value of the test statistic is 5.1742.

The critical value of the chi-squared distribution at the 5% level of significance with 8 degrees of freedom is 15.51.

Since 5.1742 < 15.51 we do not reject the null hypothesis that the mortality rate from tuberculosis in the sample is the same as that in the national population.

OR

where $m$ is the number of age groups, which in this case is 7, because we should combine age groups 65–74 and 75–84 as the expected number of deaths in age group 75–84 years is less than 5

The calculations are shown below.

| *Age-group* | *Expected deaths* | $z_x$ | $z_x^2$ |
|---|---|---|---|
| 5–14 | 18.7935 | –1.3364 | 1.7860 |
| 15–24 | 50.5460 | –0.4988 | 0.2488 |
| 25–34 | 59.8842 | –1.0188 | 1.0380 |
| 35–44 | 53.3092 | –0.4532 | 0.2054 |
| 45–54 | 39.6396 | –1.0546 | 1.1121 |
| 55–64 | 23.4140 | –0.0856 | 0.0073 |
| 65–84 | 15.2126 | 0.2019 | 0.0408 |

Therefore the value of the test statistic is 4.438.

The critical value of the chi-squared distribution at the 5% level of significance with 7 degrees of freedom is 14.07.

Since 4.438 < 14.07 we do not reject the null hypothesis that the mortality rate from tuberculosis in the sample is the same as that in the national population

(ii) (a) Small bias which is not great enough for the chi-squared test to detect.

EITHER

(b) **Signs test**

Under the null hypothesis that the mortality rate from tuberculosis in the sample is the same as that in the national population,

the number of positive signs is distributed Binomial ($m$, 0.5), where $m$ is the number of ages.

We have 1 positive sign.

The probability of 1 or fewer positive signs is given by

$$\binom{8}{0}0.5^8 + \binom{8}{1}0.5^8 = 0.0352.$$

OR (if only 7 age groups are being used)

$$\binom{7}{0}0.5^7 + \binom{7}{1}0.5^7 = 0.0625.$$

We use a two-tailed test (since too few or too many positive signs would be a problem)

so we reject the null hypothesis if the probability of 1 or fewer positive signs is less than 0.025.

Since 0.0352 (or 0.0625) > 0.025

we do not reject the null hypothesis.

OR

(b) **Cumulative deviations test**

Under the null hypothesis that the mortality rate from tuberculosis in the sample is the same as that in the national population

the test statistic

$$\frac{\sum_x \left( E_x^c \mu_x - E_x^c \mu_x^s \right)}{\sqrt{\sum_x E_x^c \mu_x^s}} \sim \text{Normal}(0,1)$$

The calculations are shown in the table below

| *Age-group* | $E_x^c \mu_x - E_x^c \mu_x^s$ | $E_x^c \mu_x^s$ |
|---|---|---|
| 5–14 | –5.7935 | 18.7935 |
| 15–24 | –3.5460 | 50.5460 |
| 25–34 | –7.8842 | 59.8842 |
| 35–44 | –3.3092 | 53.3092 |
| 45–54 | –6.6396 | 39.6396 |
| 55–64 | –0.4140 | 23.4140 |
| 65–74 | –0.3926 | 13.3926 |
| 75–84 | 1.1800 | 1.8200 |
| Σ | –26.7991 | 260.7991 |

So the value of the test statistic is $\dfrac{-26.7991}{\sqrt{260.7991}} = 1.6595$.

Using a 5% level of significance, we see that $-1.96 < 1.6596 < 1.96$.

We do not reject the null hypothesis.

(a) Individual ages at which there are unusually large differences between the sample and the national experience.

(b) **Individual standardised deviations**

Under the null hypothesis that the mortality rate from tuberculosis in the sample is the same as that in the national population

we would expect the individual deviations to be distributed Normal (0,1)

and therefore only 1 in 20 $z_x$ s should have absolute magnitudes greater than 1.96
OR
none should lie outside the range (–3, +3)
OR
diagram showing split of deviations actual versus expected.

Since the largest deviation is less in absolute magnitude than 1.96 we do not reject the null hypothesis.

(a) Sections of the data where there is appreciable bias, revealed by runs or clumps of signs of the same type.

EITHER

(b) **Grouping of signs test**

Under the null hypothesis that the mortality rate from tuberculosis in the sample is the same as that in the national population

$G$ = Number of groups of positive $z$s = 1

$m$ = number of deviations = 8 (or 7 if last two age groups combined)
$n_1$ = number of positive deviations = 1
$n_2$ = number of negative deviations = 7 (or 6 if last two age groups combined)

THEN EITHER

We want $k^*$ the largest $k$ such that

$$\sum_{t=1}^{k} \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}} < 0.05$$

The test fails at the 5% level if $G \le k^*$.

In the table in the Gold Book a value for $k^*$ is not given,
OR
The table in the Gold Book shows that $k^* = 0$,

so we are not able to reject the null hypothesis
OR
so there is no evidence of clumping.

OR

For $t = 1$

$$\begin{pmatrix} n_1 - 1 \\ t - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{which is 1}$$

So this test is automatically passed
OR
There is no evidence of clumping
OR
We cannot reject the null hypothesis.

OR

(b) **Serial correlations (lag 1)**

The calculations are shown in the tables below.

EITHER USING SEPARATE MEANS FOR THE $z_x$ AND $z_{x+1}$

| *Age group* | $z_x$ | $z_x$ | $A = z_x - \bar{z}$ | $B = z_{x+1} - \bar{z}$ | $AB$ | $A^2$ | $B^2$ |
|---|---|---|---|---|---|---|---|
| 5–14 | –1.336 | –0.499 | –0.686 | –0.164 | 0.342 | 0.470 | 0.027 |
| 15–24 | –0.499 | –1.019 | 0.152 | –0.684 | –0.155 | 0.023 | 0.468 |
| 25–34 | –1.019 | –0.453 | –0.368 | –0.118 | 0.167 | 0.136 | 0.014 |
| 35–44 | –0.453 | –1.055 | 0.197 | –0.720 | –0.208 | 0.039 | 0.518 |
| 45–54 | –1.055 | –0.086 | –0.404 | 0.249 | 0.035 | 0.163 | 0.062 |
| 55–64 | –0.086 | –0.107 | 0.565 | 0.228 | –0.061 | 0.319 | 0.052 |
| 65–74 | –0.107 | 0.875 | 0.543 | 1.210 | 0.475 | 0.295 | 1.463 |
| 75–84 | | | | | | | |
| $\bar{z}$ | –0.651 | –0.335 | | Sum | 0.595 | 1.446 | 2.604 |

0.595/√(1.446*2.604) = 0.307

Test 0.307 (√8) = 0.868 against Normal (0,1), and, since 0.868 < 1.645, we do not reject the null hypothesis.

that the mortality rate from tuberculosis in the sample is the same as that in the national population

OR USING THE FORMULA IN THE GOLD BOOK

| *Age group* | $z_x$ | $z_x$ | $A = z_x - \bar{z}$ | $B = z_{x+1} - \bar{z}$ | $AB$ | $A^2$ |
|---|---|---|---|---|---|---|
| 5–14 | –1.336 | –0.499 | –0.876 | –0.039 | 0.034 | 0.767 |
| 15–24 | –0.499 | –1.019 | –0.039 | –0.559 | 0.022 | 0.002 |
| 25–34 | –1.019 | –0.453 | –0.559 | 0.007 | –0.004 | 0.312 |
| 35–44 | –0.453 | –1.055 | 0.007 | –0.595 | –0.004 | 0.000 |
| 45–54 | –1.055 | –0.086 | –0.595 | 0.374 | –0.223 | 0.354 |
| 55–64 | –0.086 | -0.107 | 0.374 | 0.353 | 0.132 | 0.140 |
| 65–74 | –0.107 | 0.875 | 0.353 | 1.335 | 0.471 | 0.125 |
| 75–84 | 0.875 | | 1.335 | | | 1.782 |
| $\bar{z}$ | –0.460 | | | Sum | 0.428 | 3.481 |

$$\frac{\frac{1}{7}(0.428)}{\frac{1}{8}(3.481)} = 0.141$$

Test 0.141 (√8) = 0.397 against Normal (0,1), and, since 0.397 < 1.645, we do not reject the null hypothesis.

that the mortality rate from tuberculosis in the sample is the same as that in the national population

(iii) In none of the tests we have performed do we reject the null hypothesis.

Therefore it seems that the mortality from tuberculosis in the town is the same as the national force of mortality.

*In part (ii) the null hypothesis should be stated somewhere for each test. It could be stated at the beginning, or in the conclusion. As long as it is correctly stated somewhere, full credit was given. In part (iii), the comment should be consistent with the results of the tests performed in parts (i) and (ii) to gain credit.*

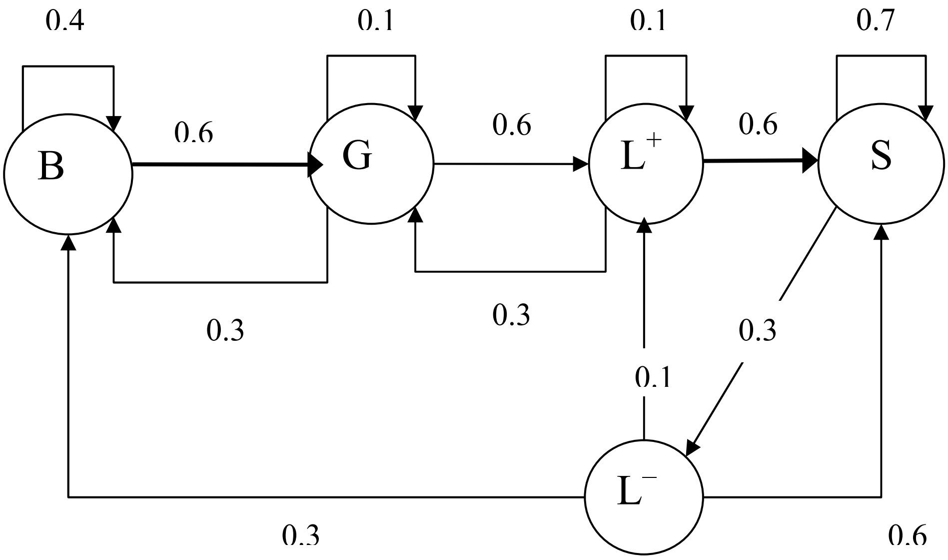
*Most candidates made a good attempt at part (i). Attempts at part (ii) were more varied. In particular, most candidates did not point out that the chi-squared test only fails to detect SMALL (but consistent) bias. If the bias is large and consistent, the chi-squared test will detect it.*

## Question 12

(i) Past history is needed to decide where to go in the chain.

If a customer is at L and reduces his or her order, you need to know what level of discount he was at the previous year to determine whether he or she drops one or two levels of discount.

(ii) The L level needs to be split into two.
$L^+$ is Loyalty Price with no reduction in demand last year
$L^-$ is Loyalty Price with reduction in demand last year




*The probabilities were not required for full credit for this diagram.*

(iii) (a)

$$\begin{array}{ccccc} B & G & L^+ & L^- & S \\ \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 \end{array}$$

$$\begin{pmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.3 & 0.1 & 0.6 & 0 & 0 \\ 0 & 0.3 & 0.1 & 0 & 0.6 \\ 0.3 & 0 & 0.1 & 0 & 0.6 \\ 0 & 0 & 0 & 0.3 & 0.7 \end{pmatrix}$$

(b) $\pi = \pi P$

$\pi_1 = 0.4\,\pi_1 + 0.3\,\pi_2 + 0.3\,\pi_4$ (1)
$\pi_2 = 0.6\,\pi_1 + 0.1\,\pi_2 + 0.3\,\pi_3$ (2)
$\pi_3 = 0.6\,\pi_2 + 0.1\,\pi_3 + 0.1\,\pi_4$ (3)
$\pi_4 = 0.3\,\pi_5$ (4)
$\pi_5 = 0.6\,\pi_3 + 0.6\,\pi_4 + 0.7\,\pi_5$ (5)
$\pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 = 1$

(4) gives $\pi_4 = 0.3\,\pi_5$

(5) gives
$$\begin{aligned} 0.6\pi_3 &= 0.3\,\pi_5 - 0.6(0.3\pi_5) \\ &= 0.12\,\pi_5 \\ \pi_3 &= 0.2\,\pi_5 \end{aligned}$$

(3) gives
$$\begin{aligned} 0.6\,\pi_2 &= 0.9\,\pi_3 - 0.1\,\pi_4 \\ &= 0.18\,\pi_5 - 0.03\,\pi_5 \\ &= 0.15\,\pi_5 \\ \pi_2 &= 0.25\,\pi_5 \end{aligned}$$

(2) gives
$$\begin{aligned} 0.6\,\pi_1 &= 0.9\,\pi_2 - 0.3\,\pi_3 \\ &= 0.9(0.25)\,\pi_5 - 0.3(0.2)\,\pi_5 \\ &= 0.225\,\pi_5 - 0.06\,\pi_5 \\ &= 0.165\,\pi_5 \\ \pi_1 &= 0.275\,\pi_5 \end{aligned}$$

$\pi_5\,(0.275 + 0.25 + 0.2 + 0.3 + 1) = 1$
$\pi_5 = 1 / 2.025$
$= 0.49382716$

| | | |
|---|---|---|
| $\pi_1 = 0.13580$ | | OR 11/81 |
| $\pi_2 = 0.12346$ | | OR 10/81 |
| $\pi_3 = 0.09877$ | ) | OR 8/81 |
| $\pi_4 = 0.14815$ | ) 0.24692 | OR 12/81 |
| $\pi_5 = 0.49383$ | | OR 40/81 |

(c) Average price for a bale of hay is

£8 × (1 × 0.1358 + 0.9 × 0.12346 + 0.8 × (0.09877 + .14815) + 0.75 ×.49383 )

= £6.5181

(iv) $$\begin{pmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.3 & 0.1 & 0.6 & 0 & 0 \\ 0 & 0.3 & 0.1 & 0 & 0.6 \\ 0.3 & 0 & 0.1 & 0 & 0.6 \\ 0 & 0 & 0 & 0.3 & 0.7 \end{pmatrix}\begin{pmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.3 & 0.1 & 0.6 & 0 & 0 \\ 0 & 0.3 & 0.1 & 0 & 0.6 \\ 0.3 & 0 & 0.1 & 0 & 0.6 \\ 0 & 0 & 0 & 0.3 & 0.7 \end{pmatrix}$$

$$= \begin{pmatrix} .16+.18 & .24+.06 & .36 & - & - \\ .12+.03 & .18+.01+.18 & .06+.06 & - & .36 \\ .09 & .03+.03 & .18+.01 & .18 & .18 \\ .12 & .18+.03 & .01 & .18 & .06+.42 \\ .09 & - & .03 & .21 & .18+.42 \end{pmatrix} = \begin{pmatrix} .34 & .3 & .36 & - & - \\ .15 & .37 & .12 & - & .36 \\ .09 & .06 & .19 & .18 & .48 \\ .12 & .21 & .01 & .18 & .48 \\ .09 & - & .03 & .21 & .67 \end{pmatrix}$$

THEN ALTERNATIVE 1

Using the long-run probabilities of being in $L^+$ and $L^-$, therefore

the chance of being at L in two years' time is

(0.19 + 0.18)*0.4 + (0.18 + 0.01)*0.6 = 0.262.

OR ALTERNATIVE 2

Assuming there is an equal probability of being $L^+$ and $L^-$,

the chance of being at L in two years' time is

(0.19 + 0.18)*0.5 + (0.18 + 0.01)*0.5 = 0.28.

OR ALTERNATIVE 3

We do not know the relative proportions in $L^+$ and $L^-$,

but for those in $L^+$ the chance of being in L in two years' time is 0.19 + 0.18 = 0.37, and for those in $L^+$ the chance of being in L in two years' time is 0.18 + 0.01 = 0.19.

OR ALTERNATIVE 4

We do not know the relative proportions in $L^+$ and $L^-$,

and so it is not possible to evaluate the overall probability that a customer in L will be in L in two years' time.

(v) A constant figure takes no account of the amount of hay which Farmer Giles has to sell: for example a drought year could produce very little which one large customer may buy in its entirety.

The amount of hay in the local market is important.

Another supplier may try a heavy discounted year to get into the market.

Customers' behaviour may depend on the discount level they are at.

There may be national trends in the demand for hay e.g. a sudden trend towards vegetarianism.

A 60% chance of increasing may be implausible, as field space is likely to be limited, so a constant increase in numbers unlikely.

Customers' behaviour may depend on the amount of hay they typically purchase.

*A common error in part (ii) was to split state G into two states as well as splitting state L. This is not required to model the system with the Markov property and so was penalised. However, candidates who split state G and then followed through with a correct matrix in part (iii)(a) and correct solutions in part (iii)(b) were not penalised again. Note that splitting state G should produce the same answer to part (iii)(b), though more work will be needed!*

*In part (iv) candidates who adopted ALTERNATIVE 4, in which they declined to give an overall answer on the grounds that they do not know the proprotions in states $L^+$ and $L^-$ , were only given credit if they presented a reasoned argument with evidence.*

*In part (v) credit was given for other sensible suggestions.*

## END OF EXAMINERS' REPORT

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINATION

7 October 2011 (am)

## Subject CT4 — Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 11 questions, beginning your answer to each question on a separate sheet.*

5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

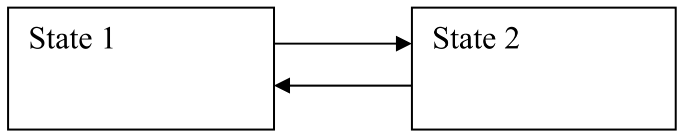
 

**1** The diagrams below show three Markov chains, where arrows indicate a non-zero transition probability.

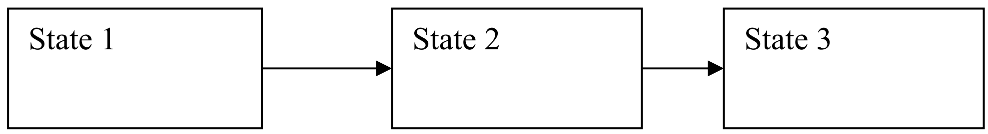
State whether each of the chains is:

(a) irreducible.
(b) periodic, giving the period where relevant. [3]

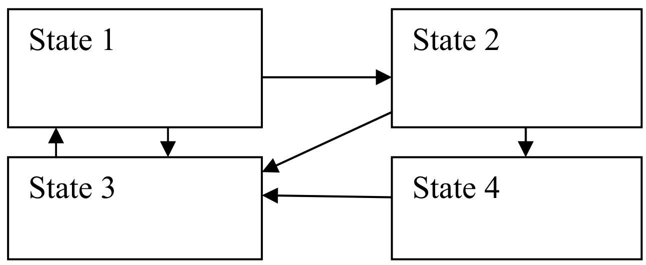
A.




B.




C.




**2** (i) Describe what is represented by each of the central rate of mortality, $m_x$ , and the initial rate of mortality, $q_x$. [2]

(ii) State the circumstance in which $m_x = \mu_x$. [1]

[Total 3]

**3** Describe how a strictly stationary stochastic process differs from a weakly stationary stochastic process. [3]

**4** A new weedkiller was tested which was designed to kill weeds growing in grass. The weedkiller was administered via a single application to 20 test areas of grass. Within hours of applying the weedkiller, the leaves of all the weeds went black and died, but after a time some of the weeds re-grew as the weedkiller did not always kill the roots.

The test lasted for 12 months, but after six months five of the test areas were accidentally ploughed up and so the trial on these areas had to be discontinued. None of these five areas had shown any weed re-growth at the time they were ploughed up.

- Ten of the remaining 15 areas experienced a re-growth of weeds at the following durations (in months): 1, 2, 2, 2, 5, 5, 8, 8, 8, 8.

- Five areas still had no weed re-growth when the trial ended after 12 months.

(i) Describe, giving reasons, the types of censoring present in the data. [2]

(ii) Estimate the probability that there is no re-growth of weeds nine months after application of the weedkiller using either the Kaplan-Meier or the Nelson-Aalen estimator. [4]
[Total 6]

**5** (i) List the factors which should be considered in assessing the suitability of a model for a particular exercise. [3]

(ii) Assess the suitability of a multiple state model with three states: Healthy, Sick and Dead, for estimating the transition intensities in an analysis of claims for sickness benefit, in the light of your answer to (i). [4]
[Total 7]

**6** A recording instrument is set up to observe a continuous time process, and stores the results for the most recent 250 transitions. The data collected are as follows:

| *State i* | *Total time spent in state i (hours)* | *Number of transitions to State A* | *State B* | *State C* |
|---|---|---|---|---|
| *A* | 35 | Not applicable | 60 | 45 |
| *B* | 150 | 50 | Not applicable | 25 |
| *C* | 210 | 55 | 15 | Not applicable |

It is proposed to fit a Markov jump model using the data.

(i) (a) State all the parameters of the model.
(b) Outline the assumptions underlying the model. [4]

(ii) (a) Estimate the parameters of the model.
(b) Write down the estimated generator matrix of the model. [4]

(iii) Specify the distribution of the number of transitions from state *i* to state *j*, given the number of transitions out of state *i*. [1]
[Total 9]

**7** A study is made of the impact of regular exercise and gender on the risk of developing heart disease among 50–70 year olds. A sample of people is followed from exact age 50 years until either they develop heart disease or they attain the age of 70 years. The study uses a Cox regression model.

(i) List reasons why the Cox regression model is a suitable model for analyses of this kind. [3]

The investigator defined two covariates as follows:

- $Z_1 = 1$ if male, 0 if female.
- $Z_2 = 1$ if takes regular exercise, 0 otherwise.

The investigator then fitted three models, one with just gender as a covariate, a second with gender and exercise as covariates, and a third with gender, exercise and the interaction between them as covariates. The maximised log-likelihoods of the three models and the maximum likelihood estimates of the parameters in the third model were as follows:

| | |
|---|---|
| null model | –1,269 |
| gender | –1,256 |
| gender + exercise | –1,250 |
| gender + exercise + interaction | –1,246 |

| *Covariate* | *Parameter* |
|---|---|
| Gender | 0.2 |
| Exercise | –0.3 |
| Interaction | –0.35 |

(ii) Show that the interaction term is required in the model by performing a suitable statistical test. [5]

(iii) Interpret the results of the model. [3]

[Total 11]

**8** A continuous-time Markov process with states {Able to work ($A$), Temporarily unable to work ($T$), Permanently unable to work ($P$), Dead ($D$)} is used to model the cost of providing an incapacity benefit when a person is permanently unable to work. The generator matrix, with rates expressed per annum, for the process is estimated as:

| | $A$ | $T$ | $P$ | $D$ |
|---|---|---|---|---|
| $A$ | –0.15 | 0.1 | 0.02 | 0.03 |
| $T$ | 0.45 | –0.6 | 0.1 | 0.05 |
| $P$ | 0 | 0 | –0.2 | 0.2 |
| $D$ | 0 | 0 | 0 | 0 |

(i) Draw the transition graph for the process. [2]

(ii) Calculate the probability of a person remaining in state $A$ for at least 5 years continuously. [2]

Define $F(i)$ to be the probability that a person, currently in state $i$, will never be in state $P$.

(iii) Derive an expression for:

(a) $F(A)$ by conditioning on the first move out of state $A$.
(b) $F(T)$ by conditioning on the first move out of state $T$. [3]

(iv) Calculate $F(A)$ and $F(T)$. [2]

(v) Calculate the expected future duration spent in state $P$, for a person currently in state $A$. [2]
[Total 11]

**9** (i) State the principle of correspondence as it applies to the estimation of mortality rates. [1]

(ii) Explain why it might be difficult to ensure the principle of correspondence is adhered to, and give a specific example of an investigation where this may be the case. [2]

An actuary was asked to investigate the mortality of lives in a particular geographical area. Data are available of the population of this area, classified by age last birthday, on 1 January in each year. Data on the number of deaths in this area in each calendar year, classified by age nearest birthday at death, are also available.

(iii) Derive a formula which would allow the actuary to estimate the force of mortality at age $x + f$, $\mu_{x+f}$, in a particular calendar year, in terms of the available data, and derive a value for $f$. [6]

(iv) List four factors other than geographical location which a government statistical office might use to subdivide data for national mortality analysis. [2]
[Total 11]

**10** (i) Describe three shortcomings of the $\chi^2$ test for comparing crude estimates of mortality with a standard table and why they may occur. [3]

The following table gives an extract of data from a mortality investigation conducted in the rural highlands of a developed country. The raw data have been graduated by reference to a standard mortality table of assured lives.

| *Age* $x$ | *Expected deaths* | *Observed deaths* | $z_x$ | $z_x^2$ |
|---|---|---|---|---|
| 60 | 36.15 | 35 | –0.191 | 0.037 |
| 61 | 28.92 | 24 | –0.915 | 0.837 |
| 62 | 31.34 | 27 | –0.775 | 0.601 |
| 63 | 38.01 | 35 | –0.488 | 0.238 |
| 64 | 26.88 | 32 | 0.988 | 0.975 |
| 65 | 37.59 | 36 | –0.259 | 0.067 |
| 66 | 33.85 | 34 | 0.026 | 0.001 |
| 67 | 26.66 | 32 | 1.034 | 1.070 |
| 68 | 22.37 | 26 | 0.767 | 0.589 |
| 69 | 18.69 | 33 | 3.310 | 10.956 |
| 70 | 18.24 | 22 | 0.880 | 0.775 |

(ii) For each of the three shortcomings you described in (i):

(a) name a test that would detect that shortcoming.
(b) carry out the test on the data above. [12]

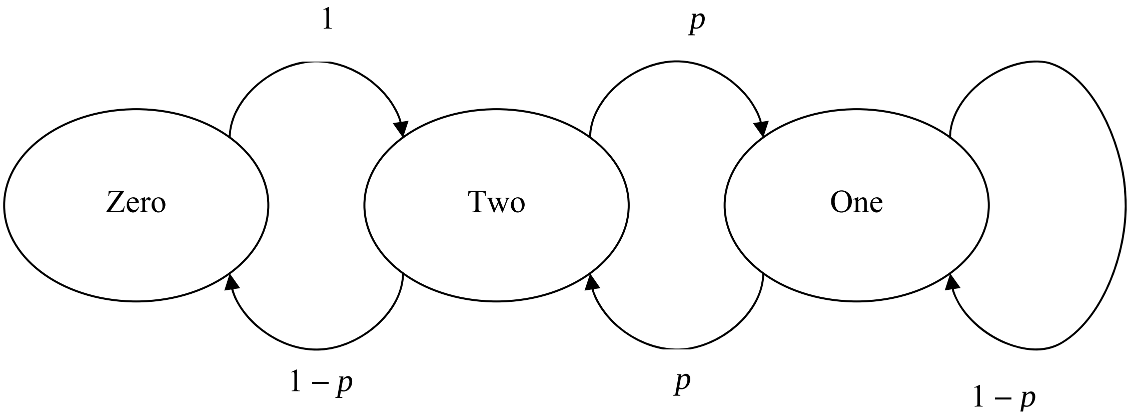
(iii) Comment on your results from (ii). [3]
[Total 18]

**11** An actuary walks from his house to the office each morning, and walks back again each evening. He owns two umbrellas. If it is raining at the time he sets off, and one or both of his umbrellas is available, he takes an umbrella with him. However if it is not raining at the time he sets off he always forgets to take an umbrella.

Assume that the probability of it raining when he sets off on any particular journey is a constant $p$, independent of other journeys.

This situation is examined as a Markov Chain with state space {0,1,2} representing the number of his umbrellas at the actuary's current location (office or home) and each time step representing one journey.

(i) Explain why the transition graph for this process is given by:




[3]

(ii) Derive the transition matrix for the number of umbrellas at the actuary's house before he leaves each morning, based on the number before he leaves the previous morning. [3]

(iii) Calculate the stationary distribution for the Markov Chain. [3]

(iv) Calculate the long run proportion of journeys (to or from the office) on which the actuary sets out in the rain without an umbrella. [2]

The actuary considers that the weather at the start of a journey, rather than being independent of past history, depends upon the weather at the start of the previous journey. He believes that if it was raining at the start of a journey the probability of it raining at the start of the next journey is $r$ ($0 < r < 1$), and if it was not raining at the start of a journey the probability of it raining at the start of the next journey is $s$ ($0 < s < 1$, $r \neq s$).

(v) Write down the transition matrix for the Markov Chain for the weather. [1]

(vi) Explain why the process with three states {0,1,2}, being the number of his umbrellas at the actuary's current location, would no longer satisfy the Markov property. [2]

(vii) Describe the additional state(s) needed for the Markov property to be satisfied, and draw a transition diagram for the expanded system. [4]

[Total 18]

**END OF PAPER**

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS’ REPORT

September 2011 examinations

## Subject CT4 — Models
## Core Technical

**Purpose of Examiners’ Reports**

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 who are 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. Although Examiners have access to the Core Reading, which is designed to interpret the syllabus, the Examiners are not required to examine the content of Core Reading. Notwithstanding that, the questions set, and the following comments, will generally be based on Core Reading.

For numerical questions the Examiners’ preferred approach to the solution is reproduced in this report. Other valid approaches are always given appropriate credit; where there is a commonly used alternative approach, this is also noted in the report. For essay-style questions, and particularly the open-ended questions in the later subjects, this report contains all the points for which the Examiners awarded marks. This is much more than a model solution – it would be impossible to write down all the points in the report in the time allowed for the question.

T J Birse
Chairman of the Board of Examiners

December 2011

## General comments on Subject CT4

Subject CT4 comprises five main sections: (1) 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); (2) stochastic processes, especially Markov chains and Markov jump processes; (3) models of a random variable measuring future lifetime; (4) the calculation of exposed to risk and the application of the principle of correspondence; (5) 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.

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.

## Comments on the September 2011 paper

The general performance was slightly worse than in April 2011 but well-prepared candidates scored well across the whole paper. As in previous diets, questions that required an element of explanation or analysis, such as Q5(ii) and Q7(iii) were less well answered than those that just involved calculation. The comments that follow the questions 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.

## Question 1

(a) A Yes, irreducible.
B No, not irreducible.
C Yes, irreducible.

(b) A Yes, <u>period is 2</u>
B No, not periodic.
C No, not periodic.

*This question was well answered, although many candidates failed to identify that C was aperiodic.*

## Question 2

(i) $m_x$ is the probability that a life alive between exact ages $x$ and $x$ dies

OR

$m_x$ is the probability of dying between exact ages $x$ and $x$ per person-year lived between exact ages $x$ and $x$

$q_x$ is the probability that a life alive at exact age $x$ dies before exact age $x$

[2]

(ii) $m_x$ and $\mu_x$ are equal when the force of mortality $\mu_{x+t}$ is constant for $0 \leq t < 1$.

*Answers to this question were disappointing. In part (i) some candidates defined $m_x$ as $\frac{q_x}{\int_0^1 {}_t p_x dt}$. For full credit, candidates who did this were required to explain what this expression means (e.g. by stating that ${}_t p_x$ is the expected amount of time spent alive between x and x+1 by a life alive at age x).*

## Question 3

A stochastic process is said to be strictly stationary if the joint distributions of $X_{t_1}, X_{t_2}, ..., X_{t_n}$ and $X_{t+t_1}, X_{t+t_2}, ..., X_{t+t_n}$ are identical for all $t, t_1, t_2, ..., t_n$ in the time set $J$ and for all integers $n$.

This means that the statistical properties of the process remain unchanged as time elapses.

Weak stationarity requires that the mean of the process, $m(t) = E(X_t)$, is constant, and

EITHER that the covariance of the process $E[(X_s - m(s))(X_t - m(t))]$ depends only on the time difference $t - s$.

OR $\mathrm{Cov}(X(t_1), X(t_2)) = \mathrm{Cov}(X(t_1 + h), X(t_2 + h)$ for all $t_1$, $t_2$ and $h > 0$.

Strict stationarity is a stringent condition which is hard to test, weak stationarity is a less stringent condition but easier to test in practice.

*This question was well answered. The last sentence was not required for full credit.*

## Question 4

(i) Right censoring: some areas never developed new weeds.

Type I censoring as the study lasts for a pre-determined time.

Random censoring as the accidental ploughing happened at a time which was not pre-determined.

Interval censoring as we do not know exactly when in each month the weed re-growth happened.

Non-informative censoring as the fact that an area was ploughed up tells us nothing about the duration to weed re-growth in any of the remaining areas.

(ii) EITHER

Kaplan-Meier estimator

| $t_j$ | $N_j$ | $D_j$ | $C_j$ | $\frac{D_j}{N_j}$ | $1-\frac{D_j}{N_j}$ |
|---|---|---|---|---|---|
| 0 | 20 | 0 | 0 | – | 1 |
| 1 | 20 | 1 | 0 | 1/20 | 19/20 |
| 2 | 19 | 3 | 0 | 3/19 | 16/19 |
| 5 | 16 | 2 | 5 | 2/16 | 14/16 |
| 8 | 9 | 4 | 5 | 4/9 | 5/9 |

Kaplan-Meier estimate of the survival function at 9 months is given by product of $1-\frac{D_j}{N_j}$ for $t_j < 9$

which is $\frac{19}{20}.\frac{16}{19}.\frac{14}{16}.\frac{5}{9} = \frac{7}{18} = 0.3889$.

OR

Nelson-Aalen estimator

| $t_j$ | $N_j$ | $D_j$ | $C_j$ | $\frac{D_j}{N_j}$ | $\sum \frac{D_j}{N_j}$ |
|---|---|---|---|---|---|
| 0 | 20 | 0 | 0 | – | 0 |
| 1 | 20 | 1 | 0 | 1/20 | 0.0500 |
| 2 | 19 | 3 | 0 | 3/19 | 0.2079 |
| 5 | 16 | 2 | 5 | 2/16 | 0.3329 |
| 8 | 9 | 4 | 5 | 4/9 | 0.7773 |

Nelson-Aalen estimate of the survival function at 9 months is given by

$\exp\left(-\sum \frac{D_j}{N_j}\right)$ for $t_j < 9$

which is $\exp(-0.7773) = 0.4596$.

*Many candidates scored highly on this question. In part (i) the reason was needed for credit. Just mentioning the type of censoring without giving a reason was not awarded any marks. In part (ii) some indication of how the estimate was arrived at (normally a statement of the formula being applied) was needed for full credit. An impressive proportion of candidates performed the calculations correctly.*

## Question 5

(i) Objectives of the modelling exercise.
Validity of the model for the purpose to which it is to be put.
Validity of the data to be used.
Possible errors associated with the model or parameters used not being a perfect representation of the real world situation being modelled.
Impact of correlations between the random variables that "drive" the model.
Extent of correlations between the results produced from the model.
Current relevance of models written and used in the past.
Credibility of the data input.
Credibility of the results output.
Dangers of spurious accuracy.
Ease with which the model and its results can be communicated.
The time and cost of constructing and maintaining the model.

(ii) The model is capable of meeting the objective, specifically the estimation of transition intensities.

The model is valid for this purpose.

The data required are the total waiting times in each of the states Healthy and Sick for the lives in the investigation during the period of the investigation, together with the number of transitions from Healthy to Sick, from Sick to Healthy, from Healthy to Dead and from Sick to Dead.

Provided these data are available, the data will be valid for the application of the model.

The model is as good a representation of the real world process as we can obtain.

The model requires that we estimate constant intensities. The results will be credible provided we estimate the intensities separately for short age intervals, over which the assumption of constant transition intensities is credible.

The concept of transition intensities is not intuitively easy for non-specialists to understand.

The results can be made easier to understand and the results clearer by converting the transition intensities to probabilities – e.g. the probability that a Healthy life aged $x$ will make a sickness claim before he or she is aged $x+t$ years.

*Some candidates scored well on part (i), which was standard bookwork, but a disappointing number did not. Answers to part (ii) were variable. To score highly, the points made in part (ii) should relate to those made in part (i). Within this general criterion, sensible points other than those listed above were given credit. Full marks could be obtained for less than is given in the model solution above.*

## Question 6

(i) (a) EITHER

The parameters are the rate of leaving state $i$, $\lambda_i$, for each $i$, and the jump-chain transition probabilities, $r_{ij}$, for $j \neq i$, where $r_{ij}$ is the conditional probability that the next transition is to state $j$ given the current state is $i$.

OR

If the rate of leaving state $i$, is $\lambda_i$, and $r_{ij}$ is the conditional probability that the next transition is to state $j$ given the current state is $i$.

The parameters are $\mu_{ij,}$ where, for $i = j$, $\mu_{ii} = -\lambda_i$ and, for $i \neq j$, $\mu_{ij} = \lambda_i\, r_{ij}$.

OR

The parameters are the six transition rates from state $i$ to state $j$ ($i \neq j$):

$\mu_{AB}$

$\mu_{AC}$

$\mu_{BA}$

$\mu_{BC}$

$\mu_{CA}$

$\mu_{CB}$

(b) The assumptions are as follows.

EITHER The holding time in each state is exponentially distributed
OR The transition intensities from each state are not time-dependent.

The parameter of this distribution varies only by state $i$, so that the distribution is independent of anything that happened prior to the arrival in current state $i$.

The destination of the jump on leaving state $i$ is independent of holding time, and of anything that happened prior to the current arrival in state $i$.

(ii) (a) The estimator [it is the maximum likelihood estimator (MLE) but this need not be stated] of $\lambda_i$, $\hat{\lambda}_i$ , is the inverse of the average duration of each visit to state $i$.

so $\hat{\lambda}_A$ = 3 per hour, $\hat{\lambda}_B$ = 1/2 per hour, $\hat{\lambda}_C$ = 1/3 per hour

The estimator [it is the MLE but this need not be stated] of $r_{ij}$, $\hat{r}_{ij}$ , is the proportion of observed jumps out of state $i$ to state $j$.

$\hat{r}_{AB} = 60/105=4/7$

$\hat{r}_{AC} = 45/105=3/7$

$\hat{r}_{BA} =50/75=2/3$

$\hat{r}_{BC} =25/75=1/3$

$\hat{r}_{CA} =55/70=11/14$

$\hat{r}_{CB} =15/70=3/14$

(b) The estimated generator matrix (in $hr^{-1}$) is:

$$\begin{pmatrix} -3 & \frac{12}{7} & \frac{9}{7} \\ \frac{1}{3} & -\frac{1}{2} & \frac{1}{6} \\ \frac{11}{42} & \frac{1}{14} & -\frac{1}{3} \end{pmatrix}$$

(iii) EITHER Binomial, with mean $n.r_{ij}$ and variance $n.r_{ij}.(1 - r_{ij})$, $n$ being the number of transitions out of state $i$.

OR Binomial ($n$, $r_{ij}$) $n$ being the number of transitions out of state $i$.

*This was a relatively straightforward question, so the Examiners were looking for accurate and incisive answers. In part (i)(b) many candidates offered vague statements about the process not depending on past history. These candidates scored only limited credit for this part. In part (ii)(a) candidates who simply wrote down the values of the transition intensities, viz:*

$\mu_{AB} = 12/7$

$\mu_{AC} = 9/7$

$\mu_{BA} = 1/3$

$\mu_{BC} = 1/6$

$\mu_{CA} = 11/42$

$\mu_{CB} = 1/14$

*scored partial credit. Some candidates combined parts (ii)(a) and (b) by simply writing down the generator matrix. If this was correct, they were awarded most of the marks for this part, but for full marks some indication of how they arrived at the numbers in the generator matrix was needed. It was extremely disappointing how few candidates were able to state the distribution in part (iii): this seems to indicate a gap in knowledge of the subject.*

## Question 7

(i) Cox's model ensures that the hazard is always positive.

Standard software packages often include Cox's model.

Cox's model allows the general "shape" of the hazard function for all individuals to be determined by the data, giving a high degree of flexibility,

The data in this investigation are censored, and Cox's model can handle censored data.

In Cox's model the hazards of individuals with different values of the covariates are proportional, meaning that they bear the same ratio to one another at all ages.

If we are not primarily concerned with the precise form of the hazard, we can ignore the shape of the baseline hazard and estimate the effects of the covariates from the data directly.

(ii) A suitable statistical test is that using the likelihood ratio statistic.

We compare the model with gender + exercise with the model with gender + exercise + the interaction.

If the log-likelihood for these two models are $L$ and $L_{\text{interaction}}$ respectively, then the test statistic is $-2(L - L_{\text{interaction}})$.

This is equal to $-2\{-1{,}250 - (-1{,}246)\} = -2(-4) = 8$.

Under the null hypothesis that the parameter on the interaction term is zero, this statistic has a chi-squared distribution with one degree of freedom (since the interaction term involves one parameter).

Since 8 > 7.879, the critical value of the chi-squared distribution at the 0.5% level (or 8 > 3.84 for the 5% level),

we reject the null hypothesis even at the 99.5% level (or 95% level) and conclude that the interaction term is required in the model.

(iii) The baseline category is females who do not take regular exercise.

The hazards of developing heart disease in the other three categories, relative to the baseline category, are as follows:

| *Gender* | *Regular exercise* | |
|---|---|---|
| Male | No | $\exp(0.2) = 1.22$ |
| Male | Yes | $\exp(0.2 - 0.3 - 0.35) = 0.64$ |
| Female | Yes | $\exp(-0.3) = 0.74$ |

Males who do not take regular exercise are more likely to develop heart disease than females.

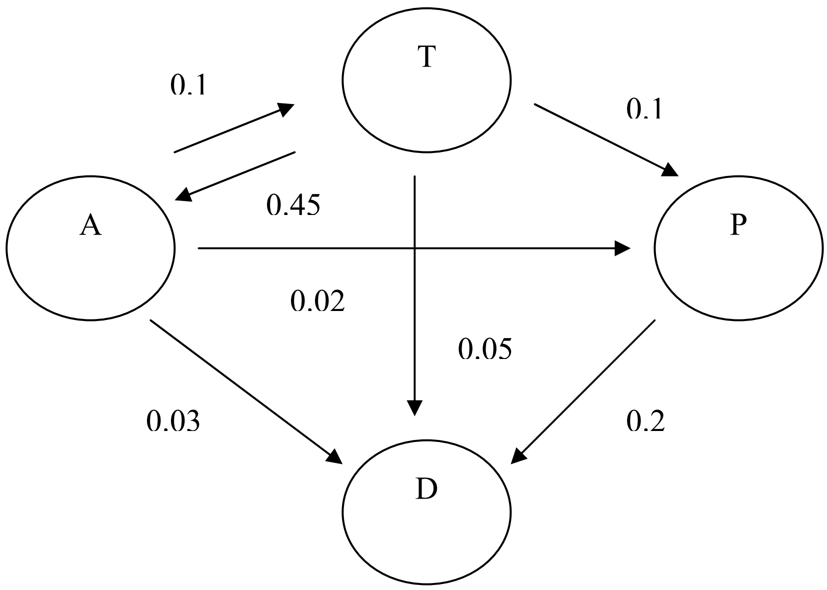
Regular exercise decreases the risk of heart disease for both males and females.

The effect of regular exercise in reducing the risk of heart disease is greater for males than for females, so much so that among those who take regular exercise, males have a lower risk of developing heart disease than females.

*There was a wide variation of performance among candidates on this question. Answers to part (i) suffered from wordiness and lack of precision, giving general descriptions of the model rather than focusing on its attractive qualities. Part (ii) was very well answered by* *many candidates. In part (iii) many candidates seemed not to understand the interpretation of the interaction term. For example, it was common to read that males had a higher risk of heart disease than females. However, this is only true for persons who do not take regular exercise. Among persons who do take regular exercise, females have a higher risk of heart disease than males.*

## Question 8

(i)




(ii) The force of leaving state $A$ is 0.15.

$$\frac{d}{dt}(P_{\overline{AA}}(t)) = -0.15P_{\overline{AA}}(t)$$
$$\frac{d}{dt}(\ln(P_{\overline{AA}}(t))) = -0.15$$
$$P_{\overline{AA}}(t) = \exp(-0.15t)$$

So the probability of staying in state $A$ for at least 5 years continuously is given by $\exp(-.75) = 0.472$.

(iii) (a) Conditioning on the first move out of $A$:

Probability 0.1/0.15 of moving to $T$, at which point probability becomes $F(T)$.

Probability 0.02/0.15 of moving to $P$, at which point certain to travel through state $P$.

Probability 0.03/0.15 of moving straight to $D$, at which point certain never to reach state $P$.

So $F(A) = 0.1/0.15*F(T)+0.02/0.15*0+0.03/0.15*1 = 2/3*F(T) + 1/5$.

(b) Similarly conditioning on first move out of $T$

Probability 0.45/0.6 to $A$ when probability becomes $F(A)$.

Probability 0.1/0.6 to $P$ when probability becomes 0.

Probability 0.05/0.6 to $D$ when probability becomes 1.

So $F(T) = 3/4 * F(A)+ 1/12$

(iv) Substituting for $F(T)$ in first equation:

$F(A) = 1/2*F(A)\ \ /18\ \ /5$

$F(A) = 23/45$

$F(T) = 7/15$

(v) Time spent in state $P$ from point of entry is exponentially distributed with rate 0.2,

so mean time spent in state $P$ from point of entry is 1/0.2 = 5 years.

So expected time spent in state $P$ for a person currently able to work is $(1 - F(A))*5 = 22/45*5 = 22/9$ years.

*Parts (i) and (ii) were well answered by most candidates. However, the majority of candidates struggled with parts (iii)–(v), many not attempting these sections. The rates were not required on the diagram in (i) for full credit. Alternative approaches to parts (iii) onwards are possible (for example involving geometric progressions) and were attempted by a few candidates. These approaches involve more complicated equations than the solution above and were rarely successfully completed.*

## Question 9

(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 deaths data at age $x$.

(ii) When the deaths data and the exposed to risk data come from different sources.

E.g. occupational mortality investigations where deaths data come from death registers and exposed to risk data from census
OR
where deaths data come from claims department of an office, whereas exposed to risk data are based on policies in force, which come from a different part of the office.

(iii) We need to adjust the exposed-to-risk to correspond to the age definition of deaths.

Let the population aged $x$ nearest birthday on 1 January in year $t$ be $P_{x,t}$.

A central exposed to risk for calendar year $t$ can be approximated by

$$E^c_{x,t} = \int_0^1 P_{x,t+s}ds \approx \frac{1}{2}(P_{x,t} + P_{x,t+1})$$

assuming that the population varies linearly over the calendar year.

Let $P_{x,t}{}^*$ be the population aged $x$ last birthday on 1 January in year $t$.

Then

$$P_{x,t} = \frac{1}{2}\left(P^*_{x,t} + P^*_{x-1,t}\right).$$

This assumes that birthdays are distributed evenly across the calendar year

If the number of deaths in year $t$ aged $x$ nearest birthday on the date of death is $\theta_{x,t}$,

then the required formula for estimating $\mu_{x+f,t}$ is thus

$$\mu_{x+f,t} = \frac{\theta_{x,t}}{\frac{1}{2}(P_{x,t} + P_{x,t+1})} = \frac{\theta_{x,t}}{\frac{1}{2}\left[\frac{1}{2}\left(P^*_{x,t} + P^*_{x-1,t}\right) + \frac{1}{2}\left(P^*_{x,t+1} + P^*_{x-1,t+1}\right)\right]}.$$

The age range at the start of the rate interval is $[x - 1, x]$, so the age range at the middle of the rate interval is $[x - ½, x + ½]$.

The average age at the middle of the rate interval is therefore $x$.

So $f = 0$.

(iv) Sex
Age
Marital status
Occupation
Socio-economic status
Ethnic origin
Educational attainment
Housing tenure
Disability, chronic health condition, limiting long-term illness

*In part (ii), candidates who stated that "different age definitions" are a reason why correspondence is difficult to achieve were given limited credit. If they went on to suggest* *that different age definitions can arise because the deaths data and the exposed-to-risk data come from different sources, and gave a relevant example, full credit was awarded. Many candidates, however, did not describe the different age definitions clearly. Part (iii) was better answered than have been exposed-to-risk questions in recent examination papers. In part (iv) "smoking behaviour" is NOT correct as a factor which a national statistical office might use to classify mortality, neither are factors such as "type of policy", "policy size" or "sales channel". Candidates are reminded to read the question carefully!*

## Question 10

(i) Outliers. Since all the information is summarised in one number, a few large deviations may be offset or hidden by a large number of small deviations.

Small bias. Since the squares of the differences are used, the sign of the differences are lost, hence small but consistent bias above or below may not be noticed.

Clumps or runs. Again because the squares of the differences are used, the sign of the differences are lost, so significant groups of (clumps or runs) of bias over ranges of the data may not be detected.

(ii) (a) A few large deviations or outliers – Individual Standardised Deviations Test.

Small but consistent bias – Signs Test OR Cumulative Deviations Test.

Clumps or runs of bias over ranges of the data - Grouping of Signs Test OR Serial Correlations Test.

(b) **Individual Standardised Deviations Test**

Under the null hypothesis that the standard table rates OR graduated rates are the true rates underlying the observed data

we would expect individual deviations to be distributed Normal (0,1).

EITHER only 1 in 20 $z_x$ should lie above 1.96 in absolute value
OR none should lie above 3 in absolute value
OR table (see below) showing split of deviations, actual versus expected.

| | $(-\infty,-2)$ | $(-2,-1)$ | $(-1,0)$ | $(0,1)$ | $(1,2)$ | $(2,+\infty)$ |
|---|---|---|---|---|---|---|
| Expected | 0.22 | 1.54 | 3.74 | 3.74 | 1.54 | 0.22 |
| Observed | 0 | 0 | 5 | 4 | 1 | 1 |

The largest deviation we have here is 3.31.

This is well outside the range −1.96 to 1.96 so we have reason to reject the null hypothesis.

**EITHER Signs Test OR Cumulative Deviations Test**

**Signs Test**

Under the null hypothesis that the standard table rates OR graduated rates are the true rates underlying the observed data

The number of positive signs amongst the $z_x$ is distributed Binomial (11, ½ )

We observe 6 positive signs.

EITHER the probability of observing 6 or more positive signs in 11 observations is 0.5
OR the probability of observing exactly 6 positive signs is 0.2256.

which implies that Pr[observing 6 or more] > 0.025 (a two-tailed test),

so we have no evidence to reject the null hypothesis.

**Cumulative Deviations Test**

Under the null hypothesis that the standard table rates OR graduated rates are the true rates underlying the observed data,

$$\text{the test statistic } \frac{\sum_x (\text{Observed deaths - Expected deaths})}{\sqrt{\sum_x \text{Expected deaths}}} \sim \text{Normal}(0,1)$$

The calculations are shown in the table below.

| *Age x* | *Expected deaths* | *Observed – expected deaths* |
|---|---|---|
| 60 | 36.15 | –1.15 |
| 61 | 28.92 | –4.92 |
| 62 | 31.34 | –4.34 |
| 63 | 38.01 | –3.01 |
| 64 | 26.88 | 5.12 |
| 65 | 37.59 | –1.59 |
| 66 | 33.85 | 0.15 |
| 67 | 26.66 | 5.34 |
| 68 | 22.37 | 3.63 |
| 69 | 18.69 | 14.31 |
| 70 | 18.24 | 3.76 |
| Totals | 318.70 | 17.30 |

The value of the test statistic is $\dfrac{17.30}{\sqrt{318.70}} = 0.969$

and, since – 1.96 < test statistic < +1.96 we have insufficient evidence to reject the null hypothesis.

**EITHER Grouping of SignsTest OR Serial Correlations Test**

**Grouping of Signs Test**

Under the null hypothesis that the standard table rates OR the graduated rates are the true rates underlying the observed data

$G$ = Number of groups of positive deviations = 2

$m$ = number of deviations = 11
$n_1$ = number of positive deviations = 6
$n_2$ = number of negative deviations = 5

THEN EITHER

We want $k^*$ the largest $k$ such that

$$\sum_{t=1}^{k} \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}} < 0.05$$

The test fails at the 5% level if $G \le k^*$.

From the Gold Book $k^* = 1$.

So we have insufficient evidence to reject the null hypothesis.

OR

For $t = 2$

$$\binom{n_1-1}{t-1} = \binom{5}{1} = 5 \quad \text{and} \quad \binom{n_2+1}{t} = \binom{6}{2} = 15$$

and $\binom{m}{n_1} = \binom{11}{6} = 462$

So Pr[$t = 2$] if the null hypothesis is true is 75/462 = 0.162, which is greater than 5% so we have insufficient evidence reject the null hypothesis.

**Serial Correlations Test (lag 1)**

Under the null hypothesis that the standard table rates OR graduated rates are the true rates underlying the observed data.
The calculations are shown in the tables below.

EITHER USING SEPARATE MEANS FOR THE $z_x$ AND $z_{x+1}$

| Age | $z_x$ | $z_x$ | $A = z_x - \bar{z}$ | $B = z_{x+1} - \bar{z}$ | $AB$ | $A^2$ | $B^2$ |
|---|---|---|---|---|---|---|---|
| 60 | –0.191 | –0.915 | –0.541 | –1.372 | 0.742 | 0.293 | 1.881 |
| 61 | –0.915 | –0.775 | –1.264 | –1.232 | 1.558 | 1.599 | 1.518 |
| 62 | –0.775 | –0.488 | –1.125 | –0.945 | 1.063 | 1.265 | 0.893 |
| 63 | –0.488 | 0.988 | –0.838 | 0.531 | –0.445 | 0.702 | 0.282 |
| 64 | 0.988 | –0.259 | 0.638 | –0.716 | –0.457 | 0.407 | 0.513 |
| 65 | –0.259 | 0.026 | –0.609 | –0.431 | 0.262 | 0.371 | 0.186 |
| 66 | 0.026 | 1.034 | –0.324 | 0.577 | –0.187 | 0.105 | 0.333 |
| 67 | 1.034 | 0.767 | 0.685 | 0.311 | 0.213 | 0.469 | 0.097 |
| 68 | 0.767 | 3.310 | 0.418 | 2.853 | 1.192 | 0.175 | 8.141 |
| 69 | 3.310 | 0.880 | 2.960 | 0.424 | 1.254 | 8.764 | 0.179 |
| 70 | 0.880 | | 0.531 | | | | |
| $\bar{z}$ | 0.350 | 0.457 | | Average | 0.520 | 1.415 | 1.402 |

0.520/√(1.415*1.402) = 0.369.

Test 0.369 (√11) = 1.223 against Normal (0,1), and, since 1.223 < 1.645, we do not reject the null hypothesis.

OR USING THE FORMULA IN THE GOLD BOOK

| Age | $z_x$ | $z_x$ | $A = z_x - \bar{z}$ | $B = z_{x+1} - \bar{z}$ | $AB$ | $A^2$ |
|---|---|---|---|---|---|---|
| 60 | –0.191 | –0.915 | –0.589 | –1.313 | 0.773 | 0.347 |
| 61 | –0.915 | –0.775 | –1.313 | –1.173 | 1.540 | 1.723 |
| 62 | –0.775 | –0.488 | –1.173 | –0.886 | 1.039 | 1.376 |
| 63 | –0.488 | 0.988 | –0.886 | 0.590 | –0.523 | 0.785 |
| 64 | 0.988 | –0.259 | 0.590 | –0.657 | –0.388 | 0.348 |
| 65 | –0.259 | 0.026 | –0.657 | –0.372 | 0.245 | 0.432 |
| 66 | 0.026 | 1.034 | –0.372 | 0.636 | –0.237 | 0.138 |
| 67 | 1.034 | 0.767 | 0.636 | 0.370 | 0.235 | 0.432 |
| 68 | 0.767 | 3.310 | 0.370 | 2.912 | 1.076 | 0.137 |
| 69 | 3.310 | 0.880 | 2.912 | 0.483 | 1.405 | 8.481 |
| 70 | 0.880 | | 0.483 | | | 0.233 |
| $\bar{z}$ | 0.350 | 0.457 | | Sum | 0.517 | 1.310 |

$$\frac{\frac{1}{10}(5.617)}{\frac{1}{11}(14.405)} = 0.395\,.$$

Test 0.395 ($\sqrt{11}$) = 1.309 against Normal (0,1), and, since 1.309 < 1.645, we do not reject the null hypothesis.

(iii) The result of the Individual Standard Deviation test suggests outliers in the data.

The actual and expected deaths are relatively low, suggesting that the population in the rural area is not very large.

The ages under consideration are also high, exacerbating this scarcity of data.

However there are at least five (actual/expected) deaths in each age group, so the data are adequate.

So this is unlikely to account for the outlier at age 69 years, which should be investigated further.

The period of the observation is not stated and could affect the results, as, for example if the observation only covered one winter a particularly bad influenza epidemic may have caused more deaths than usual (although this would likely impact all ages in this range similarly).

Both the signs and grouping of signs test suggest no bias over the whole or part of the data.

However there does seem to be a drift towards the number of observed deaths exceeding the expected at higher ages, and the number observed being smaller than expected at younger ages.

Perhaps if a larger extract from the investigation were considered or the table in its entirety, bias may be observed.

*Answers to this question were disappointing. Too many answers to part (i) were sketchy and failed to explain WHY the chi-squared test sometimes fails to detect small bias, outliers or "runs" of deviations of the same sign. In part (ii) some candidates failed to relate the tests they were performing to the deficiencies of the chi-squared test identified in part (i); other candidates performed two tests for the same deficiency (only the higher scoring of which received credit). Many candidates lost marks for vagueness in the execution of the tests. Although not all the points listed above were required in part (iii) for full credit, the number of marks available indicated that candidates were expected to go beyond the basic results of the tests. Disappointingly few did this.*

# Question 11

(i) Transitions from state "Zero"

No umbrellas to take so must be two at the other location.

Transitions from state "One"

If it does not rain, then there remains one at each location, probability $1-p$.

If it does rain, both umbrellas end up at the next destination, probability $p$.

Transitions from state "Two"

If it does not rain, then forgets to take an umbrella so none is at the next location, probability $1-p$.

If it does rain, takes one of the umbrellas to the other location, probability $p$.

(ii) One step transition matrix is:

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{pmatrix}$$

Seeking the two-step transition matrix as the square of this matrix:

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{pmatrix} . \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{pmatrix} = \begin{pmatrix} 1-p & p & 0 \\ p(1-p) & (1-p)^2+p^2 & p(1-p) \\ 0 & p(1-p) & 1-p+p^2 \end{pmatrix}$$

(iii) $$\Pi \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{pmatrix} = \Pi$$

$(1-p)\pi_3 = \pi_1$ (I)

$(1-p)\pi_2 + p\pi_3 = \pi_2$ or $\pi_2 = \pi_3$ (II)

$\pi_1 + p\pi_2 = \pi_3$ (III)

and

$\pi_1 + \pi_2 + \pi_3 = 1$ (IV)

$((1-p)+1+1)\pi_3 = 1$

$$\pi_2 = \pi_3 = \frac{1}{3-p}$$

$$\pi_1 = \frac{1-p}{3-p}$$

(iv) He gets wet if it rains on a journey when he is state "Zero".

So the long run probability is $p.\pi_1 = \dfrac{p(1-p)}{3-p}$.

(v) Denoting $R$ = raining, $NR$ = not raining

| *From / To* | *R* | *NR* |
|---|---|---|
| $R$ | $r$ | $1-r$ |
| $NR$ | $s$ | $1-s$ |

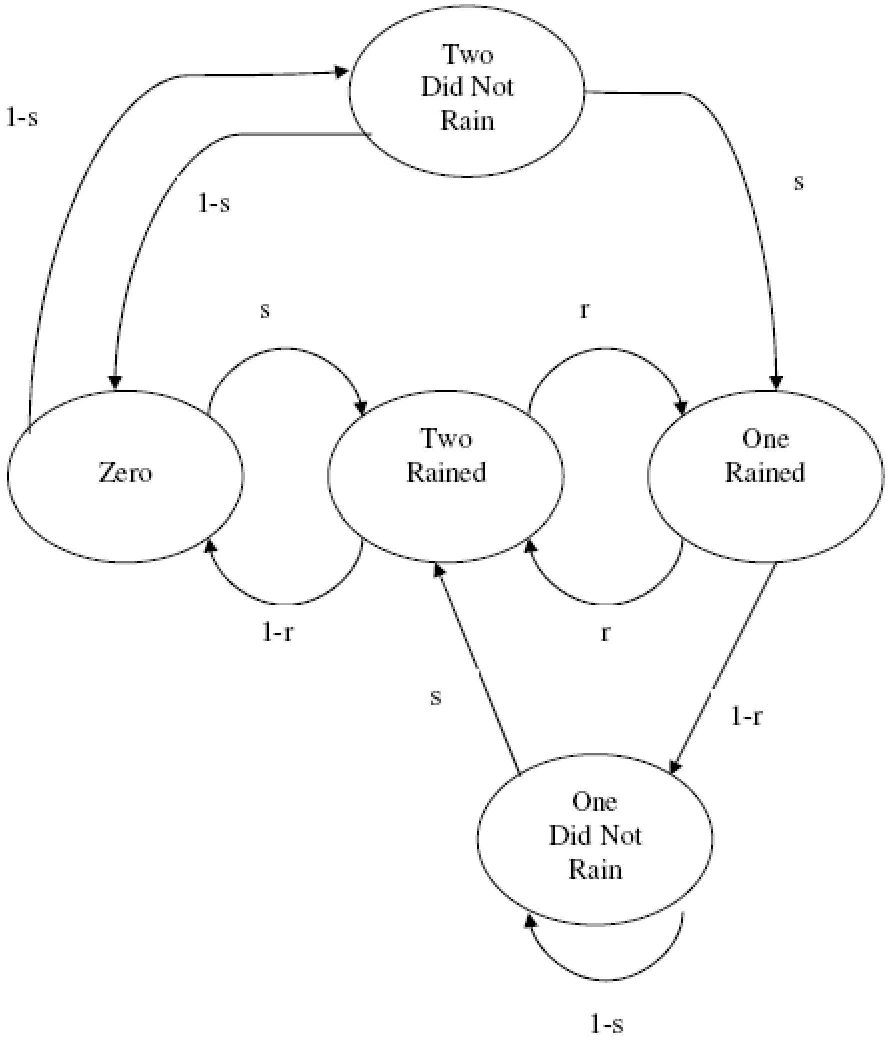
(vi) This would not satisfy the Markov property because (in states "One" and "Two") would need to know, in addition, whether it was raining or not on the last journey to determine the future evolution of the process.

e.g. if in state "Two", probability of next moving to "Zero" is 1-$r$ if it rained on the last journey and 1−$s$ if it did not. As $r$ does not equal $s$ the Markov property is not satisfied.

(vii) If we expand the states to include information about whether it rained on the last journey, then the Markov property is satisfied.

Five states are needed, as cannot be in position with zero umbrellas when it rained on last journey,

so the state space is {Zero, One Rained, One Did Not Rain, Two Rained, Two Did Not Rain}




*Many candidates scored highly on parts (i)–(iii) of this question, but a much smaller proportion made a solid effort at parts (iv)–(vii). In part (vi), candidates who simply said that the process would not satisfy the Markov property because it depended on the "past history" scored only limited credit. For full credit, it was necessary to say that what matters is whether it was raining or not on the last journey, and to give an example of transitions with differing probabilities. In part (vii), some candidates produced four-state solutions, splitting either of states One or Two, but not both. These candidates were given credit for diagrams correct for the solution they were offering.*

## END OF EXAMINERS' REPORT

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINATION

18 April 2012 (am)

## Subject CT4 – Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 12 questions, beginning your answer to each question on a separate sheet.*

5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** (i) Define a general random walk. [1]

(ii) State the conditions under which a general random walk would become a simple random walk. [1]
[Total 2]

**2** (i) Explain the reasons why data are subdivided when conducting mortality investigations. [2]

(ii) Describe the problems which can arise with subdividing data. [2]
[Total 4]

**3** A graduation of a set of crude mortality rates is tested for goodness-of-fit using a chi-squared test.

Discuss the factors to be considered in determining the number of degrees of freedom to use for the test statistic.
[4]

**4** A new drug treatment for patients suffering from a chronic skin disease with visible symptoms was tested. The drug was administered through a daily dose for the duration of the trial. As soon as the drug regime started, the symptoms disappeared in all patients, but after some time had a tendency to reappear as the agent causing the disease developed resistance to the drug. The trial lasted for six months.

The data below show the number of patients experiencing a return of their symptoms in each month after the drug regime started.

| *Month* | *Number of patient-months exposed to risk* | *Number of patients experiencing a return of their symptoms* |
|---|---|---|
| 1 | 200 | 5 |
| 2 | 190 | 8 |
| 3 | 175 | 15 |
| 4 | 150 | 10 |
| 5 | 135 | 6 |
| 6 | 125 | 3 |

(i) Calculate the hazard of symptoms returning in each month. [2]

As part of the investigation, it is desired to assess the impact of certain risk factors on the hazard of symptoms returning. It is suggested that to achieve this, the hazard could be modelled using either a Gompertz model or a semi-parametric model.

(ii) Comment on the use of each of these models in this situation. [4]
[Total 6]

**5** For a particular investigation the hazard of mortality is assumed to take the form:

$$h(t) = A + Bt$$

where $A$ and $B$ are constants and $t$ represents time.

For each life $i$ in the investigation ($i = 1, \ldots, n$) information was collected on the length of time the life was observed $t_i$ and whether the life exited due to death ($\delta_i = 1$ if the life died, 0 otherwise).

(i) Show that the likelihood of the data is given by:

$$L = \prod_{i=1}^{n} (A + Bt_i)^{\delta_i} \exp[-At_i - \frac{1}{2}Bt_i^2].$$ [3]

(ii) Derive two simultaneous equations from which the maximum likelihood estimates of the parameters $A$ and $B$ can be calculated. [3]
[Total 6]

**6** (i) List the advantages and disadvantages of using models in actuarial work. [4]

A new town is planned in a currently rural area. A model is to be developed to recommend the number and size of schools required in the new town. The proposed modelling approach is as follows:

- The current age distribution of the population in the area is multiplied by the planned population of the new town to produce an initial population distribution.
- Current national fertility and mortality rates by age are used to estimate births and deaths.
- The births and deaths are applied to the initial population distribution to generate a projected distribution of the town's population by age for each future year, and hence the number of school age children.

(ii) Discuss the appropriateness of the proposed modelling approach. [5]
[Total 9]

**7** Mr Bunn the baker made 12 pies to sell in his shop. He placed the pies in the shop at 9 a.m. During the rest of the day the following events took place.

| *Time* | *Event* |
|---|---|
| 10 a.m. | A boy bought two pies |
| 11 a.m. | A man bought three pies |
| 12 noon | Mr Bunn accidentally sat on one pie and squashed it so it could not be sold |
| 1 p.m. | A woman bought two pies |
| 2 p.m. | A dog from across the street ran into Mr Bunn’s shop and stole two pies |
| 3 p.m. | A girl on the way home from school bought one pie |
| 5 p.m. | Mr Bunn closed for the day and the remaining pie was still in the shop |

(i) Estimate the time it takes Mr Bunn to sell 40% of the pies he makes, using the Nelson-Aalen estimator. [6]

(ii) Comment on whether you think this estimate would be a good basis for Mr Bunn to plan his future production of pies. [3]
[Total 9]

**8** The mortality experience of a large company pension scheme is to be tested to see if the experience of males aged 65–72 years is consistent with a standard table. The results were collated by the firm conducting the analysis on a computer spreadsheet, with positive and negative standardised deviations being distinguished only by being in a different coloured font. Unfortunately the results have been supplied to the company in the form of a printout produced on a black-and-white printer from which it is not possible to tell the signs of the deviations.

The values of the standardised deviations shown are as follows:

0.052
0.967
2.528
0.328
1.234
0.250
1.023
0.756

(i) Suggest two tests which could be conducted from the information given. [2]

(ii) Carry out the tests you suggested in your answer to part (i). [8]
[Total 10]

**9** (i) List four factors other than age and smoker status by which life insurance mortality statistics are often subdivided. [2]

Two offices in different towns of the same life insurance company write 25-year term assurance policies. Below are data from these two offices relating to policyholders of the same age. Both deaths and policies in force are on an age last birthday basis.

| | *Gasperton* | *Great Hawking* |
|---|---|---|
| Policies in force on 1 January 2009 | 2,000 | 1,770 |
| Policies in force on 1 January 2010 | 2,100 | 1,674 |
| Deaths in calendar year 2009 | 25 | 21 |

(ii) Calculate the central death rate for the calendar year 2009 at this age for the offices in Gasperton and Great Hawking. [2]

A detailed examination of the records shows that 50% of the policyholders in Gasperton at both censuses were smokers, and 20% of policyholders in Great Hawking at both censuses were smokers. National death rates at this age for smokers in 2009 were 40% higher than those for non-smokers.

(iii) Estimate the central death rates for smokers and non-smokers in Gasperton and Great Hawking. [4]

The life insurance company charges policyholders in Gasperton and Great Hawking the same premiums for the 25-year term assurance policies. It charges smokers in both towns 40% more than non-smokers.

(iv) Comment on the company's pricing structure in the light of your results from parts (ii) and (iii) above. [3]

[Total 11]

**10** An investigation was conducted into the effect marriage has on mortality and a model was constructed with three states: 1 Single, 2 Married and 3 Dead. It is assumed that transition rates between states are constant.

(i) Sketch a diagram showing the possible transitions between states. [2]

(ii) Write down an expression for the likelihood of the data in terms of transition rates and waiting times, defining all the terms you use. [3]

The following data were collected from information on males and females in their thirties.

| | |
|---|---|
| Years spent in Married state | 40,062 |
| Years spent in Single state | 10,298 |
| Number of transitions from Married to Single | 1,382 |
| Number of transitions from Single to Dead | 12 |
| Number of transitions from Married to Dead | 9 |

(iii) Derive the maximum likelihood estimator of the transition rate from Single to Dead. [4]

(iv) Estimate the constant transition rate from Single to Dead and its variance. [2]
[Total 11]

**11** The series $Y_i$ records, for each time period $i$, whether a car driver is accident free during that period ($Y_i = 0$) or has at least one accident ($Y_i = 1$).

Define $X_i = \sum_{j=1}^{i} Y_j$ with state space $\{0,1,2,\ldots.\}$.

An insurer makes an assumption about the driver's accident proneness by considering that the probability of a driver having at least one accident is related to the proportion of previous time periods in which the driver had at least one accident as follows:

$$P(Y_{n+1} = 1) = \frac{1}{4}\left(1 + \frac{X_n}{n}\right), \quad \text{for} \quad n \geq 1$$

with $P(Y_1 = 1) = \dfrac{1}{2}$

(i) Demonstrate that the series $X_i$ satisfies the Markov property, whilst $Y_i$ does not. [2]

(ii) Explain whether the series $X_i$ is:

(a) irreducible
(b) time homogeneous

[3]

(iii) Draw the transition graph for $X_i$ covering all transitions which could occur in the first three time periods, including the transition probabilities. [4]

(iv) Calculate the probability that the driver has accidents during exactly two of the first three time periods. [2]

(v) Comment on the appropriateness of the insurer's assumption about accident proneness. [2]
[Total 13]

**12** A company operates a sick pay scheme as follows:

- Healthy employees pay a percentage of salary to fund the scheme.
- For the first two consecutive months an employee is sick, the sick pay scheme pays their full salary.
- For the third and subsequent consecutive months of sickness the sick pay is reduced to 50% of full salary.

To simplify administration the scheme operates on whole months only, that is for a particular month's payroll an employee is either healthy or sick for the purpose of the scheme.

The company's experience is that 10% of healthy employees become sick the following month, and that sick employees have a 75% chance of being healthy the next month.

The scheme is to be modelled using a Markov Chain.

(i) Explain what is meant by a Markov Chain. [1]

(ii) Identify the minimum number of states under which the payments under the scheme can be modelled using a time homogeneous Markov Chain, specifying these states. [2]

(iii) Draw a transition graph for this Markov chain. [2]

(iv) Derive the stationary distribution for this process. [4]

(v) Calculate the minimum percentage of salary which healthy employees should pay for the scheme to cover the sick pay costs. [2]

(vi) Calculate the contributions required if, instead, sick pay continued at 100% of salary indefinitely. [2]

(vii) Comment on the benefit to the scheme of the reduction in sick pay to 50% from the third month. [2]
[Total 15]

**END OF PAPER**

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS' REPORT

April 2012 examinations

## 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 who are 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. Although Examiners have access to the Core Reading, which is designed to interpret the syllabus, the Examiners are not required to examine the content of Core Reading. Notwithstanding that, the questions set, and the following comments, will generally be based on Core Reading.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report. Other valid approaches are always given appropriate credit; where there is a commonly used alternative approach, this is also noted in the report. For essay-style questions, and particularly the open-ended questions in the later subjects, this report contains all the points for which the Examiners awarded marks. This is much more than a model solution – it would be impossible to write down all the points in the report in the time allowed for the question.

T J Birse
Chairman of the Board of Examiners

July 2012

## General comments on Subject CT4

Subject CT4 comprises five main sections: (1) 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); (2) stochastic processes, especially Markov chains and Markov jump processes; (3) models of a random variable measuring future lifetime; (4) the calculation of exposed to risk and the application of the principle of correspondence; (5) 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.

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.

## Comments on the April 2012 paper

The general performance was better than in any session since April 2008. Well-prepared candidates scored highly across the whole paper. A feature of this diet was that parts of questions which required an element of explanation or interpretation (such as Q7(ii) and Q12(vii)) were better answered than in previous diets, and this largely accounted for the increased pass rate. The comments that follow the questions 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.

**1**

(i) $$X_n = \sum_{j=1}^{n} Y_j$$

where the $Y_j$ are i.i.d. random variables and $X_0 = 0$

(ii) Is simple random walk when $Y_j$ can have values +1 and −1 only

*In part (i) few candidates gave the initial condition that $X_0=0$. Many candidates were confused as to the definitions of general, simple, and symmetric random walks (for example defining a simple random walk in part (i) and then stating for part (ii) that the probabilities of $Y_j$ being +1 and -1 were both equal to 0.5).*

**2**

(i) Users of data require rates subdivided by age and other criteria.

Models are based on the assumption that we can observe groups of identical lives.

Therefore it is important that we analyse groups of lives which are homogenous (or have the same mortality).

This can, for example, help avoid anti-selection.

(ii) Small numbers in some sub-groups leading to scanty data and non-credible rates or a large variance.

Sometimes relevant factors cannot be used because the relevant information cannot be collected on the proposal form because questions are unlikely to be answered honestly,

or because the key questions are intrusive or impractical for marketing or administrative reasons or make the questionnaire too long, or cannot be asked by law.

Can be difficult to ensure that events data and exposed-to-risk data are subdivided in the same way, leading to the principle of correspondence being violated.

*Answers to this question were disappointing, even though not all the points listed above were required for full credit. In part (ii), many candidates made only the first point, about sparse data. Some candidates approached this question as practitioners or users of data rather than giving the general principles for which the question was asking. Nevertheless, if good points were made, this approach could earn full credit.*

## 3

To work out the number of degrees of freedom (d.f.) we start with the number of age groups.

We reduce the d.f. because of the constraints imposed by the graduation process.

The reduction varies according to the graduation method:

parametric formula – one d.f. lost for each parameter estimated;

standard table – one d.f. lost for each parameter fitted and a further reduction due to the constraints imposed by the choice of standard table;

graphical – two or three d.f. lost for about every 10 ages graduated.

*This question was generally well answered. Common errors were to suppose that only one d.f. is lost for the choice of standard table, and that for graphical graduation, two or three d.f. were lost in total, regardless of the number of ages being graduated.*

## 4

(i)

| | |
|---|---|
| Month 1 | 5/200 = 0.025 |
| Month 2 | 8/190 = 0.042 |
| Month 3 | 15/175 = 0.086 |
| Month 4 | 10/150 = 0.067 |
| Month 5 | 6/135 = 0.044 |
| Month 6 | 3/125 = 0.024 |

(ii) To assess the impact of risk factors, a proportional hazards model would be useful because of its simple interpretation or because it allows the effect of each individual risk factor to be assessed.

The Gompertz model can be framed as a proportional hazards model, as can a semi-parametric model (such as the Cox model).

The Gompertz model would not be appropriate here, as it has a monotonically increasing or decreasing hazard,

whereas it is clear from part (i) that the hazard of symptoms returning first rises and then falls with duration.

A semi-parametric model allows the shape of the hazard to be determined by the data.

The semi-parametric model would be better than the Gompertz in this case.

*In part (i) a minority of candidates subtracted half the deaths from the exposed-to-risk. Partial credit was given for this. Part (ii) was a higher skills question, and was poorly attempted by many candidates. Only a small proportion related their answers to the data* *given and spotted that the empirical hazard calculated in part (i) was non-monotonic and so the Gompertz model would be a poor fit. Hardly any candidates pointed out that the Gompertz model can be framed as a proportional hazards model.*

**5**

(i) The likelihood of the data is given by:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i},$$

where $f(t_i)$ is the probability density function and $S(t_i)$ is the survivor function.

Since $f(t_i)$ is related to the hazard function by

$f(t_i) = h(t_i)\, S(t_i)$

the likelihood can be rewritten:

$$L = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i).$$

Since

$$S(t_i) = \exp\left[ -\int_{r=o}^{t_i} h(r)dr \right] = \exp[-At_i - \frac{1}{2}Bt_i^{\,2}],$$

$$L = \prod_{i=1}^{n} (A + Bt_i)^{\delta_i} \exp[-At_i - \frac{1}{2}Bt_i^2] \text{ as required.}$$

(ii) The log likelihood is given by:

$$\log L = \sum_{i=1}^{n} \left[ \delta_i \log(A + Bt_i) - At_i - \frac{1}{2}Bt_i^{\,2} \right].$$

We are trying to maximise likelihood with respect to two parameters, so need partial differentials with respect to $A$ and $B$:

$$\frac{\partial}{\partial A} \log L = \sum_{i=1}^{n} \left[ \frac{\delta_i}{A + Bt_i} - t_i \right],$$

$$\frac{\partial}{\partial B}\log L = \sum_{i=1}^{n}\left[\frac{\delta_i t_i}{A+Bt_i} - \frac{1}{2}t_i^2\right].$$

The simultaneous equations satisfied by the MLEs are obtained by setting these to zero:

$$\sum_{i=1}^{n}\left[\frac{\delta_i}{A+Bt_i} - t_i\right] = 0,$$

$$\sum_{i=1}^{n}\left[\frac{\delta_i t_i}{A+Bt_i} - \frac{1}{2}t_i^2\right] = 0.$$

*In part (i) many candidates failed to explain where the components of the likelihood came from by explaining the different contributions of the lives who were observed to die and those who were not. In part (ii) credit was given for knowing the correct method even if this was not executed. Credit was also given for differentiating a second time and showing that the second derivatives were negative (and hence that we do have maxima), even though this was not required for full marks.*

## 6

(i) **Benefits**

Systems with long time frames can be studied in compressed time

Complex systems with stochastic elements can be studied (especially by simulation modelling).

Different future policies or possible actions can be compared either to see which best suits the requirements of a user or to examine different scenarios without carrying them out in practice, or to avoid potential costs associated with trialing in real life.

Models allow control over experimental conditions, so that we can reduce the variance of the results output without upsetting the mean values.

Parameters can be sensitivity tested using a model.

**Limitations**

Model development requires a lot of time and expertise, and hence can be /costly.

May need to run model lots of times (essential if it is a stochastic model).

Models more useful for comparing the results of input variations than for optimising outputs.

Models can look impressive, but can lull the user into a false sense of security. Impressive output is not a substitute for validity and close imitation of the real world.

This is more true the further into the future you project

Models rely heavily on the data input. If this is poor or lacking in credibility the output is likely to be flawed.

Users need to understand the model sufficiently well to be able to know when it is appropriate to apply it.

Interpretation of models can be difficult, and often outputs need to be seen in relative rather than absolute terms.

Models cannot take into account all possible future events (e.g. changes in legislation).

(ii) The model should be simple to apply.

The data specified are likely to be available from reliable sources.

Although it is possible that the starting point for the planned population may be wrong

Unforeseen events may take place such as a national epidemic which change the rates.

The model is relatively straightforward to explain to the planners/developers.

Should consider whether there are trends in fertility rates, rather than simply using current rates.

Mortality rates unlikely to be significant relative to the uncertainty in the projection,

because rates at ages with non-zero fertility rates should be small and child mortality rates should be low.

Current age distribution for the area may not be representative of that for the new town as, for example, rural areas may have different distributions to urban areas

Consider the type of houses being built and how they are marketing e.g. are they family houses?

May wish to consider experience of similar new towns.

May wish to consider whether national fertility rates are appropriate for a new town, where many young families may live.

Migration may affect the profile of the population, for example older families moving away and younger families buying their houses may mean the age structure remains relatively constant over time regardless of mortality and fertility rates.

The approach does not take account of non-state schooling or the possibility of children going to boarding school.

*Part (i) of the question was standard bookwork and was well answered. The quality of answers to part (ii) varied: some candidates wrote lengthy and well-argued discussions; others made only cursory attempts. In part (ii), not all the points listed above were needed for full credit, and other sensible comments could also score marks.*

**7**

(i) The sequence of events described may be summarised in the table below

| *Duration* $t_j$ | *Pies in shop* $n_j$ | *Pies bought* $d_j$ | *Pies destroyed or stolen,* $c_j$ |
|---|---|---|---|
| 1 | 12 | 2 | 0 |
| 2 | 10 | 3 | 0 |
| 3 | 7 | 0 | 1 |
| 4 | 6 | 2 | 0 |
| 5 | 4 | 0 | 2 |
| 6 | 2 | 1 | 0 |

The hazard of pies being bought is thus

| | |
|---|---|
| 2/12 at duration | 1 hour |
| 3/10 at duration | 2 hours |
| 2/6 at duration | 4 hours |
| 1/2 at duration | 6 hours |

The Nelson-Aalen estimate of the survival function, $S(t)$, is then

| *Duration* | *Nelson-Aalen estimate of S(t)* |
|---|---|
| $0 \le t < 1$ | 1 |
| $1 \le t < 2$ | exp [-2/12] = 0.8465 |
| $2 \le t < 4$ | exp [-(2/12 + 3/10)] = 0.6271 |
| $4 \le t < 6$ | exp [-(2/12 + 3/10 + 2/6)] = 0.4493 |
| $6 \le t < 8$ | exp [-(2/12 + 3/10 + 2/6 + 1/2)] = 0.2725 |

The Nelson-Aalen estimate is a step function.

We need $t$ for which $S(t) = 0.6$.

Therefore it will be 4 hours until Mr Bunn has sold 40% of his pies.

(ii) The estimate would not be a good basis on which to plan future production.

And how long it takes to sell 40% of your goods is not very relevant for future production.

It is based on only one day's experience, and a good basis for future production should be based on several days, probably involving different days of the week.

Sales of pies may vary seasonally: data from a winter's day may tell Mr Bunn little about the demand for pies in summer.

Mr Bunn might be more careful in future not to sit on his pies, and might take steps to avoid the dog from across the street stealing pies.

The proportion of pies sold will depend on the number of pies Mr Bunn stocks. He should not assume if he had twice as many pies he would still sell 40% of them in 4 hours.

Mr Bunn may vary his sales strategy, by, for example, reducing his prices

The method does, however, take account to of censored data.

*In part (i) the question said "estimate", so some indication of how the answer was arrived at was necessary, although not every detail was required. As a bare minimum full credit could be obtained for first three hazards at times 1, 2 and 4, some statement of what the Nelson-Aalen estimate of S(t) is, the fact that we are looking for S(t) = 0.6, and some numbers to demonstrate that S(t) = 0.6 happens at duration 4 hours. Answers that used the logarithm of S(t) were acceptable. Answers to part (ii) were encouraging. A substantial proportion of candidates made sensible points.*

## 8

(i) Chi-squared test (for overall goodness of fit)

(Modified) individual standardised deviations test (for outliers)

(ii) **Chi-squared test**

The null hypothesis is that the mortality among the members of the company's pension scheme is represented by the standard table.

The test statistic is $\sum_x {z_x}^2$, where the $z_x$ are the standardised deviations.

Under the null hypothesis, this statistic has a chi-squared distribution with 8 degrees of freedom.

$$\sum_x {z_x}^2 = 0.052^2 + 0.967^2 + 2.528^2 + 0.328^2 + 1.234^2 + 0.250^2$$
$$+ 1.023^2 + 0.756^2 = 10.64.$$

The critical value of the chi-squared distribution with 8 degrees of freedom at the 5% significance level is 15.51.

Since 10.64 < 15.51

we do not reject the null hypothesis.

**(Modified) individual standardised deviations test**

Under the null hypothesis (same as for the chi-squared test)

we would expect individual deviations to be distributed Normal (0,1)

Only 1 in 20 of the $z_x$ should lie above 1.96 in absolute value
OR
none should lie above 3 in absolute value
OR
about two thirds of the $z_x$ should lie between −1 and +1
OR

| Interval | (0,1) | (1,2) | (2,∞) |
|---|---|---|---|
| Actual deaths | 5 | 2 | 1 |
| Expected deaths | 5.44 | 2.24 | 0.32 |

The largest deviation we have here is 2.528 in absolute value,

which is well outside the range −1.96 to +1.96,

therefore we have reason to reject the null hypothesis.

but, since we have 8 ages we cannot say definitively whether the null hypothesis should be rejected, but the large deviation of 2.528 suggests there may be a problem.

*Many candidates scored highly on this question, though the chi-squared test was generally better done than the individual standardised deviations test. A surprising proportion of candidates thought that it was possible to perform the serial correlations test with the data given. The most common errors were to reduce the number of degrees of freedom in the chi-squared test (incorrect here as we are not testing a graduation) and a failure to spot the large deviation of 2.528, and state that this is a source of concern.*

**9**

(i) Gender

Type of policy

Level of underwriting

Duration in force

Sales channel

Policy size

Occupation

Known impairments

Postcode/geographical area

Education

Socio-economic class / income

Marital status

(ii) For Gasperton we have, using the census formula central death rate

$$= \frac{25}{\frac{1}{2}(2,000+2,100)} = 0.0122 .$$

For Great Hawking we have central death rate

$$= \frac{21}{\frac{1}{2}(1,770+1,674)} = 0.0122 .$$

(iii) Let the death rate for smokers in Gasperton be $\gamma_s$, and that for non-smokers be $\gamma_n$.

We therefore have

$$0.5\gamma_s + 0.5\gamma_n = 0.0122$$

$$\gamma_s = 1.4\gamma_n$$

and hence

$$0.5(1.4)\gamma_n + 0.5\gamma_n = 0.0122$$

$$\gamma_n = \frac{0.0122}{1.2} = 0.0102$$

$$\gamma_s = \frac{0.0122(1.4)}{1.2} = 0.0142$$

Let the death rate for smokers in Great Hawking be $\zeta_s$, and that for non-smokers be $\zeta_n$.

We therefore have

$$0.2\zeta_s + 0.8\zeta_n = 0.0122$$

$$\zeta_s = 1.4\zeta_n$$

and hence

$$0.2(1.4)\zeta_n + 0.8\zeta_n = 0.0122$$

$$\zeta_n = \frac{0.0122}{1.08} = 0.0113$$

$$\zeta_s = \frac{0.0122(1.4)}{1.08} = 0.0158$$

(iv) The company would do better to vary the premiums on the basis of geographical area,

as it is clear that death rates in Great Hawking for both smokers and non-smokers are higher than those in Gasperton.

If the company does not differentiate its prices on the basis of geographical area, it may lose business in Gasperton to a rival company which does differentiate; conversely in Great Hawking it may attract new business from rival companies, but will underprice the product and hence risk its life assurance fund becoming insolvent.

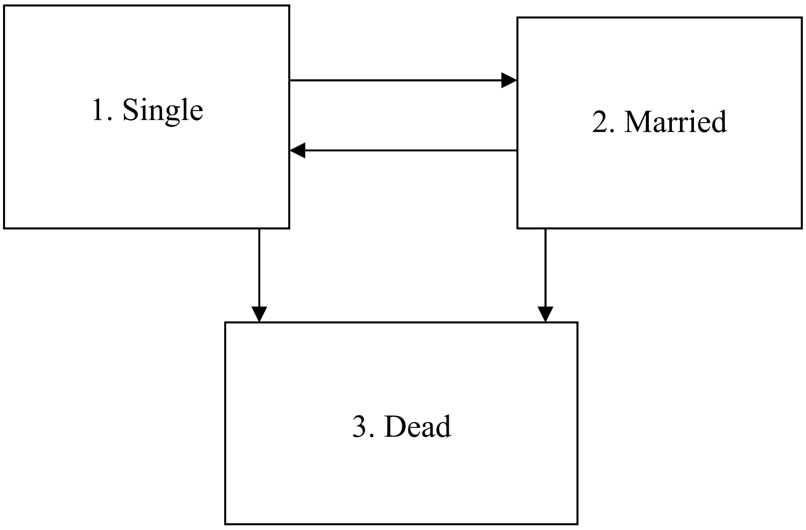
There are relatively little data, so it might be worth adopting a "wait and see" approach.

1.4 times the death rate will not translate as 1.4 times the premium. The difference may me relatively small, (although it is a 25 year term assurance so it probably is pretty significant).

*Most candidates scored highly on parts (i) and (ii). Part (iii) was very poorly answered. A large number of candidates misinterpreted the question as meaning that the ratio of the* <u>*numbers of deaths*</u> *to smokers and non-smokers was 1.4. This works for Gasperton because there are equal numbers of smokers and non-smokers in the exposed-to-risk, but for Great Hawking it produces incorrect results. Only a minority of candidates made a serious attempt at part (iv). Credit was given for any sensible comments in part (iv) which were consistent with the answers given to parts (ii) and (iii).*

## 10

(i)




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

where

$\mu^{ij}$ is the transition intensity from state *i* to state *j*

$v^i$ is the total observed waiting time in state *i*

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

(iii) Taking the logarithm of the likelihood we get

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

Differentiate with respect to $\mu^{13}$

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

Setting this to zero we obtain

$$\hat{\mu}^{13} = \frac{d^{13}}{\nu^1}.$$

To check it is a maximum differentiate again giving

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

(iv) The maximum likelihood estimate of $\mu^{13}$ is 12/10,298 = 0.001165.

The variance is $-1\Big/\dfrac{d^2\ln(L)}{(d\mu^{13})^2} = 12\big/10{,}298^2 = 1.13 \times 10^{-7}$.

*This was the best answered question on the paper, with most candidates scoring at least 9 of the 11 marks available. In part (ii) many candidates omitted the constant of proportionality. In part (iv) the question says "estimate", so we needed some indication of where the answers came from for full marks.*

## 11

(i) $X_{n+1} = X_n + Y_{n+1} = X_n + f(X_n)$

so the series $X_i$ depends only on the current state and hence satisfies the Markov property.

$$Y_{n+1} = \frac{1}{4}\left(1 + \frac{X_n}{n}\right) = \frac{1}{4}\left(1 + \frac{\sum_{j=1}^{n} Y_j}{n}\right)$$

and hence depends on all the previous values of $Y_i$.

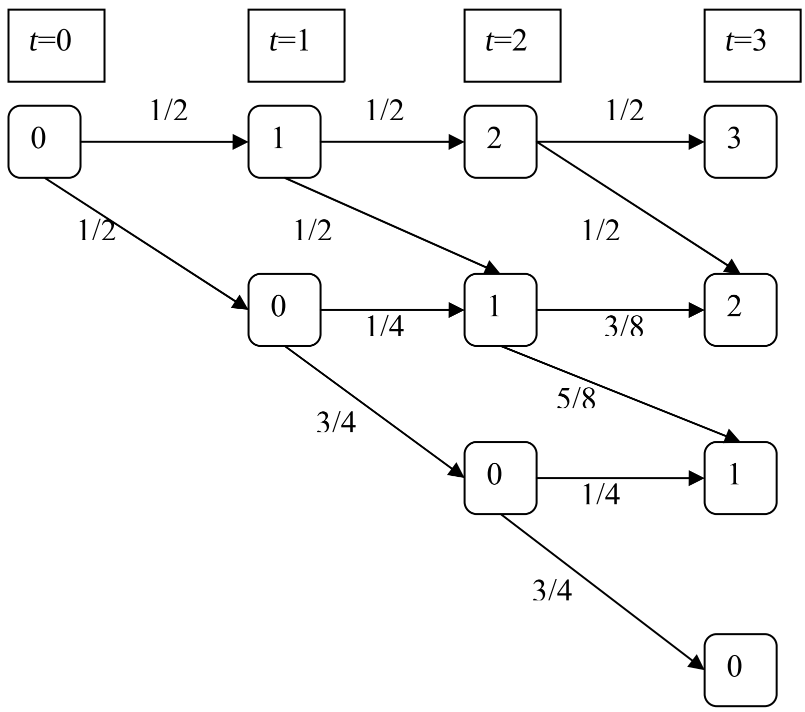
(ii) (a) It is not possible for the cumulative number of accidents to reduce (OR the cumulative number of accidents is an increasing/ non-decreasing function)

and so the process is not irreducible.

(b) The probabilities depend on the number of time periods $n$

so the process is not time homogeneous

(iii)




(iv) From the diagram above (or otherwise) it can be seen that there are three paths to the 2 accidents by time 3 box.

Required probability = Pr(0–0–1–2)+Pr(0–1–1–2)+Pr(0–1–2–2)

$$=\frac{1}{2}.\frac{1}{4}.\frac{3}{8}+\frac{1}{2}.\frac{1}{2}.\frac{3}{8}+\frac{1}{2}.\frac{1}{2}.\frac{1}{2}=\frac{3+6+8}{64}=\frac{17}{64}$$

(v) It is reasonable to assume that probability of having an accident depends on the number of previous accidents.

It is also reasonable that the effect of a previous accident should wear off over time.

There are likely to be other factors which have a significant effect on the probability of an accident, such as the fact that people who have recently had an accident might drive more carefully.

May want to give more weighting to recent years.

*This was a demanding question, and only a minority of candidates scored highly. In part (i) few answers were sufficiently rigorous, many either re-stating the question or simply stating the Markov property. Part (ii)(a) was well answered, but in part (ii)(b) many candidates did not understand the term "time homogeneous". Most candidates made only sketchy attempts at part (iii) and (iv). Credit was given for calculations in part (iv) which demonstrated that candidates knew the correct method, even though the numbers were incorrect.*

## 12

(i) A process with a discrete state space and discrete time space

where the future development is only dependent on the current state occupied.

OR

$P[X_t \in A \mid X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n] = P[X_t \in A \mid X_s = x]$
for all times $s_1 < s_2 < \ldots < s_n < s < t$, all states $x_1, x_2, \ldots, x_n, x$ in $S$ and all subsets $A$ of $S$.

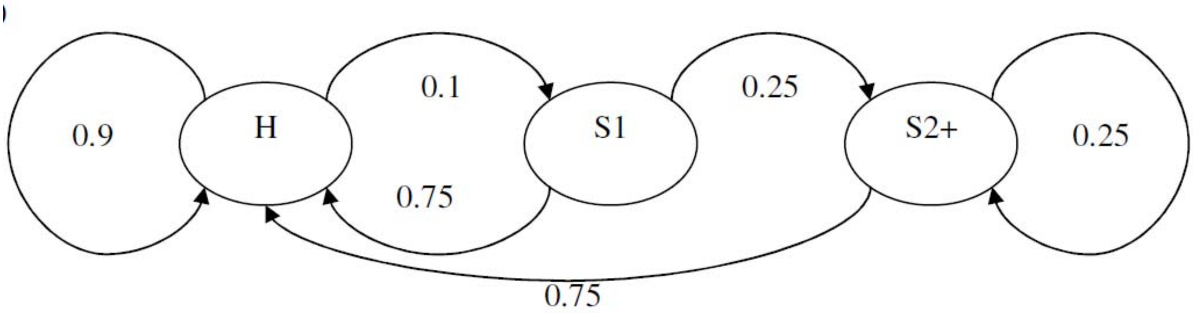
THEN EITHER THE THREE STATE SOLUTION

(ii) The sick pay depends on the duration of sickness, so to model with a time homogeneous Markov chain needs as a minimum the states:

Healthy (H)
Sick month 1 (S1)
Sick month 2 or more (S2+)

So the minimum number of states is 3.

(iii)




(iv) If using H, S1, S2+ then the stationary distribution π is given by:

$$\pi\begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.75 & 0 & 0.25 \\ 0.75 & 0 & 0.25 \end{pmatrix} = \pi$$

$$\pi_H = 0.9\pi_H + 0.75\pi_{S_1} + 0.75\pi_{S_{2+}}$$
$$\pi_{S_1} = 0.1\pi_H$$
$$\pi_{S_{2+}} = 0.25\pi_{S_1} + 0.25\pi_{S_{2+}}$$
$$\pi_H + \pi_{S_1} + \pi_{S_{2+}} = 1$$

$$3\pi_{S_{2+}} = \pi_{S_1}$$
$$30\pi_{S_{2+}} = \pi_H$$

Implies

$$\pi_{S_{2+}} = \frac{1}{34},\ \pi_{S_1} = \frac{3}{34},\ \pi_H = \frac{15}{17}$$

(v) Let percentage of salary when healthy be *a*%

Then in the stationary state looking at payments for the <u>next</u> month we need

Expected income = Expected outgo.

Probability healthy **a*% of salary = Probability of 100% sick pay*100% of salary + Probability on 50% sick pay*50% of salary:

$$\left(\frac{15}{17}*0.9+\frac{2}{17}*0.75\right)a = \left(\frac{3}{34}*0.25+\frac{1}{34}*0.25*50\%+\frac{15}{17}*0.1\right)$$

$0.88235a = 0.113971$
$a$ =12.917% of salary.

(vi) Now just need a two state version {H,S}

$$\pi_H = \frac{15}{17},\ \pi_S = \frac{2}{17}$$

and need contribution rate > 2/15 = 13.333% of salary.

OR THE FOUR STATE SOLUTION

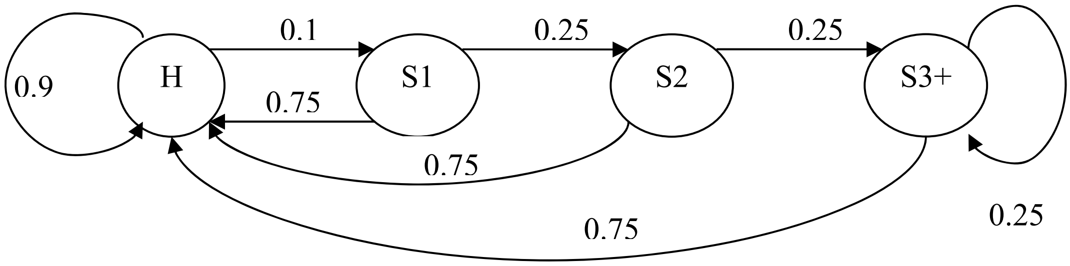
(ii) The sick pay depends on the duration of sickness, so to model with a time homogeneous Markov chain needs as a minimum the states:

Healthy (H)
Sick month 1 (S1)
Sick month 2 (S2)
Sick month 3 or more (S3+)

So the minimum number of states is 4.

[2]

(iii)




(iv) If using H, S1, S2, S3+ then the stationary distribution π is given by:

$$\pi\begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.75 & 0 & 0.25 & 0 \\ 0.75 & 0 & 0 & 0.25 \\ 0.75 & 0 & 0 & 0.25 \end{pmatrix} = \pi$$

$$\pi_H = 0.9\pi_H + 0.75\pi_{S_1} + 0.75\pi_{S_2} + 0.75\pi_{S_{3+}}$$

$$\pi_{S_1} = 0.1\pi_H$$
$$\pi_{S_2} = 0.25\pi_{S_1}$$
$$\pi_{S_{3+}} = 0.25\pi_{S_2} + 0.25\pi_{S_{3+}}$$
$$\pi_H + \pi_{S_1} + \pi_{S_2} + \pi_{S_{3+}} = 1$$

$$\pi_H = 10\pi_{S_1}$$

$$\pi_{S_2} = \frac{1}{4}\pi_{S_1}$$

$$0.75\pi_{S_{3+}} = 0.25\pi_{S_2}$$

$$\pi_{S_{3+}} = \frac{1}{3}.\frac{1}{4}\pi_{S_1}$$

$$\pi_{S_1}\left\{10+1+\frac{1}{4}+\frac{1}{12}\right\}=1$$

Implies

$$\pi_{S_1} = \frac{12}{136} = \frac{3}{34},\ \pi_H = \frac{120}{136} = \frac{15}{17},\ \pi_{S_2} = \frac{3}{136}\ \pi_{S_{3+}} = \frac{1}{136}$$

(v) Let percentage of salary when healthy be $a$%

Expected income = Expected outgo

Probability healthy *$a$% of salary = Probability of 100% sick pay*100% of salary
Probability on 50% sick pay*50% of salary

$$\frac{15}{17}a = \left(\frac{3}{34}+\frac{3}{136}\right)+\frac{1}{2}\left(\frac{1}{136}\right)$$

$0.88235a = 0.113971$

$a$ =12.917% of salary.

(vi) Now all those not healthy get 100% so

$$\frac{15}{17}a = \frac{2}{17}$$

and need contribution rate > 2/15 = 13.333% of salary

(vii) The reduction in cost is calculated as 3.23%.

This is not particularly significant either relative to the likely uncertainty in the assumptions or because recovery rates are so high.

The reduction in sick pay is likely to encourage employees to try to get back into work.

*This question was well answered despite its complexity. Most candidates went for the four state solution, and there were many correct answers to parts (i)-(iv). In parts (v) and (vi) a* *common mistake was to fail to divide by the proportion of healthy employees, as only healthy employees (i.e. those not receiving sick pay) contribute to the scheme. Answers to part (vii) often included sensible comments that gained credit, even if some candidates answered as if the scheme had an unlimited supply of funds!*

**END OF EXAMINERS' REPORT**

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINATION

27 September 2012 (am)

## Subject CT4 – Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 10 questions, beginning your answer to each question on a separate sheet.*

5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** Describe two benefits and two limitations of using models in actuarial work. [4]

**2** A large company wishes to construct a model of sickness rates among its employees to use in evaluating the present and future financial health of its sick pay scheme. Outline factors which the company should take into consideration when developing the model. [5]

**3** (i) State the principle of correspondence as it applies to mortality rates. [1]

A life insurance company has the following data:

| | *Number of policies in force on* | | | |
|---|---|---|---|---|
| *Age last birthday* | *1 January 2009* | *1 January 2010* | *1 July 2010* | *1 January 2011* |
| 49 | 2,000 | 2,100 | 2,300 | 2,500 |
| 50 | 2,100 | 2,200 | 2,300 | 2,400 |
| 51 | 2,300 | 2,400 | 2,500 | 2,600 |

*Number of deaths classified by age next birthday and calendar year*

| *Age next birthday* | *2009* | *2010* |
|---|---|---|
| 49 | 175 | 200 |
| 50 | 200 | 225 |
| 51 | 225 | 235 |

(ii) Estimate, using these data, the force of mortality at age 50 next birthday for the period 1 January 2009 to 1 January 2011. [5]

(iii) State the exact age to which your answer to part (ii) relates. [1]
[Total 7]

**4** (i) State one advantage of a semi-parametric model over a fully parametric one. [1]

(ii) Write down a general expression for the Cox proportional hazards model, defining all the terms you use. [2]

A life office is trying to understand the impact of certain factors on the lapse rates of its policies. It has studied the lapse rates on a block of business subdivided by:

- sex of policyholder (Male or Female)
- policy type (Term Assurance or Whole Life)
- sales channel (Internet, Direct Sales Force or Independent Financial Adviser)

The office has fitted a Cox proportional hazards model to the data and has calculated the following regression parameters:

| *Covariate* | *Regression parameter* |
|---|---|
| Female | 0.2 |
| Male | 0 |
| Term Assurance | −0.1 |
| Whole Life | 0 |
| Internet | 0.4 |
| Independent Financial Adviser | −0.2 |
| Direct Sales Force | 0 |

(iii) State the sex/sales channel/policy type combination to which the baseline hazard relates. [1]

A Term Assurance is sold to a Female by an Independent Financial Adviser.

(iv) Calculate the probability that this Term Assurance is still in force after five years given that 60% of Whole Life policies bought on the Internet by Males have lapsed by the end of year five. [4]
[Total 8]

**5** A no claims discount system operates with three levels of discount, 0%, 15% and 40%. If a policyholder makes no claim during the year he moves up a level of discount (or remains at the maximum level). If he makes one claim during the year he moves down one level of discount (or remains at the minimum level) and if he makes two or more claims he moves down to, or remains at, the minimum level.

The probability for each policyholder of making two or more claims in a year is 25% of the probability of making only one claim.

The long-term probability of being at the 15% level is the same as the long-term probability of being at the 40% level.

(i) Derive the probability of a policyholder making only one claim in a given year. [4]

(ii) Determine the probability that a policyholder at the 0% level this year will be at the 40% level after three years. [2]

(iii) Estimate the probability that a policyholder at the 0% level this year will be at the 40% level after 20 years, without calculating the associated transition matrix. [3]
[Total 9]

**6** (i) Define the stationary distribution of a Markov chain. [2]

A baseball stadium hosts a match each evening. As matches take place in the evening, floodlights are needed. The floodlights have a tendency to break down. If the floodlights break down, the game has to be abandoned and this costs the stadium \$10,000. If the floodlights work throughout one match there is a 5% chance that they will fail and lead to the abandonment of the next match.

The stadium has an arrangement with the Floodwatch repair company who are brought in the morning after a floodlight breakdown and charge \$1,000 per day. There is a 60% chance they are able to repair the floodlights such that the evening game can take place and be completed without needing to be abandoned. If they are still broken the repair company is used (and paid) again each day until the lights are fixed, with the same 60% chance of fixing the lights each day.

(ii) Write down the transition matrix for the process which describes whether the floodlights are working or not. [1]

(iii) Derive the long run proportion of games which have to be abandoned. [3]

The stadium manager is unhappy with the number of games being abandoned, and contacts the Light Fantastic repair company who are estimated to have an 80% chance of repairing floodlights each day. However Light Fantastic will charge more than Floodwatch.

(iv) Calculate the maximum amount the stadium should be prepared to pay Light Fantastic to improve profitability. [4]

[Total 10]

**7** The volatility of equity prices is classified as being High ($H$) or Low ($L$) according to whether it is above or below a particular level. The volatility status is assumed to follow a Markov jump process with constant transition rates $\varphi_{LH} = \mu$ and $\varphi_{HL} = \rho$.

(i) Write down the generator matrix of the Markov jump process. [1]

(ii) State the distribution of holding times in each state. [1]

A history of equity price volatility is available over a representative time period.

(iii) Explain how the parameters $\mu$ and $\rho$ can be estimated. [2]

Let ${}_t p_s^{ij}$ be the probability that the process is in state $j$ at time $s$+$t$ given that it was in state $i$ at time $s$ ($i, j = H, L$), where $t \geq 0$. Let ${}_t p_s^{\overline{ii}}$ be the probability that the process remains in state $i$ from time $s$ to time $s$+$t$.

(iv) Write down Kolmogorov's forward equations for $\frac{\partial}{\partial t} {}_tP_s^{\overline{LL}}$, $\frac{\partial}{\partial t} {}_tP_s^{LL}$ and $\frac{\partial}{\partial t} {}_tP_s^{LH}$ . [2]

Equity price volatility is Low at time zero.

(v) Derive an expression for the time after which there is a greater than 50% chance of having experienced a period of high equity price volatility. [2]

(vi) Solve the Kolmogorov equation to obtain an expression for ${}_tP_0^{LL}$. [4]

[Total 12]

**8** (i) Describe a situation when graduation of raw mortality data using a parametric formula might be appropriate and explain why. [2]

(ii) (a) State another method of graduation.
(b) Suggest a situation in which its use may be appropriate. [1]

A large insurance company has graduated the mortality experience of part of its business. The original data and the graduated rates are as follows.

| *Age* | *Exposed to risk* | *Number of deaths* | *Graduated rates* ($\hat{q}_s$) |
|---|---|---|---|
| 40 | 1284 | 4 | 0.00240 |
| 41 | 2038 | 4 | 0.00266 |
| 42 | 1952 | 12 | 0.00297 |
| 43 | 2158 | 7 | 0.00332 |
| 44 | 2480 | 11 | 0.00371 |
| 45 | 1456 | 7 | 0.00415 |
| 46 | 2100 | 12 | 0.00464 |
| 47 | 1866 | 16 | 0.00519 |
| 48 | 1989 | 15 | 0.00577 |
| 49 | 1725 | 10 | 0.00642 |

(iii) Test this graduation for overall goodness of fit. [5]

(iv) Discuss whether it may be necessary to test for smoothness. [2]

(v) Test the data for individual outliers. [3]

[Total 13]

**9** A certain town runs a training course for traffic wardens each year. The course lasts for 30 days, but the examination which enables someone to qualify as a traffic warden can be sat any day during the course. In 2011 there were 13 participants who started the training course. The following table has been compiled to show the day each candidate qualified or the day each candidate who did not qualify left the course.

| *Candidate* | *Day qualified* | *Day left without qualifying* |
|---|---|---|
| A | | 30 |
| B | 5 | |
| C | | 21 |
| D | 19 | |
| E | 12 | |
| F | | 30 |
| G | 1 | |
| H | | 19 |
| I | 12 | |
| J | | 30 |
| K | 15 | |
| L | | 10 |
| M | 24 | |

(i) Explain whether the following types of censoring are present:

- interval censoring
- right censoring
- informative censoring

[3]

(ii) Calculate the Kaplan-Meier estimate of the non-qualification function. [6]

(iii) Sketch a graph of the Kaplan-Meier estimate, labelling the axes. [2]

When the data were gathered, the reasons for exit of candidates *D* and *H* were accidentally transposed, and those for candidates *B* and *L* were also accidentally transposed.

(iv) Explain how your answer to part (ii) would change if you had access to the correct (i.e. untransposed) data for candidates *D*, *H*, *B* and *L*. [3]

[Total 14]

**10** On a small distant planet lives a race of aliens. The aliens can die in one of two ways, either through illness, or by being sacrificed according to the ancient custom of the planet. Aliens who die from either cause may, some time later, become zombies.

(i) Draw a multiple-state diagram with four states illustrating the process by which aliens die and become zombies, labelling the four states and the possible transitions between them. [2]

(ii) Write down the likelihood of the process in terms of the transition intensities, the numbers of events observed and the waiting times in the relevant states, clearly defining all the terms you use. [4]

(iii) Derive the maximum likelihood estimator of the death rate from illness. [3]

The aliens take censuses of their population every ten years (where the year is an "alien year", which is the length of time their planet takes to orbit their sun). On 1 January in alien year 46,567, there were 3,189 live aliens in the population. On 1 January in alien year 46,577 there were 2,811 live aliens in the population. During the intervening ten alien years, a total of 3,690 aliens died from illness and 2,310 were sacrificed, and the annual death rates from illness and sacrifice were constant and the same for each alien.

(iv) Estimate the annual death rates from illness and from sacrifice over the ten alien years between alien years 46,567 and 46,577. [2]

The rate at which aliens who have died from either cause become zombies is 0.1 per alien year.

(v) Calculate the probabilities that an alien alive in alien year 46,567 will, ten alien years later:

(a) still be alive
(b) be dead but not a zombie

[7]
[Total 18]

## END OF PAPER

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS' REPORT

September 2012 examinations

## 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.

D C Bowie
Chairman of the Board of Examiners

December 2012

## General comments on Subject CT4

Subject CT4 comprises five main sections: (1) 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); (2) stochastic processes, especially Markov chains and Markov jump processes; (3) models of a random variable measuring future lifetime; (4) the calculation of exposed to risk and the application of the principle of correspondence; (5) the reasons why mortality (or other decrement) 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.

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.

## Comments on the September 2012 paper

The general performance was slightly better than in the September 2010 or September 2011 sessions, but substantially inferior to that in April 2012. Nevertheless, well-prepared candidates scored highly across the whole paper. The comments that follow the questions concentrate on areas where candidates could have improved their performance.

**1**

**Benefits**

Systems with long time frames can be studied in compressed time

Complex systems with stochastic elements can be studied (especially by simulation modelling).

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

Models allow control over experimental conditions, so that we can reduce the variance of the results output without upsetting the mean values.

**Limitations**

Model development requires a lot of time and expertise, and hence can be costly.

Models more useful for comparing the results of input variations than for optimising outputs.

Models can look impressive, but can lull the user into a false sense of security. Impressive output is not a substitute for validity and close imitation of the real world.

Models rely heavily on the data input. If this is poor or lacking in credibility the output is likely to be flawed.

Models rely heavily on the assumptions used, poor assumptions can invalidate the model output.

Users need to understand the model sufficiently well to be able to know when it is appropriate to apply it.

Interpretation of models can be difficult.

Models cannot take into account all possible future events, e.g. changes in legislation.

*Many candidates scored full marks on this question. The question asked for TWO benefits and TWO limitations, so credit was given for the most fully described two of each. Extra marks for the benefits could not be transferred to the limitations to make up a shortfall, and vice versa.*

**2**

The nature of the existing sickness data the company possesses. The model can only be as complex as the data will allow it to be.

Whether the company has made any previous attempts to model sickness rates among its employees, and how successful they were.

The complexity of the model – e.g. whether it should be stochastic or deterministic. More complex models will be costlier to prepare and run, but eventually there may be diminishing returns to additional complexity.

General trends in sickness at the national level may need to be built in.

The definition of sickness and level of benefits payable under the scheme.

Does the company plan to change the characteristics of the employees? For example, does it plan to recruit more mature persons?

The ease of communication of the model.

The budget and resources available for the construction of the model.

Capability of staff. Will outside consultants be required?

By whom will the model be used? Will they be capable of understanding and using it?

Does the model need to interface with models of other aspects of the company's business (e.g. taking data from other systems)?

The independence of sickness rates should be taken into account e.g. in the event of an epidemic claims cannot be considered independent.

*Other relevant points were given credit. The Examiners were looking for comments which made reference to the scenario proposed in the question. Many candidates simply reproduced one of the lists in the Core Reading (unit 1 page 4 and unit 1 page 6) without relating to the scenario in the question. Answers along these lines scored limited credit.*

**3**

(i) The principle of correspondence states that a life should be included in the denominator of the rate at time $t$ if and only if, were that life to die at time $t$, his or her death would be counted in the numerator.

(ii) In order for the exposed to risk to correspond to the deaths data, it needs to be on an age next birthday basis.

The exposed to risk at age $x$ next birthday may be approximated using the census approximation.

$$E_x^c = \int_0^2 P_{x,t}\,dt$$

Using the trapezium rule (i.e. assuming the population varies linearly between "census" dates) this may be evaluated as

$$E_x^c = \frac{1}{2}(P_{x,1/1/09} + P_{x,1/1/10}) + \frac{1}{4}(P_{x,1/1/10} + P_{x,1/7/10}) + \frac{1}{4}(P_{x,1/7/10} + P_{x,1/1/11})$$
$$= \frac{1}{2}P_{x,1/1/09} + \frac{3}{4}P_{x,1/1/10} + \frac{1}{2}P_{x,1/7/10} + \frac{1}{4}P_{x,1/1/11}$$

where $P_{x,t}$ is the population aged $x$ next birthday at time $t$.

But, in this case, we have data on an age last birthday basis.

If $P^*_{x,t}$ is the population aged $x$ last birthday at time $t$, then

$$P_{x,t} = P^*_{x-1,t}$$

and the exposed to risk becomes

$$E_x^c = \frac{1}{2}P^*_{x-1,1/1/09} + \frac{3}{4}P^*_{x-1,1/1/10} + \frac{1}{2}P^*_{x-1,1/7/10} + \frac{1}{4}P^*_{x-1,1/1/11}$$

So, using the data given, the exposed to risk we need at age 50 is

$$E_x^c = \frac{1}{2}(2{,}000) + \frac{3}{4}(2{,}100) + \frac{1}{2}(2{,}300) + \frac{1}{4}(2{,}500) = 4{,}350$$

and the estimated force of mortality at age 50 next birthday is

$$\hat{\mu}_{50} = \frac{200 + 225}{4{,}350} = 0.0977$$

(iii) The estimate $\hat{\mu}_{50}$ applies to the middle of the rate interval,

which is exact age 49.5 years.

*In part (ii) the question said "estimate" so some indication of how the answer was arrived at was required for full credit. The correct numerical answer on its own was insufficient. Some candidates noted that a correct exposed-to-risk could be calculated without using the July 2010 population figures. This was given full credit, provided a valid explanation of why the July population figures were not needed was given. In part (iii) the question said "state" so the full mark was awarded for 49.5. In part (iii) for full credit the answer had to be consistent with what the candidate had done in part (ii).*

## 4

(i) We do not need to know the general shape of the hazard/distribution.

(ii) $h(t, z_i) = h_0(t)\exp\left(\beta z_i^T\right)$

$h(t, z_i)$ is the hazard at time $t$ (or just $h(t)$ is OK)
$h_0(t)$ is the baseline hazard
$z_i$ are covariates
$\beta$ is a vector of regression parameters

(iii) Baseline hazard refers to a male sold a whole life policy by the direct sales force.

(iv) For the male policy

the probability still in force is 0.4.

Sum of parameters for male is 0.4

$$0.4 = \exp\left\{-\int_0^5 h_0(t)\exp(0.4)dt\right\} = \exp\left\{-1.49\int_0^5 h_0(t)dt\right\}$$

So $$\int_0^5 h_0(t)dt = \frac{\ln 0.4}{-1.49}$$

And for the female policy sum of parameters is -0.1

THEN EITHER

We therefore want $\exp\left\{-\int_0^5 h_0(t)\exp(-0.1)dt\right\} = \exp\left\{-0.905 * \frac{\ln 0.4}{-1.49}\right\}$

$= 0.57364$

OR

We therefore want $\left\{0.4^{e^{-0.4}}\right\}^{e^{-0.1}} = 0.4^{e^{-0.5}} \qquad = 0.57364$

*Parts (i)-(iii) of this question were well answered. Answers to part (iv) were variable. Common errors included working with the probability of having lapsed (i.e. 1 minus the probability of still being in force), and omission of the integral.*

**5**

(i) Let $x = \frac{5}{4}c$

where $c$ is the probability of exactly one claim in a year and $x$ is the probability of one or more claims in a year.

The transition matrix is

$$\begin{pmatrix} x & 1-x & 0 \\ x & 0 & 1-x \\ \frac{c}{4} & c & 1-x \end{pmatrix}$$

Using $\pi = \pi P$ we get

$$\pi_1 = x\pi_1 + x\pi_2 + {}^{c}\!/_{4}\,\pi_3$$
$$\pi_2 = (1-x)\pi_1 + c\pi_3$$
$$\pi_3 = (1-x)\pi_2 + (1-x)\pi_3$$

The equation for $\pi_3$ gives

$$\pi_2(1-x) = \pi_3\{1-(1-x)\} = \pi_3 x$$

$$\pi_2 = \pi_3 \frac{x}{1-x}$$

So $x = 1 - x$ from which $x = 0.5$ and $c = 0.4$

So the probability of exactly one claim in any given year is 0.4.

(ii) EITHER

Using the transition matrix

$$M = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.1 & 0.4 & 0.5 \end{pmatrix}$$

$$M^2 = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.1 & 0.4 & 0.5 \end{pmatrix}\begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0.1 & 0.4 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.3 & 0.45 & 0.25 \\ 0.3 & 0.25 & 0.45 \end{pmatrix}$$

The required probability is therefore

$(0.5 \times 0.25) + (0.5 \times 0.25) + (0 \times 0.45) = 0.25$

OR

We require the probability of no claims in either of years 2 and 3 (since only this will leave the policyholder at the 40% level at the end of year 3).

The probability of one or more claims is 0.5 (from the solution to part (i)).

So the probability of no claims is 0.5, and the probability of no claims in years 2 and 3 is $0.5 \times 0.5 = 0.25$.

(iii) After 20 years the probabilities of being at any level will be close to the stationary probability distribution

From part (i) we know that $\pi_2 = \pi_3$.

Using $\pi = \pi P$ we get

$0.5\pi_1 + 0.5\pi_2 + 0.1\pi_3 = \pi_1$,

so $\pi_2 = \frac{5}{6}\pi_1$.

Since $\pi_1 + \pi_2 + \pi_3 = 1$,+½

we have $\pi_1 = \frac{3}{8}$, $\pi_2 = \pi_3 = \frac{5}{16}$.

So the probability of being at the 40% level after 20 years is estimated as 0.3125.

*This question proved more difficult for candidates that the Examiners had envisaged, and answers were disappointing. Various alternative specifications of the matrix in part (i) were acceptable. In all three parts of this question some indication of how each result was obtained was required. Candidates who just wrote down the numerical answers did not score full credit. The solution to part (ii) could be found by drawing a diagram and tracing the possible routes through: this is perfectly valid and is arguably the quickest way to the correct answer. In part (iii) some indication that the answer is an estimate was required. This could be provided by saying, for example, that after 20 years the probabilities of being at any level will be close to the stationary probability distribution.*

**6**

(i) Let $S$ be the state space. We say that $\{\pi_j \mid j \in S\}$ is a stationary probability distribution for a Markov chain with transition matrix $P$ if the following hold for all $j \in S$:

$$\pi_j = \sum_{i \in S} \pi_i p_{ij}, \text{ OR } \pi = \pi \text{ P}$$

$$\sum_{j \in S} \pi_j = 1.$$

$\pi_i \geq 0$

(ii) With state space {Working, Broken}

$$\text{Transition matrix } A = \begin{pmatrix} 0.95 & 0.05 \\ 0.6 & 0.4 \end{pmatrix}$$

(iii) This requires the stationary distribution $\pi$ which satisfies

$\pi A = \pi$

$$0.95\pi_W + 0.6\pi_B = \pi_W$$
$$0.05\pi_W + 0.4\pi_B = \pi_B$$

and $\pi_W + \pi_B = 1$

$$\pi_W = 12\pi_B$$
$$\pi_W = 12/13$$
$$\pi_B = 1/13$$

So 1 in 13 games is cancelled.

OR

The average number of games for which the lights work before breaking down is 1/0.05 = 20 games.

Once they have broken down the expected number of games for which they will be out of action is 1/0.6 = 5/3 games.

Therefore the proportion of games for which the lights are out of action is

$$\frac{5/3}{20+(5/3)}=\frac{5}{65}=\frac{1}{13}$$

So 1 in 13 games is cancelled.

(iv) First we need to find the new stationary probabilities.

Transition matrix $A'=\begin{pmatrix}0.95 & 0.05\\ 0.8 & 0.2\end{pmatrix}$

$$0.95\pi_{W'}+0.8\pi_{B'}=\pi_{W'}$$

and $\pi_{W'}+\pi_{B'}=1$

Giving $\pi_{W'}=16/17$ $\pi_{B'}=1/17$

Lost income (including fees to repair company):

with Floodwatch: (\$10,000 + \$1,000)*1/13
with Light Fantastic: (\$10,000 + $X$)*1/17 where $X$ is fee to be negotiated.

So need: 11000/13 = (10000 + $X$)/17

$X$ = \$4,384.62 per day.

*In part (i) no credit was given for wordy description of "what happens in the long run". In part (iii) the question said "derive", so we needed an explanation of where the answer came from: only limited credit was given for writing down a numerical answer (even if correct) without explanation. Moreover, in part (iii) calculating the stationary distribution was not sufficient for full credit: we were looking for the correct element to be identified and its value indicated i.e. an explicit statement that "1 in 13 games is cancelled". Parts (i), (ii) and (iii) of this question were well answered, and many candidates also evaluated the proportion of games that would be cancelled under the new floodlight regime in part (iv). Few were able to compute the daily saving, however.*

**7**

(i) With state space {L, H} we have generator matrix

$$A = \begin{pmatrix} -\mu & \mu \\ \rho & -\rho \end{pmatrix}$$

(ii) The holding times are exponentially distributed with parameter $\mu$ in state $L$ and $\rho$ in state $H$.

(iii) EITHER

The time spent in state $L$ before the next visit to $H$ has mean $1/\mu$.

Therefore a reasonable estimate for $\mu$ is the reciprocal of the mean length of each visit:

= (Number of transitions from $L$ to $H$) / (Total time spent in state $L$)

Similarly estimate for $\rho$ is the reciprocal of the mean length of each visit:

= (Number of transitions from $H$ to $L$) / (Total time spent in state $H$)

OR

Using the maximum likelihood estimator for $\mu$, we have:

(Number of transitions from L to H)/Total time spent in state L).

Similarly, the MLE of $\rho$ is

(Number of transitions from H to L)/Total time spent in state H).

(iv) $$\frac{\partial}{\partial t}\, {}_tP_s^{\overline{LL}} = -\mu\, {}_tP_s^{\overline{LL}}$$

$$\frac{\partial}{\partial t}\, {}_tP_s^{LL} = -\mu\, {}_tP_s^{LL} + \rho\, {}_tP_s^{LH}$$

$$\frac{\partial}{\partial t}\, {}_tP_s^{LH} = \mu\, {}_tP_s^{LL} - \rho\, {}_tP_s^{LH}$$

(v) $$\frac{\partial}{\partial t}\, {}_tP_s^{\overline{LL}} = -\mu\, {}_tP_s^{\overline{LL}}$$

so ${}_tP_0^{\overline{LL}} = \exp(-\mu t)$

Looking for time when ${}_tP_0^{\overline{LL}} = 1/2$

$1/2 = \exp(-\mu T)$
$T = \ln(2)/\mu$

(vi) Observe that ${}_tP_0^{LL} + {}_tP_0^{LH} = 1$

so, substituting, we have

$$\frac{\partial}{\partial t}\,{}_tP_0^{LL} = -\mu\,{}_tP_0^{LL} + \rho(1 - {}_tP_0^{LL})$$

$$\frac{\partial}{\partial t}\left[\exp((\mu+\rho)t)\,{}_tP_0^{LL}\right] = \rho\exp((\mu+\rho)t)$$

$$\exp((\mu+\rho)t)\,{}_tP_0^{LL} = \frac{\rho}{\mu+\rho}\exp((\mu+\rho)t) + \text{constant}$$

And in state $L$ at time zero so $const = \dfrac{\mu}{\mu+\rho}$

$${}_tP_0^{LL} = \frac{\rho}{\mu+\rho} + \frac{\mu}{\mu+\rho}\exp(-(\mu+\rho)t)$$

*Few candidates scored highly on this question. In particular, very few made a serious attempt at parts (v) and (vi). In part (iv), there was confusion among some candidates between ${}_tp_0^{LL}$ and ${}_tp_0^{\overline{LL}}$ and a common error was to write down $\frac{\partial}{\partial t}\,{}_tP_s^{\overline{LL}} = \exp(-\mu t)$.*
*Many candidates did not attempt part (v) even though is is relatively straightforward.*
*In part (vi) working through with ${}_tP_0^{LH}$ then at the end taking one minus the answer is a valid approach .*

## 8

(i) When preparing standard tables OR when graduating data from a large industrywide scheme, or a national population

because there will be lots of data available.

(ii) (a) EITHER Graphical graduation OR Graduation with reference to a standard table

(b) EITHER
Graphical graduation may be suitable for a analysis of a newly discovered insect (as data will be scanty and an existing table will not exist)

OR

Graduation with reference to a standard table is useful if data are scanty and a suitable standard table exists (e.g. for female pensioners from a small scheme).

(iii) To test for overall goodness of fit we use the $\chi^2$ test.

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

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

| *Age* | *Exposed to risk* | *Observed deaths* | *Graduated rates* $(\hat{q}_s)$ | *Expected deaths* | $z_x$ | $z_x^2$ |
|---|---|---|---|---|---|---|
| 40 | 1,284 | 4 | .00240 | 3.0816 | 0.5232 | 0.2737 |
| 41 | 2,038 | 4 | .00266 | 5.4211 | -0.6103 | 0.3725 |
| 42 | 1,952 | 12 | .00297 | 5.7974 | 2.5760 | 6.6360 |
| 43 | 2,158 | 7 | .00332 | 7.1646 | -0.0615 | 0.0038 |
| 44 | 2,480 | 11 | .00371 | 9.2008 | 0.5932 | 0.3518 |
| 45 | 1,456 | 7 | .00415 | 6.0424 | 0.3896 | 0.1518 |
| 46 | 2,100 | 12 | .00464 | 9.7440 | 0.7227 | 0.5223 |
| 47 | 1,866 | 16 | .00519 | 9.6845 | 2.0294 | 4.1184 |
| 48 | 1,989 | 15 | .00577 | 11.4765 | 1.0401 | 1.0818 |
| 49 | 1,725 | 10 | .00642 | 11.0745 | -0.3229 | 0.1043 |
| | | | | Total | 6.8794 | 13.6163 |

The observed test statistic is 13.62

The number of age groups is 10, but we lose an unknown number of degrees for the graduation, perhaps 2. So *m* = 8, say.

The critical value of the chi-squared distribution with 8 degrees of freedom at the 5% level is 15.51.

Since 13.62 < 15.51

we do not reject the null hypothesis.

(iv) It is not necessary to test for smoothness if the graduation was performed using a parametric formula or a standard table, provided that a small number of parameters were used in the formula, or in the function linking to the rates in the standard table.

It will be necessary to test for smoothness if the graduation was performed graphically but this is unlikely to be the case with data from a large insurance company.

(v) The null hypothesis is that the graduated rates are the same as the true underlying rates in the block of business. (i.e the same as part (iii))

We would expect the individual deviations to be distributed Normal (0,1)

and therefore only 1 in 20 $z_x$s should have absolute magnitude greater than 1.96 (or none should be outside −3 to +3)

Looking at the $z_x$s we see that the largest one is 2.576 and the next is 2.0294

Since they are both greater in magnitude than 1.96

we have sufficient evidence to reject the null hypothesis.

*In part (ii)(b) credit was given either for a valid reason or an appropriate example: both are not required. In part (iii) some candidates combined ages 40 and 41, on the basis that the expected deaths at age 40 are fewer than 5, and the statement in the Core Reading at the bottom of Unit 11, p. 10. This was fine. The relevant numbers for the combined 40-41 year age group will be*

| | |
|---|---|
| *Observed deaths* | *8* |
| *Expected deaths* | *8.5027* |
| $z_x$ | *0.1724* |
| $z_x^2$ | *0.0297* |
| *Chi-squared* | *12.9998* |

*Because we now only have 9 age groups, we should test against the chi-squared distribution with fewer than 9 degrees of freedom. In part (iv) no credit was given for performing a test for smoothness. Very limited credit was given for impressionistic comments on the putative smoothness or otherwise of the data given. In part (v) few candidates specified the null hypothesis or the distribution of the individual deviations under the null hypothesis.*

## 9

(i) Interval
No. We are counting in days and we know which day each event occurred.

Right
Yes. The end of the course at day 30 cut short the investigation when not all candidates had qualified.

Informative
Possible. Those who left during the 30 days will probably take longer to qualify than those who stayed.

(ii) The data can be re-arranged as shown below.

| *Day* | *Candidate* | *Event* |
|---|---|---|
| 1 | G | Qualified |
| 5 | B | Qualified |
| 10 | L | Left |
| 12 | E | Qualified |
| 12 | I | Qualified |
| 15 | K | Qualified |
| 19 | D | Qualified |
| 19 | H | Left |
| 21 | C | Left |
| 24 | M | Qualified |
| 30 | A | Left |
| 30 | F | Left |
| 30 | J | Left |

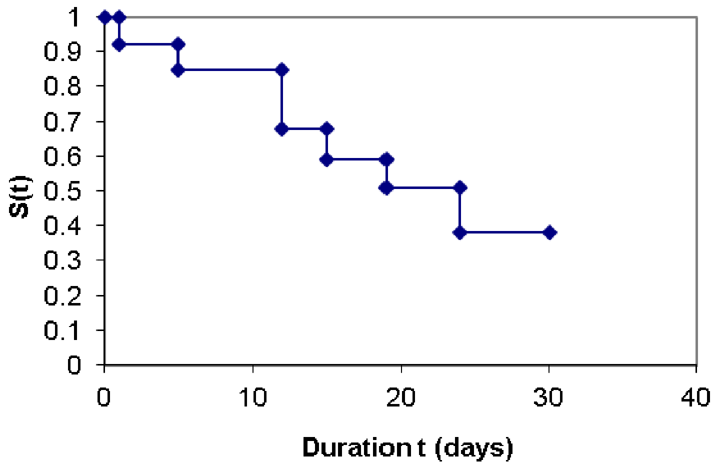
The Kaplan-Meier Estimate is $\hat{S}(t) = \prod_{t_j \le t} 1 - \frac{d_j}{n_j}$

| $t_j$ | $N_j$ | $D_j$ | $C_j$ | $\frac{D_j}{N_j}$ | $1-\frac{D_j}{N_j}$ |
|---|---|---|---|---|---|
| 0 | 13 | 0 | 0 | | |
| 1 | 13 | 1 | 0 | 1/13 | 12/13 |
| 5 | 12 | 1 | 1 | 1/12 | 11/12 |
| 12 | 10 | 2 | 0 | 2/10 | 8/10 |
| 15 | 8 | 1 | 0 | 1/8 | 7/8 |
| 19 | 7 | 1 | 2 | 1/7 | 6/7 |
| 24 | 4 | 1 | 0 | 1/4 | 3/4 |

Then the Kaplan-Meier estimate of the survival function is

| $t$ | $\hat{S}(t)$ |
|---|---|
| $0 \le t < 1$ | 1.000 |
| $1 \le t < 5$ | 0.923 |
| $5 \le t < 12$ | 0.846 |
| $12 \le t < 15$ | 0.677 |
| $15 \le t < 19$ | 0.592 |
| $19 \le t < 24$ | 0.508 |
| $24 \le t \le 30$ | 0.381 |

(iii) A suitable graph is shown below.




(iv) Since qualifications are assumed to happen before censorships, swapping $D$ and $H$ will have no effect at all, as the order of the two events will simply be reversed, OR is equivalent to re-labelling D as H and H as D, which clearly does not affect the calculations.

Swapping $B$ and $L$ will reduce the value of $N_j$ at time 10 (replacing time 5)

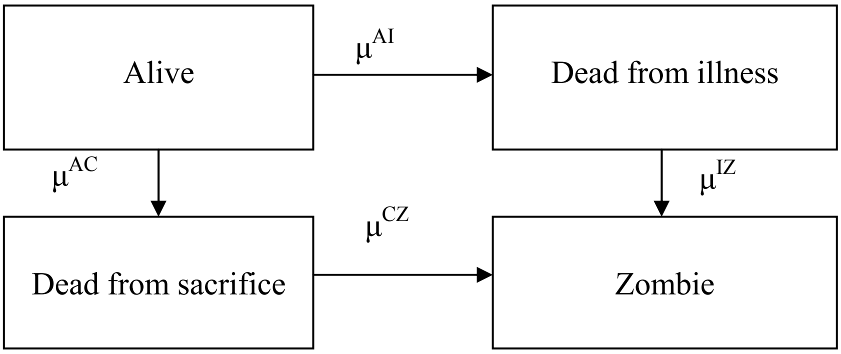
which will increase the value of the hazard at duration 10 compared with that previously at duration 5

This will increase $S(t)$ from durations 5 to immediately before duration 10 and reduce it at durations 10 and over.

*In part (i) candidates who gave alternative answers as to whether the form of censoring is present were given credit if the reason was sensible and consistent. So, for example, candidates who stated that interval censoring was present because we do not know exactly when within the day events occurred were given credit. The most common error in part (ii) was performing the calculations with "leaving" as the event. The question asked for the "non-qualification function". This means the survival function with qualification as the event, by analogy with an analysis of mortality in which the survival function can be described as a "non-death" function when death is the event.*

**10**

(i)




(ii) Let the states be labelled as follows:

Alive, $A$
Dead from illness, $I$
Dead from sacrifice, $C$
Zombie, $Z$

Let the number of transitions observed between states $i$ and $j$ be $d^{ij}$

and let the transition rate between states $i$ and $j$ be $\mu^{ij}$.

Let the observed waiting time in state $i$ be $v^i$

The likelihood of the data can be written as follows:

$$L \propto \exp[(-\mu^{AI} - \mu^{AC})v^A]\exp(-\mu^{CZ}v^C)\exp(-\mu^{IZ}v^I)(\mu^{AI})^{d^{AI}}(\mu^{AC})^{d^{AC}}(\mu^{CZ})^{d^{CZ}}(\mu^{IZ})^{d^{IZ}}$$

(iii) Taking logarithms of the likelihood we have:

$$\log_e L = -\mu^{AI}v^A + d^{AI}\log(\mu^{AI}) + \text{terms not depending on } \mu^{AI}.$$

Differentiating this with respect to $\mu^{AI}$ gives:

$$\frac{d(\log_e L)}{d\mu^{AI}} = -v^A + \frac{d^{AI}}{\mu^{AI}},$$

and setting the derivative equal to zero produces the maximum likelihood estimate of $\mu^{AI}$:

$$\hat{\mu}^{AI} = \frac{d^{AI}}{v^A}.$$

This is a maximum as the second derivative

$$\frac{d^2(\log_e L)}{(d\mu^{AI})^2} = -\frac{d^{AI}}{(\mu^{AI})^2}$$

is necessarily negative.

(iv) Using the census formula, we estimate $v^A$ as follows

$$v^A = 0.5(P_0 + P_{10}) = 0.5(3{,}189 + 2{,}811) = 3{,}000.$$

assuming the population of aliens varies linearly over the ten years between the censuses.

The estimated annual death rate from illness is therefore

$$\frac{369}{3000} = 0.123,$$

and the estimated rate of death through sacrifice over the ten years is

$$\frac{231}{3000} = 0.077.$$

(v) (a) The probability that an alien is still alive in ten years' time is given by the formula

$$_{10}p_x^{AA} = \exp\left[-\int_0^{10}(\mu^{AI} + \mu^{AC})du\right] = \exp[-(0.077 + 0.123)10]$$

$$= \exp(-2) = 0.135.$$

(b) Since we are only interested in whether the alien is dead, not what cause (s)he died from,

and since the rate at which aliens become zombies does not depend on cause of death, we can combine the two states "Dead from illness" and "Dead from sacrifice", into a single state "Dead".

For an alien to be Dead in 10 years time (s)he must have survived for $u$ alien years ($0 < u < 10$), died at time $u$, and then survived in the Dead state (i.e. not become a Zombie) for a duration equal to $10-u$ alien years.

The probability density of this happening for any given value of $u$ is

exp $[-0.2u]$ survival Alive for a period $u$
$\times$
0.2 $du$ (here we ignore $o(du)$)
$\times$
exp$[-0.1(10-u)]$ survival Dead for a period $10 - u$

which is

$$0.2 * \exp[-(1+0.1u)]du = 0.2 * \exp(-1)\exp(-0.1u)du$$

$$= 0.0736\exp(-0.1u)du$$

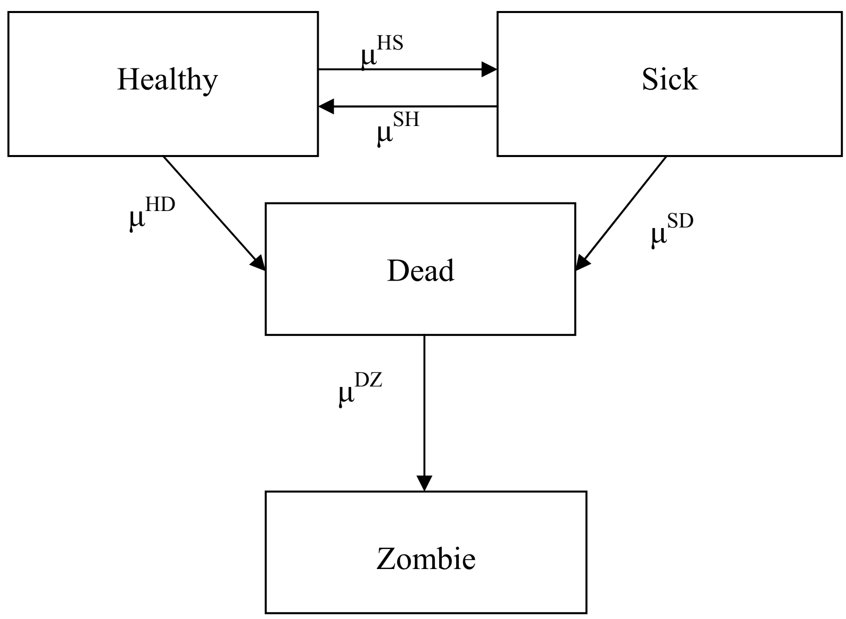
The required probability is obtained by integrating this expression over all values of $u$ from 0 to 10.

This is

$$\int_0^{10} 0.0736\exp(-0.1u)du = \frac{0.0736}{-0.1}\left[\exp(-0.1u)\right]_0^{10}$$

$$= \frac{0.0736}{0.1}\left[1-0.3678)\right] = 0.465$$

*Parts (i), (ii) and (iii) of this question were well answered by most candidates, but there were few good attempts at parts (iv), (v) and (vi). A minority of candidates produced an alternative transition diagram in part (i) as follows:*



*Full credit was given for this, and for answers to parts (ii) and (iii) which were consistent with it. In part (iii) some candidates derived the maximum likelihood estimate by applying the correct method to the wrong transition. In part (v)(b) it was possible to write the integral as follows:*

$$\int_0^{10} \exp[-0.2(10-w)] * 0.2 * \exp(-0.1w)\, dw.$$

*The evaluation is:*

$$0.2 * \int_0^{10} \exp(-2+0.2w)\exp(-0.1w)\, dw$$
$$= 0.2\int_0^{10} \exp(-2+0.1w)\, dw$$
$$= 0.2\exp(-2)\int_0^{10} \exp(0.1w)\, dw$$
$$= \frac{0.2}{0.1}\exp(-2)[\exp(1)-\exp(0)]$$
$$= 2 * 0.1353 * (2.718-1) = 0.465$$

**END OF EXAMINERS' REPORT**

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINATION

24 April 2013 (pm)

## Subject CT4 – Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 11 questions, beginning your answer to each question on a separate sheet.*

5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** Describe the differences between deterministic and stochastic models. [4]

**2** In the context of a survival model:

(i) Define right censoring, Type I censoring and Type II censoring. [3]

(ii) Give an example of a practical situation in which censoring would be informative. [2]

[Total 5]

**3** For both of the following sets of four stochastic processes, place each process in a separate cell of the following table, so that each cell correctly describes the state space and the time space of the process placed in it. Within each set, all four processes should be placed in the table.

| | | *Time Space* | |
|---|---|---|---|
| | | *Discrete* | *Continuous* |
| *State Space* | *Discrete* | | |
| | *Continuous* | | |

(a) General Random Walk, Compound Poisson Process, Counting Process, Poisson Process

(b) Simple Random Walk, Compound Poisson Process, Counting Process, White Noise

[5]

**4** The mortality of a certain species of furry animal has been studied. It is known that at ages over five years the force of mortality, μ, is constant, but the variation in mortality with age below five years of age is not understood. Let the proportion of furry animals that survive to exact age five years be ${}_5p_0$.

(i) Show that, for furry animals that die at ages over five years, the average age at death in years is $\frac{5\mu+1}{\mu}$. [1]

(ii) Obtain an expression, in terms of μ and ${}_5p_0$, for the proportion of all furry animals that die between exact ages 10 and 15 years. [3]

A new investigation of this species of furry animal revealed that 30 per cent of those born survived to exact age 10 years and 20 per cent of those born survived to exact age 15 years.

(iii) Calculate μ and ${}_5p_0$. [3]

[Total 7]

**5** Population censuses in a certain country are taken each year on the President's birthday, provided that the President's astrological advisor deems the taking of a census favourable. Censuses record the age of every inhabitant in completed years (that is, *curtate* age). Deaths in this country are registered as they happen, and classified according to age nearest birthday at the time of death.

Below are some data from the three most recent censuses.

| *Age in completed years* | *Population 2006 (thousands)* | *Population 2009 (thousands)* | *Population 2010 (thousands)* |
|---|---|---|---|
| 64 | 300 | 320 | 350 |
| 65 | 290 | 310 | 330 |
| 66 | 280 | 300 | 320 |

Between the censuses of 2006 and 2009 there were a total of 3,000 deaths to inhabitants aged 65 nearest birthday, and between the censuses of 2009 and 2010 there were a total of 1,000 deaths to inhabitants aged 65 nearest birthday.

(i) Estimate, stating any assumptions you make, the death rate at age 65 years for each of the following periods:

- the period between the 2006 and 2009 censuses
- the period between the 2009 and 2010 censuses [6]

(ii) Explain the exact age to which your estimates apply. [1]

[Total 7]

**6** (i) State the form of the hazard function for the Cox Regression Model, defining all the terms used. [2]

(ii) State two advantages of the Cox Regression Model. [2]

Susanna is studying for an on-line test. She has collected data on past attempts at the test and has fitted a Cox Regression Model to the success rate using three covariates:

| | |
|---|---|
| Employment | $Z_1$ = 0 if an employee, and 1 if self-employed |
| Attempt | $Z_2$ = 0 if first attempt, and 1 if subsequent attempt |
| Study time | $Z_3$ = 0 if no study time taken, and 1 if study time taken |

Having analysed the data Susanna estimates the parameters as:

| | |
|---|---|
| Employment | 0.4 |
| Attempt | –0.2 |
| Study time | 1.15 |

Bill is an employee. He has taken study time and is attempting the test for the second time. Ben is self-employed and is attempting the test for the first time without taking study time.

(iii) Calculate how much more or less likely Ben is to pass, compared with Bill. [3]

Susanna subsequently discovers that the effect of the number of attempts is different for employees and the self-employed.

(iv) Explain how the model could be adjusted to take this into account. [2]
[Total 9]

**7** The Shining Light company has developed a new type of light bulb which it recently tested. 1,000 bulbs were switched on and observed until they failed, or until 500 hours had elapsed. For each bulb that failed, the duration in hours until failure was noted. Due to an earth tremor after 200 hours, 200 bulbs shattered and had to be removed from the test before failure.

The results showed that 10 bulbs failed after 50 hours, 20 bulbs failed after 100 hours, 50 bulbs failed after 250 hours, 300 bulbs failed after 400 hours and 50 bulbs failed after 450 hours.

(i) Calculate the Kaplan-Meier estimate of the survival function, $S(t)$, for the light bulbs in the test. [6]

(ii) Sketch the Kaplan-Meier estimate calculated in part (i). [2]

(iii) Estimate the probability that a bulb will not have failed after each of the following durations: 300 hours, 400 hours and 600 hours. If it is not possible to obtain an estimate for any of the durations without additional assumptions, explain why. [3]
[Total 11]

**8** During a football match, the referee can caution players if they commit an offence by showing them a yellow card. If a player commits a second offence which the referee deems worthy of a caution, they are shown a red card, and are sent off the pitch and take no further part in the match. If the referee considers a particularly serious offence has been committed, he can show a red card to a player who has not previously been cautioned, and send the player off immediately.

The football team manager can also decide to substitute one player for another at any point in the match so that the substituted player takes no further part in the match. Due to the risk of a player being sent off, the manager is more likely to substitute a player who has been shown a yellow card. Experience shows that players who have been shown a yellow card play more carefully to try to avoid a second offence.

The rate at which uncautioned players are shown a yellow card is 1/10 per hour.

The rate at which those players who have already been shown a yellow card are shown a red card is 1/15 per hour.

The rate at which uncautioned players are shown a red card is 1/40 per hour.

The rate at which players are substituted is 1/10 per hour if they have not been shown a yellow card, and 1/5 if they have been shown a yellow card.

(i) Sketch a transition graph showing the possible transitions between states for a given player. [2]

(ii) Write down the compact form of the Kolmogorov forward equations, specifying the generator matrix. [3]

A football match lasts 1.5 hours.

(iii) Solve the Kolmogorov equation for the probability that a player who starts the match remains in the game for the whole match without being shown a yellow card or a red card. [2]

(iv) Calculate the probability that a player who starts the match is sent off during the match without previously having been cautioned. [3]

Consider a match that continued indefinitely rather than ending after 1.5 hours.

(v) (a) Derive the probability that in this instance a player is sent off without previously having been cautioned.

(b) Explain your result. [2]

[Total 12]

**9** A life office compared the mortality of its policyholders in the age range 30 to 60 years inclusive with a set of mortality rates prepared by the Continuous Mortality Investigation (CMI). The mortality of the life office policyholders was higher than the CMI rates at ages 30–35, 38–41, 45–50 and 54–59 years inclusive, and lower than the CMI rates at all other ages in the age range.

(i) Perform two tests of the null hypothesis that the underlying mortality of the life office policyholders is represented by the CMI rates. [7]

(ii) Comment on your results from part (i). [2]

(ii) Explain the problem which duplicate policies cause in the context of the CMI mortality investigations. [3]
[Total 12]

**10** (i) State the Markov property. [1]

A certain non-fatal medical condition affects adults. Adults with the condition suffer frequent episodes of blurred vision. A study was carried out among a group of adults known to have the condition. The study lasted one year, and each participant in the study was asked to record the duration of each episode of blurred vision. All participants remained under observation for the entire year.

The data from the study were analysed using a two-state Markov model with states:

1. not suffering from blurred vision.
2. suffering from blurred vision.

Let the transition rate from state $i$ to state $j$ at time $x+t$ be $\mu_{x+t}^{ij}$, and let the probability that a person in state $i$ at time $x$ will be in state $j$ at time $x+t$ be ${}_t p_x^{ij}$.

(ii) Derive from first principles the Kolmogorov forward equation for the transition from state 1 to state 2. [5]

The results of the study were as follows:

| | |
|---|---|
| Participant-days in state 1 | 21,650 |
| Participant-days in state 2 | 5,200 |
| Number of transitions from state 1 to state 2 | 4,330 |
| Number of transitions from state 2 to state 1 | 4,160 |

Assume the transition intensities are constant over time.

(iii) Calculate the maximum likelihood estimates of the transition intensities from state 1 to state 2 and from state 2 to state 1. [1]

(iv) Estimate the probability that an adult with the condition who is presently not suffering from blurred vision will be suffering from blurred vision in 3 days' time. [6]
[Total 13]

**11** (i) Explain what is meant by a time inhomogeneous Markov chain and give an example of one. [2]

A No Claims Discount system is operated by a car insurer. There are four levels of discount: 0%, 10%, 25% and 40%. After a claim-free year a policy holder moves up one level (or remains at the 40% level). If a policy holder makes one claim in a year he or she moves down one level (or remains at the 0% level). A policy holder who makes more than one claim in a year moves down two levels (or moves to or remains at the 0% level). Changes in level can only happen at the end of each year.

(ii) Describe, giving an example, the nature of the boundaries of this process. [2]

(iii) (a) State how many states are required to model this as a Markov chain.

(b) Draw the transition graph. [2]

The probability of a claim in any given month is assumed to be constant at 0.04. At most one claim can be made per month and claims are independent.

(iv) Calculate the proportion of policyholders in the long run who are at the 25% level. [6]

(v) Discuss the appropriateness of the model. [3]

[Total 15]

## END OF PAPER

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS' REPORT

April 2013 examinations

## 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.

D C Bowie
Chairman of the Board of Examiners

July 2013

## General comments on Subject CT4

Subject CT4 comprises five main sections: (1) 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); (2) stochastic processes, especially Markov chains and Markov jump processes; (3) models of a random variable measuring future lifetime; (4) the calculation of exposed to risk and the application of the principle of correspondence; (5) 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.

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.

## Comments on the April 2013 paper

The general performance was slightly inferior to that in April 2011 or April 2012, but better than that in September 2012. Despite this, well-prepared candidates scored highly across the whole paper, with an above average proportion of candidates scoring 70 per cent or more. The comments that follow the questions concentrate on areas where candidates could have improved their performance. Candidates approaching the subject for the first time are advised to include revision of these areas in their preparation.

## 1

A stochastic model is one that recognises the random nature of the input components.

A model that does not contain any random component is deterministic in nature.

In a deterministic model, the output is determined once the set of fixed inputs and the relationships between them have been defined.

By contrast, in a stochastic model the output is random in nature. The output is only a snapshot or an estimate of the characteristics of the model for a given set of inputs.

A deterministic model is really just a special (simplified) case of a stochastic model.

A deterministic model will give one set of results of the relevant calculations for a single scenario; a stochastic model will be run many times with the same input and gives distributions of the relevant results for a distribution of scenarios

The results for a deterministic model can often be obtained by direct calculation.

The results of stochastic models often require Monte Carlo simulation, although some stochastic models can have an analytical solution.

Correlations can be important in stochastic models as they indicate when the behaviour of one variable is associated with that of another.

Stochastic models are more complex and more difficult to interpret than deterministic models and so require more expertise, expense and computer power.

*Not all the points listed above were required for full marks. Credit was also given for sensible points not included in the above list.*

## 2

(i) **Right censoring.** The duration to the event is not known exactly, but is known to exceed some value.
OR
the censoring mechanism cuts short observations in progress.

**Type I censoring.** The durations at which observations will be censored are specified in advance.

**Type II censoring**. Observation continues until a pre-determined number/proportion of individuals have experienced the event of interest.

(ii) An investigation of mortality based on life office data in which individuals are censored who discontinue paying their premiums.

Those whose premiums lapse tend, on average, to be in better health than do those who carry on paying their premiums.

*In part (ii) any suitable example was given credit. However, for full credit it was necessary to describe a comparison between the risk of the event happening in the censored and uncensored observations (e.g. "in better health than" or "less likely to die than"). Most candidates made a good attempt at this question.*

## 3

(a)

<table>
<tr><th></th><th></th><th colspan="2">Time Space</th></tr>
<tr><th></th><th></th><th>Discrete</th><th>Continuous</th></tr>
<tr><td rowspan="2">State Space</td><td>Discrete</td><td>Counting process</td><td>Poisson process</td></tr>
<tr><td>Continuous</td><td>General random walk</td><td>Compound Poisson process</td></tr>
</table>

(b)

<table>
<tr><th></th><th></th><th colspan="2">Time Space</th></tr>
<tr><th></th><th></th><th>Discrete</th><th>Continuous</th></tr>
<tr><td rowspan="2">State Space</td><td>Discrete</td><td>Simple random walk</td><td>Counting process</td></tr>
<tr><td>Continuous</td><td>White noise</td><td>Compound Poisson process</td></tr>
</table>

*This question was answered well, with many candidates scoring full marks. Some candidates lost marks by failing to follow the instructions in the question precisely. To obtain full credit, candidates were required to place the processes in grids like those shown above with ONE process in each of the four cells. What is shown above is the only solution which fulfils this criterion for groups (a) and (b). In some cases, processes could correctly be placed in cells other than those shown in the grids above, and credit was given for each process thus classified correctly.*

**4**

(i) If the force of mortality, μ, is constant, then the expected waiting time is $\frac{1}{\mu}$.

Hence expected age at death is $5 + \frac{1}{\mu} = \frac{5\mu+1}{\mu}$.

[1]

(ii) EITHER

We need ${}_{10}p_0 - {}_{15}p_0$.

Since ${}_{x}p_0 = {}_{x-5}p_5 \cdot {}_5p_0$

and for $x > 5$, ${}_{x}p_5 = e^{-\mu x}$,

then

$${}_{10}p_0 - {}_{15}p_0 = {}_5p_5 \cdot {}_5p_0 - {}_{10}p_5 \cdot {}_5p_0 = {}_5p_0 e^{-5\mu} - {}_5p_0 e^{-10\mu} = {}_5p_0(e^{-5\mu} - e^{-10\mu}).$$

OR

We need ${}_{10}p_0 \cdot {}_5q_{10}$

$= {}_{10}p_0(1 - {}_5p_{10})$

Since for $x > 5$, ${}_{x}p_5 = e^{-\mu x}$,

$${}_{10}p_0(1 - {}_5p_{10}) = {}_5p_0({}_5p_5 - {}_{10}p_5) = {}_5p_0(e^{-5\mu} - e^{-10\mu})$$

[3]

(iii) EITHER

${}_5p_0 e^{-5\mu} = 0.3$ and ${}_5p_0 e^{-10\mu} = 0.2$.

So $\quad \frac{{}_5p_0 e^{-5\mu}}{{}_5p_0 e^{-10\mu}} = \frac{0.3}{0.2}$

and $e^{-5\mu} = 1.5 e^{-10\mu}$

so that $-5\mu = \log_e 1.5 - 10\mu$

$5\mu = 0.4055$
$\mu = 0.0811.$

Therefore ${}_5p_0 e^{-5(0.0811)} = 0.3$

and $\quad {}_5p_0 = \dfrac{0.3}{e^{-5(0.0811)}} = 0.4500.$

OR

${}_{10}p_0 = {}_5p_0 \cdot {}_5p_5 = 0.3$

With a constant force after age 5 years, ${}_5p_5 = {}_5p_{10}$,

so ${}_{15}p_0 = {}_5p_0 \cdot {}_{10}p_5 = {}_5p_0 \cdot {}_5p_5 \cdot {}_5p_{10} = {}_5p_0({}_5p_5)^2 = 0.2$ .

Hence ${}_5p_5 = \dfrac{0.2}{0.3}$

and ${}_5p_0 = \dfrac{0.3}{{}_5p_5} = \dfrac{(0.3)^2}{0.2} = 0.45.$

Then $\mu = -\dfrac{\log_e {}_5p_5}{5} = \dfrac{0.4055}{5} = 0.0811.$

*Answers to this question were extremely disappointing. Few candidates could even attempt part (i) correctly, and there were similarly few correct attempts at parts (ii) and (iii). In part (iii) the question asked "calculate" so candidates giving both correct numerical answers scored full credit. If one of either μ or ${}_5p_0$ was correct, a minimum of +2 was scored. Where candidates made the same theoretical error in parts (ii) and (iii), the error was only penalised once.*

**5**

(i) We adjust the exposed to risk so that the age definition corresponds with that of the deaths data.

Let the population at age 65 nearest birthday be $P_{65}$ and let the central exposed to risk at age 65 nearest birthday be $E_{65}^c$.

In 2006 $P_{65}$ = 0.5(300,000 + 290,000) = 295,000
In 2009 $P_{65}$ = 0.5(320,000 + 310,000) = 315,000

In 2010 $P_{65}$ = 0.5(350,000 + 330,000) = 340,000,

assuming that birthdays are uniformly distributed across calendar time.

Using the census approximation (trapezium method) for the period 2006-2009 then

assuming that the population varies linearly between census dates,

$E^c_{65}$ = 1.5(295,000 + 315,000) = 915,000

and for the period 2009–2010

$E^c_{65}$ = 0.5(315,000 + 340,000) = 327,500.

Assuming that the force of mortality is constant within each year of age

$\mu_{65} = \dfrac{3,000}{915,000} = 0.003279$ for the period 2006–2009, and

$\mu_{65} = \dfrac{1,000}{327,500} = 0.003053$ for the period 2009–2010.

We also assuming that the President doesn't change (so the birthday is on the same day each year), or if the President does change the new President's birthday is the same as the birthday of the old President.

(ii) The rate interval is the life year, starting at age $x$ – 0.5.

The age in the middle of the rate interval is thus $x$, so the estimate relates to exact age 65 years.

*A common error in part (i) was to use equal time periods, whereas the period 2006-2009 is three years and 2009-2010 only one year. For full credit, the assumptions had to appear in the script close to the relevant bit of calculation. Candidates who listed many assumptions, both necessary and unnecessary, in a block at the end of the answer were penalised. In part (i), some candidates calculated $q_x$ rather than $\mu_x$. Full credit was given for this provided that the initial exposed-to-risk was used as the denominator. In part (ii) the age to which $q_x$ applies is 64.5 years (i.e. the age at the start of the rate interval), and for full credit the answers to parts (i) and (ii) had to be consistent.*

**6**

(i) $\lambda(t{:}Z_{\mathrm{i}}) = \lambda_0(t) \exp(\beta Z_i^T)$

Where:
$\lambda(t{:}Z_{\mathrm{i}})$ is the hazard at time $t$
$\lambda_0(t)$ is the baseline hazard
$Z_i$ is a vector of covariates
$\beta$ is a vector of regression parameters

(ii) It ensures the hazard is always positive.

The log-hazard is linear.

You can ignore the shape of the baseline hazard and calculate the effect of covariates directly from the data.

It is widely available in standard computer packages OR is a popular, well-established model.

(iii) Ben, self-employed, first attempt, no study has hazard $\lambda_0(t) \exp(0.4)$

Bill, employee, re-sit, study leave has hazard $\lambda_0(t) \exp(0.95)$

So Ben is only $\exp(-0.55) = 57.7\%$ as likely to pass as Bill OR 42.3% less likely to pass than Bill.

OR

Bill is 73% more likely to pass than Ben

(iv) The model could be adjusted by including a covariate measuring the interaction between the number of attempts and employment status.

The covariate would be equal to $Z_1Z_2$ and would take the value 1 for a self-employed person on his or her second or subsequent attempt, and 0 otherwise.

The effect of the number of attempts for an employee would be equal to $\exp(\beta_2)$, where $\beta_2$ is the parameter related to $Z_2$, For a self-employed person, the effect of the number of attempts would be equal to $\exp(\beta_2 + \beta_3)$, where $\beta_3$ is the parameter related to the interaction term.

*This question was well answered by many candidates. In part (iii) the question asked candidates to "calculate" so the correct numerical answer scored full credit. However a common error was to use ambiguous or incorrect wording in the final comparison (e.g. Bill* *is 57.7 per cent less likely to pass than Ben). In part (iv) no credit was given for the addition of covariates with no bearing on the interaction term. However redundant parameters were not penalised provided the modification to the model allowed the interaction to be quantified.*

**7**

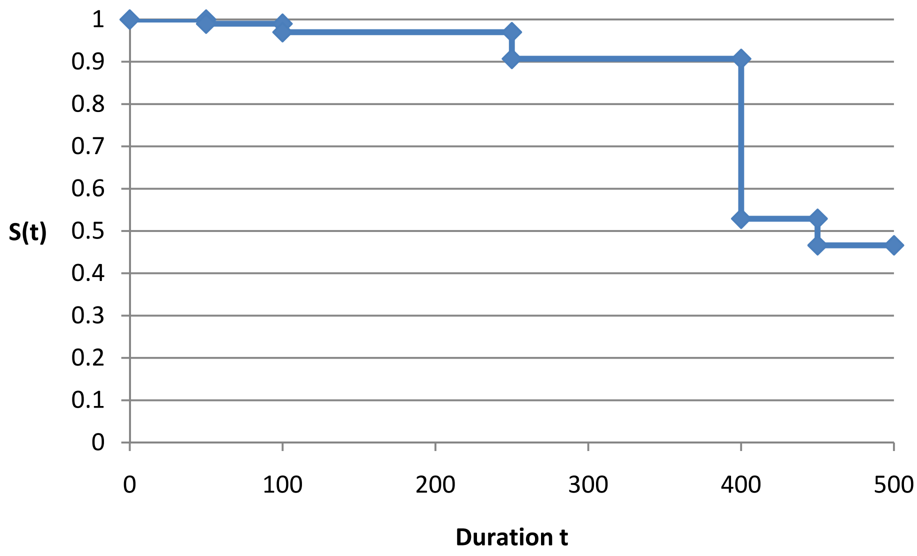
(i)

| $t_j$ | $N_j$ | $d_j$ | $c_j$ | $d_j / N_j$ | $1 - d_j / N_j$ | |
|---|---|---|---|---|---|---|
| 0 | 1,000 | | | | | |
| 50 | 1,000 | 10 | 0 | 0.0100 | 0.9900 | or 99/100 |
| 100 | 990 | 20 | 0 | 0.0202 | 0.9798 | or 97/99 |
| 200 | 970 | 0 | 200 | | | |
| 250 | 770 | 50 | 0 | 0.0649 | 0.9351 | or 72/77 |
| 400 | 720 | 300 | 0 | 0.4167 | 0.5833 | or 7/12 |
| 450 | 420 | 50 | 370 | 0.1190 | 0.8810 | or 37/42 |

The Kaplan-Meier estimate is $\hat{S}(t) = \prod_{t_j \le t} (1 - \frac{d_j}{n_j})$

| $t$ | *Kaplan-Meier estimate of S(t)* | |
|---|---|---|
| $0 \le t < 50$ | 1.0000 | or 1 |
| $50 \le t < 100$ | 0.9900 | or 99/100 |
| $100 \le t < 250$ | 0.9700 | or 97/100 |
| $250 \le t < 400$ | 0.9070 | or 1,746/1,925 |
| $400 \le t < 450$ | 0.5291 | or 291/550 |
| $450 \le t < 500$ | 0.4661 | or 3,589/7,700 |

(ii)



(iii) $S(300) = 0.9070$.

$S(400) = 0.5291$.

$S(600)$ cannot be estimated without additional assumptions

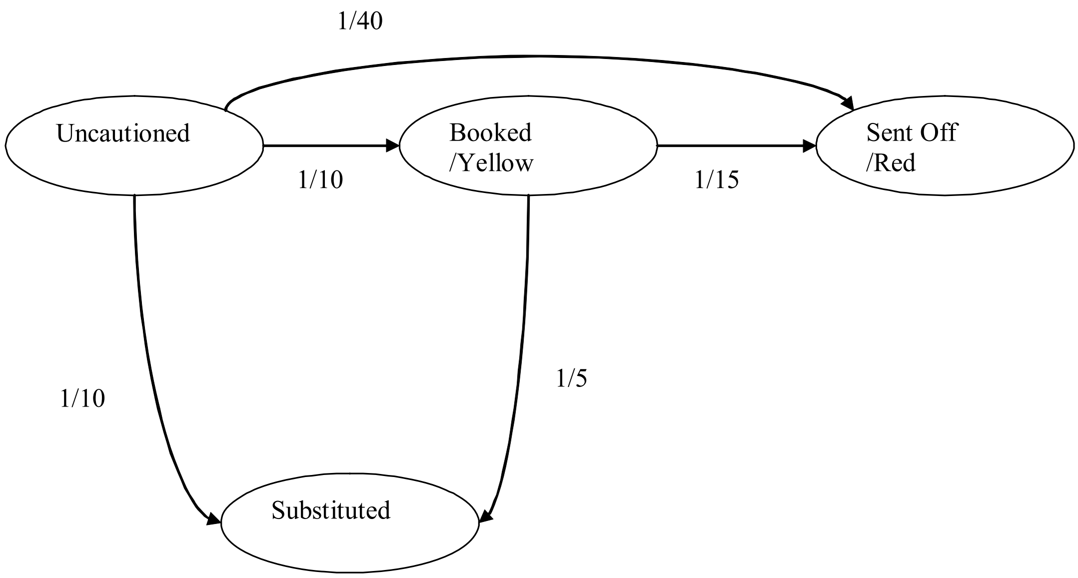
as it lies outside the range of our data.

[3]
[Total 11]

*This question was very well answered, with many candidates scoring 10 or more marks out of a possible 11. Some of the sketches in part (ii) were very scrappy and were penalised: though great accuracy was not required, the sketch did need to be sufficiently clear to demonstrate that the candidate understood the nature of the function being plotted. In part (iii) some candidates suggested an assumption which would enable them to give an answer for S(600). Such candidates were given full credit provided they explained why the assumption was needed, and provided that the stated assumption was consistent with the numerical answer offered.*

**8**

(i)



(ii) $\frac{d}{dt}P(t) = P(t)A$

where generator matrix

$$A = \begin{pmatrix} -9/40 & 1/10 & 1/40 & 1/10 \\ 0 & -4/15 & 1/15 & 1/5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

In order of states $\{U, Y, R, S\}$

(iii) $\frac{d}{dt}P_{UU}(t) = -\frac{9}{40}P_{UU}(t)$

$P_{UU}(t) = \exp\left(-\frac{9}{40}t\right) + const = \exp\left(-\frac{9}{40}t\right)$ as looking for probability on pitch throughout match.

At end $t = 3/2$ so require exp(−27/80) = 71.36%.

(iv) Prob[sent off without being booked] = $\int_{s=0}^{3/2} P_{UU}(s).\frac{1}{40}ds$

$$= \int_{s=0}^{3/2} \exp(-\frac{9}{40}s).\frac{1}{40}ds$$

$$= \left[-\frac{1}{9}\exp\left(-\frac{9}{40}s\right)\right]_0^{3/2} = \frac{1}{9}\left(1 - \exp\left(-\frac{27}{80}\right)\right) = 0.03183$$

(v) (a) EITHER

The upper limit of the integral tends to infinity

so result becomes 1/9.

OR

We need

Pr[sent off directly]/Pr[leaves state U]

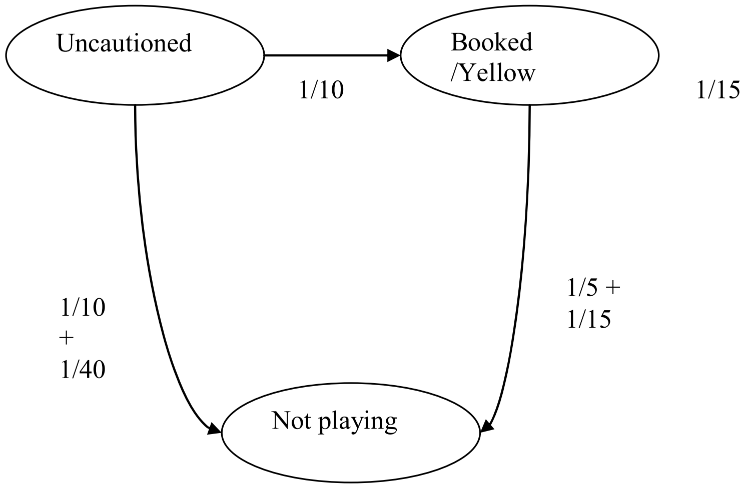
$$= \frac{1/40}{9/40} = \frac{1}{9}$$

(b) This is the ratio of the transition rate to "straight to sent off" to the total transition rate out of state $U$.

*This was one of the more demanding questions on the paper, and a high proportion of candidates struggled to get past part (ii). In part (i) the transition rates were not required. In part (iii) the question asked candidates to solve an equation, so for full credit the equation had to be written down, and the method of solution described. In part (v) candidates who used the rationale in (b) to do the calculation in (a) scored full credit.*

*Candidates who interpreted the question in a manner not intended, but instead combined the categories "sent off" and "substituted" were not penalised. This interpretation leads to a three state solution for parts (i) and (ii) as follows.*

*(i)*




*(ii)* $$\frac{d}{dt}P(t) = P(t)A$$

*where generator matrix*

$$A = \begin{pmatrix} \frac{-9}{40} & \frac{1}{10} & \frac{1}{8} \\ 0 & \frac{-4}{15} & \frac{4}{15} \\ 0 & 0 & 0 \end{pmatrix}$$

*In order of states {U, Y, N}* [3]

*This was given full credit in parts (i), (ii) and (iii). The answer to part (iii) is the same as for the four-state solution. Credit was given in parts (iv) and (v) for following this alternative through correctly.*

**9**

(i) **Signs Test**

Under the null hypothesis that the underlying mortality of the life office policyholders is the same as the CMI mortality,

the number of positive deviations is distributed Binomial($m$,0.5)

THEN EITHER ALTERNATIVE 1 (NORMAL APPROXIMATION)

Here we have $m = 31$, so as $m > 20$ we can use the Normal approximation, that the number of positive deviations is distributed Normal ($m$/2, $m$/4).

the number of positive deviations is Normal $\left(\frac{31}{2}, \frac{31}{4}\right)$.

In this case we have 22 positive deviations.

The $z$-score corresponding to 22 is $\frac{22-15.5}{\sqrt{7.75}} = \frac{6.5}{2.78} = 2.33$

OR Using a continuity correction

The $z$-score corresponding to 22 is $\frac{21.5-15.5}{\sqrt{7.75}} = \frac{6}{2.78} = 2.16$

Using a 2-tailed test, we reject the null hypothesis at the 5% level of significance if $|z|>1.96$.

Since 2.33 (or 2,16) > 1.96 we reject the null hypothesis.

OR ALTERNATIVE 2 (EXACT CALCULATION)

In this case we have 22 positive deviations.
The probability of observing exactly 22 positive deviations is 0.009388
OR
The probability of observing ≥ 22 positive deviations is 0.014725

Using a 2-tailed test, we reject the null hypothesis at the 5% level of significance if the probability is <0.025

Since 0.014725 < 0.025 we reject the null hypothesis.

**Grouping of Signs Test**

Define the test statistic:

$G$ = Number of groups of positive deviations.

THEN EITHER ALTERNATIVE 1 (NORMAL APPROXIMATION)

Since $m$ = 31 (which is ≥ 20), we can use a Normal approximation as follows:

$$G \sim \text{Normal}\left(\frac{n_1(n_2+1)}{n_1+n_2}, \frac{(n_1 n_2)^2}{(n_1+n_2)^3}\right).$$

In this case $m$ = 31, $n_1$ = 22 and $n_2$ = 9.

Thus $G \sim \text{Normal}(7.10, 1.32)$.

We have 4 groups of positive signs.

The $z$-score corresponding to 4 is $\dfrac{4-7.10}{\sqrt{1.32}} = \dfrac{-3.10}{1.15} = -2.70$

Using a 1-tailed test, we reject the null hypothesis at the 5% level of significance if $z < -1.645$.

Since $-2.70 < -1.645$ we reject the null hypothesis.

OR ALTERNATIVE 2 USING TABLE IN GOLD BOOK

$m$ = total number of deviations = 31
$n_1$ = number of positive deviations = 22
$n_2$ = number of negative deviations = 9

We want $k^*$ the largest $k$ such that

$$k = \sum_{t=1}^{x} \binom{n_1-1}{t-1}\binom{n_2+1}{t} \Big/ \binom{n_1+n_2}{n_1} < 0.05$$

We have 4 groups of positive signs.

The test fails at the 5% level if $G \le k^*$

From the table in the Gold Book $k^*$ = 4

Since $G$ is not greater than this, we reject the null hypothesis

(ii) The life office's rates are, overall, different from the CMI rates (actually they are higher).

Additional tests are needed to examine the magnitude of the difference between the two sets of rates.

The shape of the life office's mortality rates is also rather different from the CMI schedule, and this might require further investigation,
OR
The Grouping of Signs test suggests clumping of the deviations.

It is possible that the difference between the shape of the two sets of rates is so small in magnitude as to be negligible.

(iii) We can no longer be sure that we are observing a collection of independent claims.

It is quite possible that two distinct death claims are the result of the death of the same life.

The effect of this is to increase the variance of the number of claims,

by a factor which may depend on age.

This may affect tests based on standardised deviations.

*Answers to this question were very disappointing, especially part (i) . The two tests were often performed in a rather cursory way, with important steps being missed out. In part (i) the null hypothesis only needed stating once. Common errors were to work with only 30 ages or, more seriously, to use eight ages (i.e. to treat each run of consecutive ages of the same sign as a single age).*

## 10

(i) The future development of the process depends only on the state currently occupied and not on any previous history.

OR

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

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 $S$.

(ii) Condition on the state occupied at $x+t$, using the Markov assumption:

$$_{t+dt}p_x^{12} = {}_tp_x^{11}\,{}_{dt}p_{x+t}^{12} + {}_tp_x^{12}\,{}_{dt}p_{x+t}^{22}$$

But by Law of Total Probability $_{dt}p_{x+t}^{22} = 1 - {}_{dt}\,p_{x+t}^{21}$, so

$$_{t+dt}p_x^{12} = {}_tp_x^{11}\,{}_{dt}p_{x+t}^{12} + {}_tp_x^{12}(1 - {}_{dt}\,p_{x+t}^{21}).$$

Now, since by assumption, for small *dt*, ${}_{dt}p^{ij}_{x+t} = \mu^{ij}_{x+t}dt + o(dt)$,

where $\lim_{dt\to 0}\dfrac{o(dt)}{dt} = 0$,

we can substitute to give

$$\begin{aligned}{}_{t+dt}p^{12}_x &= {}_tp^{11}_x\mu^{12}_{x+t}dt + {}_tp^{12}_x(1-\mu^{21}_{x+t}dt) + o(dt)\\ &= {}_tp^{11}_x\mu^{12}_{x+t}dt + {}_tp^{12}_x - {}_tp^{12}_x\mu^{21}_{x+t}dt + o(dt)\end{aligned}$$

so that

$${}_{t+dt}p^{12}_x - {}_tp^{12}_x = {}_tp^{11}_x\mu^{12}_{x+t}dt - {}_tp^{12}_x\mu^{21}_{x+t}dt + o(dt).$$

Dividing by *dt* and taking limits gives

$$\lim_{dt\to 0}\left[\frac{{}_{t+dt}p^{12}_x - {}_tp^{12}_x}{dt}\right] = \frac{d}{dt}\,{}_tp^{12}_x = \lim_{dt\to 0}\left[\frac{{}_tp^{11}_x\mu^{12}_{x+t}dt}{dt} - \frac{{}_tp^{12}_x\mu^{21}_{x+t}dt}{dt} + \frac{o(dt)}{dt}\right]$$

$$= {}_tp^{11}_x\mu^{12}_{x+t} - {}_tp^{12}_x\mu^{21}_{x+t}.$$

(iii) State 1 to state 2: $\mu^{12} = 4{,}330/21{,}650 = 0.2$.

State 2 to state 1: $\mu^{21} = 4{,}160/5{,}200 = 0.8$.

(iv) $\dfrac{d}{dt}\,{}_tp^{12}_x = {}_tp^{11}_x\mu^{12} - {}_tp^{12}_x\mu^{21} = \mu^{12}(1-{}_tp^{12}_x) - {}_tp^{12}_x\mu^{21}$

and substituting the values from the answer to part (iii) gives

$$\frac{d}{dt}\,{}_tp^{12}_x = 0.2(1-{}_tp^{12}_x) - 0.8\,{}_tp^{12}_x = 0.2 - {}_tp^{12}_x.$$

$$\frac{d}{dt}\,{}_tp^{12}_x + {}_tp^{12}_x = 0.2$$

$$\left[\frac{d}{dt}\,{}_tp^{12}_x\right]e^t + {}_tp^{12}_xe^t = \frac{d}{dt}({}_tp^{12}_xe^t) = 0.2e^t$$

$${}_tp^{12}_xe^t = 0.2e^t + C$$

Since ${}_0p^{12}_x = 0$

$C = -0.2$

and ${}_t p_x^{12} = 0.2 - 0.2e^{-t}$.

For $t$ = 3 days, ${}_t p_x^{12} = 0.2 - 0.2e^{-3} = 0.1900$.

*In part (ii) minor variations on the exact derivation given above were permitted, but all the steps were required for full credit. In part (iii) some candidates attempted a solution using integral equations. A relatively common example argued that the required probability could be obtained from:*

$${}_3 p_0^{12} = \int_0^3 e^{-0.2w}.0.2.e^{-0.8(3-w)}dw = 0.2\int_0^3 e^{-2.4}e^{0.6w}dw = 0.2e^{-2.4}\int_0^3 e^{0.6w}dw.$$

*Evaluating this integral produces* ${}_3 p_0^{12} = \frac{0.2}{0.6}e^{-2.4}\left[e^{0.6w}\right]_0^3 = \frac{0.2}{0.6}(e^{-0.6} - e^{-2.4}) = 0.1527$.

*This is incorrect as it ignores the possibility that lives might oscillate between states 1and 2 between t = 0 and t = 3. It only considers those lives who move between states 1 and 2 with exactly one transition and do not return to state 1. However, this alternative shows considerable understanding of the process, and was given some credit.*

## 11

(i) A Markov chain is a discrete time, discrete space Markov process

For a time-inhomogeneous Markov chain, the transition probabilities depend on the absolute values of time, rather than just the time difference.

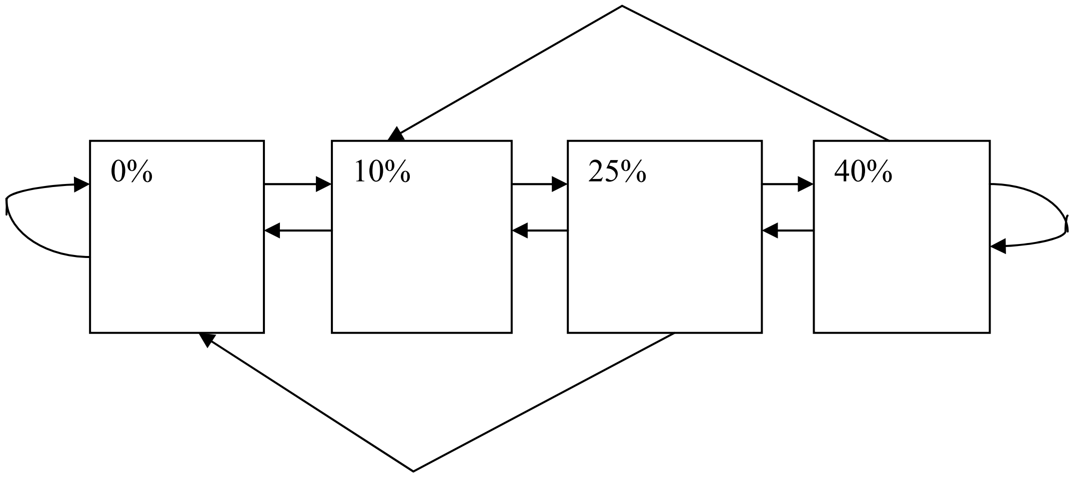
The value of "time" can be represented by many factors, for example the time of year, age or duration.

An example might be a No Claims Discount scheme where the probability of a claim reflects trends in accident frequency over time.

(ii) Both boundaries are mixed as policyholders can either stay in that state for consecutive periods or move back to another state.

E.g. When at the maximum 40% level, a policyholder who makes no claim will stay there the next year, whereas one who makes one claim will drop to the 25% level and one who makes more than one claim will drop to the 10% level.

(iii) Four states are required: 0%, 10%, 25% and 40%.




(iv) Prob [no claims in year] = $0.96^{12} = 0.6127$
Prob [exactly 1 claim in year] = $0.96^{11}$ (0.04) 12 = 0.3064
Prob [more than one claim in a year] = 1 – (0.6127 + 0.3064) = 0.0809

$$\pi\begin{pmatrix} 0.3873 & 0.6127 & 0 & 0 \\ 0.3873 & 0 & 0.6127 & 0 \\ 0.0809 & 0.3064 & 0 & 0.6127 \\ 0 & 0.0809 & 0.3064 & 0.6127 \end{pmatrix} = \pi$$

$$\pi_1 = 0.3873\pi_1 + 0.3873\pi_2 + 0.0809\pi_3 \qquad (1)$$
$$\pi_2 = 0.6127\pi_1 + 0.3064\pi_3 + 0.0809\pi_4 \qquad (2)$$
$$\pi_3 = 0.6127\pi_2 + 0.3064\pi_4 \qquad (3)$$
$$\pi_4 = 0.6127\pi_3 + 0.6127\pi_4 \qquad (4)$$
$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1. \qquad (5)$$

From (4) $\pi_4(1 - 0.6127) = 0.6127\pi_3$
so $\pi_4 = 1.5820\pi_3$

From (3) $\pi_2 = \pi_3(1 - 0.3064 \times 1.5820) / 0.6127$
so $\pi_2 = 0.8411\pi_3$

and (1) gives $\pi_1 = \pi_3(0.0809 + 0.3873 \times 0.8411) / (1 - 0.3837)$
so $\pi_1 = 0.6637\pi_3$

Using (5) we get $\pi_3\left(0.6637+0.8411+1+1.5820\right)=1$

so $\pi_3=0.2447$

so $\pi_4=1.5820\times 0.2447$ $\pi_4=0.3871$

In the long run 24.47% of policyholders are at the 25% level.

(v) Equal probability of an accident in every month is pretty unlikely.

Perhaps more accidents in winter when driving conditions are worse, or in summer, when mileage is higher.

The probability of a second claim may differ from the first and may be dependent upon the level the person is at (e.g. does it make a difference to the future premium?)

Claim probability may depend upon policyholder age/sex or car size/age, and on many other factors (occupation, geographical area, marital status, mileage, where car is stored, etc.)

Claim levels may be affected by the past history of a person's claims (so the process is no longer Markov).

Unrealistic to assume at most one claim per month.

*Parts (i), (iii) and (v) of this question were well answered, though in part (i) it was not often clear how the examples given operated in discrete time. Part (ii) was very poorly attempted. In part (iv) a common error was to assume that the 0.04 claim rate is annual. This gave the answer that just under 4 per cent of policyholders were at the 25 per cent level. Candidates who made this error were penalised for using incorrect probabilities, but were given full credit for solving the equations to obtain the steady-state probabilities. In part (v) sensible suggestions other than those listed were given credit.*

## END OF EXAMINERS' REPORT

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINATION

2 October 2013 (pm)

## Subject CT4 – Models
## Core Technical

*Time allowed: Three hours*

***INSTRUCTIONS TO THE CANDIDATE***

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all nine questions, beginning your answer to each question on a separate sheet.*

5. *Candidates should show calculations where this is appropriate.*

***Graph paper is NOT required for this paper.***

*AT THE END OF THE EXAMINATION*

*Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.*

*In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.*

**1** Data are often subdivided when investigating mortality statistics.

(i) Explain why this is done. [2]

(ii) Discuss one potential problem with sub-dividing mortality data. [2]

(iii) List four factors which are commonly used to sub-divide mortality data. [2]
[Total 6]

**2** The two football teams in a particular city are called United and City and there is intense rivalry between them. A researcher has collected the following history on the results of the last 20 matches between the teams from the earliest to the most recent, where:

U indicates a win for United;
C indicates a win for City;
D indicates a draw.

UCCDDUCDCUUDUDCCUDCC

The researcher has assumed that the probability of each result for the next match depends only on the most recent result. He therefore decides to fit a Markov chain to this data.

(i) Estimate the transition probabilities for the Markov chain. [3]

(ii) Estimate the probability that United will win at least two of the next three matches against City. [3]
[Total 6]

**3** (i) Define a Poisson process. [2]

A bus route in a large town has one bus scheduled every 15 minutes. Traffic conditions in the town are such that the arrival times of buses at a particular bus stop may be assumed to follow a Poisson process.

Mr Bean arrives at the bus stop at 12 midday to find no bus at the stop. He intends to get on the first bus to arrive.

(ii) Determine the probability that the first bus will not have arrived by 1.00 pm the same day. [2]

The first bus arrived at 1.10 pm but was full, so Mr Bean was unable to board it.

(iii) Explain how much longer Mr Bean can expect to wait for the second bus to arrive. [1]

(iv) Calculate the probability that at least two more buses will arrive between 1.10 pm and 1.20 pm. [2]
[Total 7]

**4** (i) State what needs to be assessed when considering the suitability of a model for a particular purpose. [5]

A city has care homes for the elderly. When an elderly person in the city is no longer able to cope alone at home they can move into a care home where they will be looked after until they die.

The city council runs some of the care homes and is reviewing the number of rooms it needs in all the council-run care homes. The following model has been proposed.

The age distribution of the city population over the age of 60 is taken from the latest census. Statistics from the national health service on the average age of entry into a care home and the proportion of elderly who go into council-run care homes are applied to the current population to give the rate at which people enter care homes. A recent national mortality table is then applied to give the rate at which care home residents die.

(ii) Discuss the suitability of this model. [4]

[Total 9]

**5** A motor insurer offers a No Claims Discount scheme which operates as follows. The discount levels are {0%, 25%, 50%, 60%}. Following a claim-free year a policyholder moves up one discount level (or stays at the maximum discount). After a year with one or more claims the policyholder moves down two discount levels (or moves to, or stays in, the 0% discount level).

The probability of making at least one claim in any year is 0.2.

(i) Write down the transition matrix of the Markov chain with state space {0%, 25%, 50%, 60%}. [2]

(ii) State, giving reasons, whether the process is:

(a) irreducible.
(b) aperiodic.

[2]

(iii) Calculate the proportion of drivers in each discount level in the stationary distribution. [4]

The insurer introduces a "protected" No Claims Discount scheme, such that if the 60% discount is reached the driver remains at that level regardless of how many claims they subsequently make.

(iv) Explain, without doing any further calculations, how the answers to parts (ii) and (iii) would change as a result of introducing the "protected" No Claims Discount scheme. [3]

[Total 11]

**6** (i) Explain what is meant by censoring in the context of a mortality investigation. [1]

A trial was conducted on the effectiveness of a new cream to treat a skin condition. 100 sufferers applied the cream daily for four weeks or until their symptoms disappeared if this happened sooner. Some of the sufferers left the trial before their symptoms disappeared.

(ii) Describe two types of censoring that are present and state to whom they apply. [2]

The following data were collected.

| *Number of sufferers* | *Day symptoms disappeared* |
|---|---|
| 2 | 6 |
| 1 | 7 |
| 1 | 10 |
| 2 | 14 |

| *Number of sufferers* | *Day they left the trial* |
|---|---|
| 3 | 2 |
| 1 | 10 |
| 3 | 13 |

(iii) Calculate the Nelson-Aalen estimate of the survival function for this trial. [5]

(iv) Sketch the survival function, labelling the axes. [2]

(v) Estimate the probability that a person using the cream will still have symptoms of the skin condition after two weeks. [1]
[Total 11]

**7** (i) Explain why the Gompertz model is commonly used in investigations of human mortality. [1]

The following model of mortality was used in an investigation of the effects of where someone lives and income on the risk of death.

$$\log_e \mu_x = \alpha + \beta_0 x + \beta_1 U + \beta_2 I \,,$$

where $\mu_x$ is the force of mortality at age $x$, $U$ takes the value 1 if the person lives in an urban area and 0 if the person lives in a rural area, $I$ is annual income in US dollars, and $\alpha$, $\beta_0$, $\beta_1$ and $\beta_2$ are parameters.

(ii) Show that the model is both a Gompertz model and a proportional hazards model. [3]

The estimates of the parameters were $\alpha = -9.0$ $\beta_0 = 0.09$, $\beta_1 = 0.3$ and $\beta_2 = -0.0001$.

(iii) Calculate the predicted force of mortality for an urban resident aged 40 years with an annual income of \$20,000. [2]

(iv) Calculate the additional income that an urban resident must have in order to have the same force of mortality as a rural resident of the same age. [2]

(v) Calculate the 10-year survival probability for an urban resident aged 40 years whose annual income is $20,000. [2]

(vi) Determine the age of a rural resident with the same income as an urban resident aged 40 years, who has the same chance of surviving for the next 10 years. [4]

[Total 14]

**8** Outside an apartment block there is a small car park with three parking spaces. A prospective purchaser of an apartment in the block is concerned about how often he would return in his car to find that there was no empty parking space available. He decides to model the number of parking spaces free at any time using a time homogeneous Markov Jump Process where:

- The probability that a car will arrive seeking a parking space in a short interval $dt$ is $A.dt + o(dt)$.

- For each car which is currently parked, the probability that its owner drives the car away in a short interval $dt$ is $B.dt + o(dt)$.

where $A, B > 0$.

(i) Specify the state space for the above process. [1]

(ii) Draw a transition graph of the process. [2]

(iii) Write down the generator matrix for the process. [2]

(iv) Derive the probability that, given all the parking spaces are full, they will remain full for at least the next two hours. [2]

(v) Explain what is meant by a jump chain. [1]

(vi) Specify the transition matrix for the jump chain associated with this process. [2]

Suppose there are currently two empty parking spaces.

(vii) Determine the probability that all the spaces become full before any cars are driven away. [1]

(viii) Derive the probability that the car park becomes full before the car park becomes empty. [3]

(ix) Comment on the prospective purchaser's assumptions regarding the arrival and departure of cars. [3]

[Total 17]

**9** (i) (a) State three different methods of graduating crude mortality data.

(b) Give, for each method, one advantage and one disadvantage. [5]

An insurance company has graduated the experience of one block of its life business against a standard table, the following is an extract of the data.

| *Age x* | *Exposed to risk* | *Observed deaths* | *Graduated rates* |
|---|---|---|---|
| 30 | 36,254 | 26 | 0.000590 |
| 31 | 37,259 | 20 | 0.000602 |
| 32 | 28,057 | 23 | 0.000617 |
| 33 | 31,944 | 23 | 0.000636 |
| 34 | 30,005 | 26 | 0.000660 |
| 35 | 28,389 | 12 | 0.000689 |
| 36 | 36,124 | 31 | 0.000724 |
| 37 | 28,152 | 22 | 0.000765 |
| 38 | 24,001 | 25 | 0.000813 |
| 39 | 30,448 | 31 | 0.000870 |

(ii) Carry out a test for overall goodness of fit. [5]

(iii) Carry out two other statistical tests to check the validity of the graduation. [6]

(iv) Discuss, with reference to the tests you have performed, whether it would be reasonable for the company to use the graduated rates to price life insurance policies. [3]

[Total 19]

## END OF PAPER

# INSTITUTE AND FACULTY OF ACTUARIES

# EXAMINERS' REPORT

September 2013 examinations

## 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.

D C Bowie
Chairman of the Board of Examiners

December 2013

## General comments on Subject CT4

Subject CT4 comprises five main sections: (1) 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); (2) stochastic processes, especially Markov chains and Markov jump processes; (3) models of a random variable measuring future lifetime; (4) the calculation of exposed to risk and the application of the principle of correspondence; (5) 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.

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.

## Comments on the September 2013 paper

The general performance was very similar to that in April 2013, and better than that in September 2012. Well-prepared candidates scored highly across the whole paper, with an above average proportion of candidates scoring 70 per cent or more. The comments that follow the questions concentrate on areas where candidates could have improved their performance. Candidates approaching the subject for the first time are advised to include these areas in their revision.

**1**

(i) All our models and analyses are based on the assumption that we can observe groups of identical lives (or at least, lives whose mortality characteristics are the same).

In practice, this is never possible.

However, we can at least subdivide our data according to characteristics known, from experience, to have a significant effect on mortality.

This ought to reduce the heterogeneity of each class so formed.

(ii) The number of lives in each subdivision may become small. This will lead to estimates of mortality that are unreliable, with large standard errors.

OR

Information about the factors which affect mortality may be unavailable because it was not asked on the insurance proposal form, or population census

OR

Information about the factors which affect mortality may be unreliable because respondents gave inaccurate or false answers to questions.

(iii) Sex
Age
Type of policy (which often reflects the reason for insuring)
Smoker/non-smoker status
Level of underwriting
Duration in force
Sales channel
Policy size
Occupation of policyholder
Known impairments
Postcode/geographical location
Marital status

*Answers to part (i) of this question were disappointing, with few candidates relating the need for homogeneity to the models we use. Parts (ii) and (iii) were generally well answered. In part (ii) the instruction was to describe a single limitation, so no credit was given for second or subsequent limitations. In part (iii) credit was given in some cases for wording different from that indicated, such as "state of health", or for certain other factors which are known to affect mortality, and about which information is asked, for example, in population censuses. However, genetic factors were not given credit.*

**2**

(i) Need to rearrange data as tally chart of next states:

| *Previous state* | *Number where next state is:* U | C | D |
|---|---|---|---|
| U | 1 | 11 | 111 |
| C | 11 | 111 | 11 |
| D | 11 | 111 | 1 |

So the transition probabilities are estimated as:

| *From/To* | *U* | *C* | *D* |
|---|---|---|---|
| U | 1/6 | 1/3 | 1/2 |
| C | 2/7 | 3/7 | 2/7 |
| D | 1/3 | 1/2 | 1/6 |

(ii) The possible sequences with at least 2 wins for United are:

UUU, UUC, UUD, DUU, CUU, UDU, UCU

The probabilities if the last match was won by City are:

UUU = 2/7*1/6*1/6 = 1/126
UUC = 2/7*1/6*1/3 = 1/63
UUD = 2/7*1/6*1/2 = 1/42
DUU = 2/7*1/3*1/6 = 1/63
CUU = 3/7*2/7*1/6 = 1/49
UDU = 2/7*1/2*1/3 = 1/21
UCU = 2/7*1/3*2/7 = 4/147

OR (quicker)

UUX = 2/7*1/6 = 1/21
DUU = 2/7*1/3*1/6 = 1/63
CUU = 3/7*2/7*1/6 = 1/49
UDU = 2/7*1/2*1/3 = 1/21
UCU = 2/7*1/3*2/7= 4/147

where X refers to any result

Total = 140/882 = 10/63 = 0.15873

*Answers to this question were generally disappointing. In both parts (i) and (ii) the question said "estimate" so some explanation of where the answer is coming from was required for full credit (e.g. in part (i) a statement that $n_{ij}/n_i$ is needed, or a suitable diagram were acceptable). A common error was to use 8 as the denominator for the C row. A more serious error was to use 19 as the denominator for all the transition probabilities. Many candidates* *did not take account of the fact that City had won the last match in the string given and thus only used pairs, rather than triplets, of probabilities.*

**3**

(i) 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 $t$.

Events occur singly and may occur at any time;

the probability that an event occurs during the short time interval from time $t$ to time $t+h$ is approximately equal to $\lambda h$ for small $h$, where the parameter $\lambda$ is the rate of the Poisson process.

OR

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

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

and $o(h)$ is such that $\lim_{h \to 0} \frac{o(h)}{h} = 0$.

OR

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

$N_0 = 0$
$N_t$ has independent increments
$N_t$ has Poisson distributed stationary increments

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

(ii) The probability that no bus arrives in the first 60 minutes is $e^{-60/15} = 0.0183$.

By the memoryless property / by independence of increments / because the holding times are exponential.

(iii) The expected time between buses is 15 minutes.

By independence of increments / memoryless property this is the time Mr Bean can expect to wait for the second bus.

(iv) The rate at which buses arrive per 10 minute period is 10/15.

Therefore the probability of no buses arriving between 1.10 and 1.20 p.m. is $e^{-10/15} = 0.5134$.

The probability of one bus arriving is $e^{-10/15}(10/15) = 0.3423$.

The probability of two or more buses arriving is therefore

$1 - 0.5134 - 0.3423 = 0.1443$.

*Answers to this question were disappointing. Most candidates managed to score reasonably well on part (i). In parts (ii) and (iii) some explanations of the answers were required. In part (iv) several candidates calculated the probability of exactly two buses arriving. Candidates who used an incorrect rate in part (ii) could score full credit for part (iv) if they correctly calculated the probability of two or more buses arriving given the incorrect rate.*

## 4

(i) The objectives of the modelling exercise.

The validity of the model for the purpose to which it is to be put.

The validity of the data to be used.

The validity of assumptions used.

The possible errors associated with the model or parameters used not being a perfect representation of the real world situation being modelled.

The impact of correlations between the random variables that "drive" the model.

The extent of correlations between the various results produced from the model.

The current relevance of models written and used in the past.

The credibility of the data input.

The credibility of the results output.

The dangers of spurious accuracy.

The cost of buying or constructing, and of running the model.

The ease of use and availability of suitable staff to use it.

Risk of model being used incorrectly or with wrong inputs.

The ease with which the model and its results can be communicated.

Compliance with the relevant regulations.

(ii) The objectives are to determine the number of rooms the council needs.

But we have no information about "down time" between occupants, or requirement for seasonal variation or any built-in surplus, and so on.

So with these data alone the requirements cannot be fulfilled.

Local mortality may be different from national experience.

Care home residents may experience significantly different mortality to the national population (the council may have data on deaths to care home residents which could be used to adjust the national mortality table).

The data are readily available and should be reliable.

However they may not be suitable for projecting more than a couple of years into the future because for example:

- the distribution of the local population may be skewed, e.g. there may be a huge number of 55 to 59 year olds compared to 60 to 65 year olds;
- age of entry into care homes is likely to change as medical advances help people stay healthier longer;
- the proportion of people going into council homes versus private homes may change as economic conditions change;
- the national mortality table may have no mortality improvements built in.
- social habits may change e.g. families may start living more as a unit so adult children/grandchildren may be available to care for the elderly at home for longer, especially if the ethnic mix of the city changes;
- the size/age mix of the city may change.

The average age at entry into a care home needs to be converted to a distribution by age. This may be subjective.

There might be different types of room for different levels of care needed. In this case facilities may be inadequate to meet needs even if there are sufficient rooms in total.

The data used for the model are taken from different sources so should not be unduly correlated.

We are not told of any models written in the past. If these existed, it would be useful to compare its past projections with what has happened in reality.

The resultant "number of rooms" occupied at any one time will need to be adjusted for example more elderly may decide to enter homes at the start of winter when it becomes harder/more expensive to stay warm at home, or when epidemics of influenza happen.

The model is relatively simple to explain.

The local mix of public and private care homes available locally may affect the proportion of elderly who go into council homes.

*Most candidates made a good attempt at part (i). Answers to part (ii) were more variable, and tended to focus more on the problems with the mortality data rather than issues connected with the supply and provision of care homes. In both parts of this question, not all the points listed were required for full marks.*

**5**

(i) $$P = \begin{pmatrix} 0.2 & 0.8 & 0 & 0 \\ 0.2 & 0 & 0.8 & 0 \\ 0.2 & 0 & 0 & 0.8 \\ 0 & 0.2 & 0 & 0.8 \end{pmatrix}$$

where the levels are ordered 0%, 25%, 50%, 60%.

(ii) (a) The chain is irreducible as it is clear that any state can eventually be reached from any other state.

(b) The process is aperiodic because, for example, the process can loop round in the 0% or 60% states giving no set return period to any state.

(iii) Stationary distribution $\pi$ satisfies $\pi = \pi P$

$$0.2\pi_0 + 0.2\pi_{25} + 0.2\pi_{50} = \pi_0 \quad (1)$$
$$0.8\pi_0 + 0.2\pi_{60} = \pi_{25} \quad (2)$$
$$0.8\pi_{25} = \pi_{50} \quad (3)$$
$$0.8\pi_{50} + 0.8\pi_{60} = \pi_{60} \quad (4)$$

Also $\pi_0 + \pi_{25} + \pi_{50} + \pi_{60} = 1 \quad (5)$

Working in terms of $\pi_{60}$

$$\pi_{50} = 0.25\pi_{60}$$

$$\pi_{25} = \frac{5}{16}\pi_{60}$$

$$\pi_0 = \frac{9}{64}\pi_{60}$$

Hence $\dfrac{(64+16+20+9)}{64}\pi_{60} = 1$

So the e stationary distribution is $\dfrac{1}{109}\begin{bmatrix} 9 \\ 20 \\ 16 \\ 64 \end{bmatrix}$

and the proportion of drivers at each level is

| | |
|---|---|
| 0% | 9/109 = 0.08257 |
| 25% | 20/109 = 0.18349 |
| 50% | 16/109 = 0.14679 |
| 60% | 64/109 = 0.58716. |

(iv) The 60% discount level becomes an absorbing state and so it is no longer irreducible.

However it is still aperiodic because you cannot get out of the absorbing state 60% and the other states still have no period.

The process would now be stationary when all drivers are in the absorbing 60% discount level.
OR
The new stationary distribution is [0,0,0,1] because the 60% state is now absorbing.

*This question was well answered, with many candidates scoring close to full marks. In part (iii) the correct numerical probabilities scored full marks, provided that it was clear to which level each probability applied. In part (iv) some candidates made vague statements that the probability of being in the 60% state would increase. While this is true, it was not given full credit, as the key point is that the stationary distribution has <u>everyone</u> in the 60% state.*

**6**

(i) Censoring is the mechanism which prevents us from knowing when an individual entered the investigation or the exact date of death.

(ii) Right Censoring. The trial is cut short after four weeks when some patients had still not recovered.
OR
The trial is cut short when some patients left the trial before their symptoms disappeared.

Type I Censoring. Censoring times are known in advance for all those patients still not recovered at the end of the trial.
Random Censoring. The time at which patients left the trial before their symptoms disappeared is a random variable.

Non-Informative Censoring. There is no reason to believe that those who left the trial had more or less chance of being cured by the cream than those who remained.

(iii) Rearranging the data:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Day | 0 | 2 | 6 | 7 | 10 | 10 | 13 | 14 |
| People in trial | 100 | 100 | 97 | 95 | 94 | 93 | 92 | 89 |
| No of exits | 0 | 3 | 2 | 1 | 1 | 1 | 3 | 2 |
| Reason for exit | | Left | cured | cured | cured | left | left | cured |

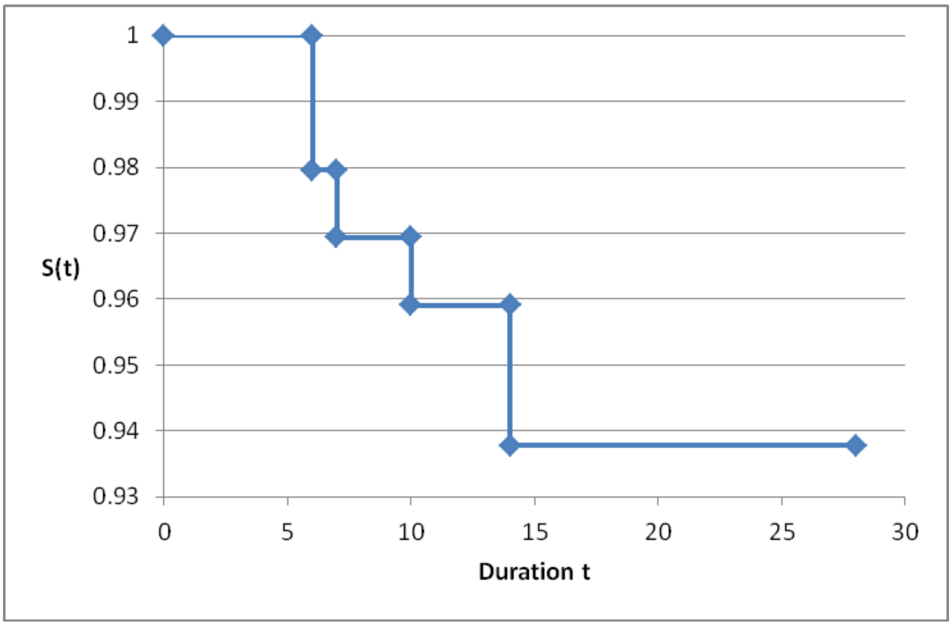
The Nelson-Aalen estimate is $\Lambda_t = \sum_{x_j \le x} \frac{d_j}{n_j}$

| $t_j$ | $n_j$ | $d_j$ | $c_j$ | $d_j/n_j$ | $\Lambda_t$ |
|---|---|---|---|---|---|
| 0 | 100 | 0 | 0 | | |
| 2 | 100 | 0 | 3 | | |
| 6 | 97 | 2 | 0 | 2/97 | .020619 |
| 7 | 95 | 1 | 0 | 1/95 | .031145 |
| 10 | 94 | 1 | 1 | 1/94 | .041783 |
| 13 | 92 | 0 | 3 | | |
| 14 | 89 | 2 | 0 | 2/89 | .064255 |

Since $S(t) = \exp\left(-\Lambda_t\right)$ we have

| $t$ | $S(t)$ |
|---|---|
| $0 \le t < 6$ | 1 |
| $6 \le t < 7$ | 0.97959 |
| $7 \le t < 10$ | 0.96934 |
| $10 \le t < 14$ | 0.95908 |
| $14 \le t < 28$ | 0.93777 |

(iv)




(v) The survival probability at $t = 14$ is 0.93777, so there is approximately a 94% chance of still having symptoms after two weeks.

*Many candidates answered this question well. In part (ii) Informative Censoring was acceptable if a sensible argument was made for it, for example those who left may be allergic to the cream and therefore less likely to be cured by it than those who remain. In part (ii) some candidates did not answer both parts of the question (that is, both describing the censoring and stating to whom it applied). In part (iii) and part (iv) it was expected that candidates would recognise that no information about what happened after 28 days could be gained from the data. In part (v) the answer given should be consistent with the S(t) estimated in part (iii).*

**7**

(i) The Gompertz model is simple to understand and to apply, having only two parameters.

It also fits human mortality at older ages well (e.g. 30–85 years).

(ii) $\log_e \mu_x = \alpha + \beta_0 x + \beta_1 U + \beta_2 I$

So $\mu_x = \exp(\alpha + \beta_0 x + \beta_1 U + \beta_2 I) = \exp(\beta_0 x)\exp(\alpha + \beta_1 U + \beta_2 I)$

This is equal to $Bc^x$ where $B = \exp(\alpha + \beta_1 U + \beta_2 I)$ and $c = \exp \beta_0$, hence Gompertz.

$\mu_x = \exp(\alpha + \beta_0 x + \beta_1 U + \beta_2 I) = \exp(\alpha + \beta_0 x)\exp(\beta_1 U + \beta_2 I)$

EITHER

Hence the force of mortality factorises into a term $\exp(\alpha+\beta_0 x)$ depending on age $x$ but not the covariates, and a term $\exp(\beta_1 U + \beta_2 I)$ depending on the covariates but not $x$, SO proportional hazards.

OR

Consider any two individuals, $i$ and $j$, with values of the covariates $U_i$ and $I_i$, and $U_j$ and $I_j$ respectively. Then the hazards for individuals $i$ and $j$ at age $x$ are

$$\mu_{x,i} = \exp(\alpha + \beta_0 x)\exp(\beta_1 U_i + \beta_2 I_i)$$

and

$$\mu_{x,j} = \exp(\alpha + \beta_0 x)\exp(\beta_1 U_j + \beta_2 I_j)$$

The ratio between the hazards is thus

$$\frac{\mu_{x,i}}{\mu_{x,j}} = \frac{\exp(\alpha + \beta_0 x)\exp(\beta_1 U_i + \beta_2 I_i)}{\exp(\alpha + \beta_0 x)\exp(\beta_1 U_j + \beta_2 I_j)} = \frac{\exp(\beta_1 U_i + \beta_2 I_i)}{\exp(\beta_1 U_j + \beta_2 I_j)},$$

which does not depend on $x$, hence proportional hazards.

(iii) $\log_e \mu_{40} = -9 + 0.09(40) + 0.3 - 0.0001(20{,}000) = -7.1$

so $\mu_{40} = 0.000825$.

(iv) $\mu_x = \exp(\alpha + \beta_0 x + \beta_1 U + \beta_2 I)$

Let the income of the urban resident be $I_U$ and that of the rural resident be $I_R$.

$$\begin{aligned}
&\exp(\alpha)\exp(\beta_0 x)\exp(\beta_1 + \beta_2 I_U) = \exp(\alpha)\exp(\beta_0 x)\exp(\beta_2 I_R)\\
&\exp(\beta_1 + \beta_2 I_U) = \exp(\beta_2 I_R)\\
&\exp(0.3 - 0.0001 I_U) = \exp(-0.0001 I_R)\\
&0.3 - 0.0001 I_U = -0.0001 I_R\\
&3{,}000 = I_U - I_R
\end{aligned}$$

So the difference is \$3,000.

Survival probability is

$$\exp\left\{-\left[\frac{e^{0.09s}}{0.09}\right]_{40}^{50} e^{-9}e^{0.3}e^{-0.0001(20,000)}\right\} = \exp\left[-\frac{0.00002254(90.017-36.598)}{0.09}\right]$$

$$= \exp(-0.01338) = 0.9867.$$

(vi) Since $S_x(t) = \exp\left[-\int_x^{x+t} \mu_s ds\right]$,

then if the rural resident is *a* years older than the urban resident we have

$$\exp\left[-\int_x^{x+t} e^{0.09s}e^{\alpha}e^{\beta_1}e^{\beta_2 I} ds\right] = \exp\left[-\int_{x+a}^{x+a+t} e^{0.09s}e^{\alpha}e^{\beta_2 I} ds\right]$$

Therefore

$$e^{\alpha}e^{\beta_2 I}\int_{x+a}^{x+a+t} e^{0.09s} ds = e^{\alpha}e^{\beta_2 I}\int_x^{x+t} e^{\beta_1}e^{0.09s} ds$$

$$\int_{x+a}^{x+a+t} e^{0.09s} ds = \int_x^{x+t} e^{\beta_1}e^{0.09s} ds$$

$$\left[\frac{e^{0.09s}}{0.09}\right]_{x+a}^{x+a+t} = \left[\frac{e^{\beta_1}e^{0.09s}}{0.09}\right]_x^{x+t}$$

$$e^{0.09(x+a+t)} - e^{0.09(x+a)} = e^{\beta_1}(e^{0.09(x+t)} - e^{0.09x})$$

$$e^{0.09x}e^{0.09a}(e^{0.09t} - 1) = e^{\beta_1}e^{0.09x}(e^{0.09t} - 1)$$

$$e^{0.09a} = e^{\beta_1}$$

$$a = \beta_1 / 0.09 = 0.3/0.09 = 3.33$$

So the rural dweller is aged 40 + 3.33 = 43.33 years.

*In part (i) very few candidates made the point that the Gompertz model is simple and convenient to use. Part (ii) was very poorly answered. When demonstrating that the model was a proportional hazards (PH) model, many candidates simply factorised the expression as* $\mu_{x,i} = \exp(\alpha)\exp(\beta_0 x + \beta_1 U_i + \beta_2 I_i)$ *and said that therefore exp(α) was the baseline hazard. This is incorrect because the second term includes both duration and the covariates. It was* *acceptable in part (ii) only to break up the equation once as long as the argument was developed further for both the Gompertz and the PH cases. A common error in part (v) was to assume that the hazard was constant at its value at age 40 years. This produced a survival probability of 0.99178. In part (vi) the derivation shown in above was not required for full credit. Candidates who spotted that, if 40+a is the age of the rural dweller in years, then* $e^{0.09a} = e^{\beta_1}$*, scored full credit.*

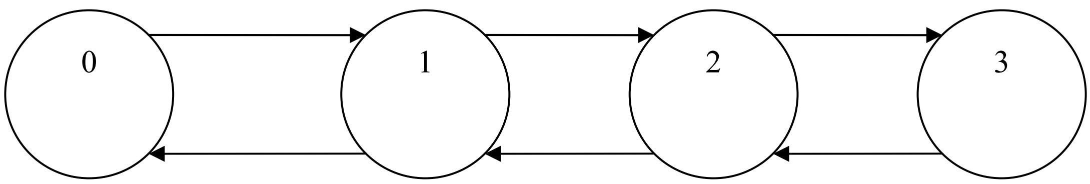
*Since the question was missing a comma after "were* $\alpha = -9.0$*"a small number of candidates interpreted the parameters differently i.e.* $\alpha = 0.09$*,* $\beta_0 = -0.01$*,* $\beta_1 = 0.3$ *and* $\beta_3 = -0.0001$*. This interpretation was given full credit, if followed through correctly.*

*In part (vi) the approach using* ${}_t p_x = \left[\exp\left(\frac{-B}{\log c}\right)\right]^{c^x(c^t-1)}$ *was acceptable.*

**8**

(i) The state space is {0,1,2,3} where the number indicates the number of available spaces.

(ii)




(iii) $$\begin{pmatrix} -3B & 3B & 0 & 0 \\ A & -A-2B & 2B & 0 \\ 0 & A & -A-B & B \\ 0 & 0 & A & -A \end{pmatrix}$$

where the order of the rows/columns is {0, 1, 2, 3}.

(iv) $\frac{d}{dt}P_{00}(t) = -3BP_{00}(t)$ (as probability of returning to state 0 not of interest)

OR

$$P_{00}(t) = \exp\left[-\int_0^t 3B\,dt\right]$$

$$P_{00}(t) = \exp(-3Bt)$$

$$P_{00}(2) = \exp(-6B)$$

(v) If a Markov jump process $X_t$ is examined only at the times of transition, the resulting process is called the jump chain associated with $X_{t.}$
OR
A jump chain is each distinct state visited in the order visited where the time set is the times when states are moved between.

(vi)
$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ A/_{A+2B} & 0 & 2B/_{A+2B} & 0 \\ 0 & A/_{A+B} & 0 & B/_{A+B} \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

where the order of the rows/columns is $\{0, 1, 2, 3\}$.

(vii) This is $\dfrac{A}{A+B}.\dfrac{A}{A+2B}$

(viii) Consider the paths by which the car park can become full before it becomes empty

Required probability = $P_{21}P_{10} + P_{21}P_{12}P_{21}P_{10} + P_{21}P_{12}P_{21}P_{12}P_{21}P_{10} + ......$

$$= \frac{A}{A+B}.\frac{A}{A+2B}\left[1+\frac{A}{A+B}.\frac{2B}{A+2B}+\frac{A}{A+B}.\frac{2B}{A+2B}\frac{A}{A+B}.\frac{2B}{A+2B}+....\right]$$

$$= \frac{A}{A+B}.\frac{A}{A+2B}\Big/\left[1-\frac{A}{A+B}.\frac{2B}{A+2B}\right]$$

OR

This can be done by defining a function of the probability full before empty from the current state, say $D_x$

Then $D_0 = 1$ and $D_3 = 0$

and $D_1 = D_0.\dfrac{A}{A+2B} + D_2.\dfrac{2B}{A+2B}$

$$D_2 = D_1.\frac{A}{A+B} + D_3.\frac{B}{A+B}$$

Solving these gives

$$D_2 = \frac{A}{A+B}.\frac{A}{A+2B}\Big/\left[1-\frac{A}{A+B}.\frac{2B}{A+2B}\right]$$

(ix) A time inhomogeneous model may be more appropriate.

Residents may come and go at particular times, for example if they drive to work.

They are unlikely to be moving their car as regularly in the middle of the night

Independent arrivals questionable because a family might have two cars arriving/leaving at the same time OR people might arrive and wait until a space becomes available thus leading to a queue

The Markov assumption may not be valid because neighbours may know from at experience when cars are moved and time their arrival accordingly.

The model assumes those parking cars are competent drivers, and do not park so as to take up 2 spaces.

*The problem can be worked in terms of the number of occupied spaces. This was not given full credit for part (i) as the question said "model the number of spaces free, but could gain full credit for the other parts. In part (ii) it was not necessary to mark the probabilities on the diagram. A common error was to omit the 2s and 3s in the matrix in part (iii). Part (v) was not as well answered as might have been expected, with many candidates writing vague descriptions which did not make it clear that they understood what a jump chain is.*

*Overall, this question was poorly answered by many candidates. A large proportion of candidates did not attempt parts (iii)–(viii).*

## 9

(i) Graduation by parametric formula.

Advantage: If a small number of parameters is used the resultant rates are automatically smooth;
OR sometimes when comparing several investigations it is useful to fit the same parametric formula to all of them;
OR the approach is well suited to the production of standard tables from large amounts of data.

Disadvantage: It can be difficult to find a suitable curve which fits the experience at all ages;
OR care is needed when extrapolating from ages where there is most data.

Graduation by reference to a standard table.

Advantage: Provided a simple function is chosen the resultant rates are automatically smooth;
OR it can be useful to fit relatively small data sets when a suitable standard table exists;

OR the standard table can be very good at deciding the shape of the graduation at extreme ages where data are sparse.

Disadvantage: It can be difficult to find a suitable standard table for the data;
OR it is not suitable for the preparation of standard tables.

Graphical graduation.

Advantage: It can be used for small data where no suitable standard table exists;
OR can allow for known features of the experience for example the accident hump.

Disadvantage: It is hard to achieve accuracy;
OR it takes a skilled practitioner;
OR it is very difficult to achieve adequate smoothness.

(ii) To test for overall goodness of fit we use the $\chi^2$ test.

The null hypothesis is that the graduated rates are the underlying rates of the experience.

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

| *Age* | *Exposed to risk* | *Observed deaths* | *Table Rates* | *Expected deaths* | $z_x$ | $z_x^2$ |
|---|---|---|---|---|---|---|
| 30 | 36,254 | 26 | 0.000590 | 21.38986 | 0.9968 | 0.9936 |
| 31 | 37,259 | 20 | 0.000602 | 22.42992 | −0.5131 | 0.2632 |
| 32 | 28,057 | 23 | 0.000617 | 17.31117 | 1.3673 | 1.8695 |
| 33 | 31,944 | 23 | 0.000636 | 20.31638 | 0.5954 | 0.3545 |
| 34 | 30,005 | 26 | 0.000660 | 19.80330 | 1.3925 | 1.9390 |
| 35 | 28,389 | 12 | 0.000689 | 19.56002 | −1.7094 | 2.9220 |
| 36 | 36,124 | 31 | 0.000724 | 26.15378 | 0.9476 | 0.8980 |
| 37 | 28,152 | 22 | 0.000765 | 21.53628 | 0.0999 | 0.0100 |
| 38 | 24,001 | 25 | 0.000813 | 19.51281 | 1.2422 | 1.5430 |
| 39 | 30,448 | 31 | 0.000870 | 26.48976 | 0.8763 | 0.7679 |
| | | | | Total | 5.2956 | 11.5608 |

The observed test statistic is 11.56

The number of age groups is 10, but we lose some degrees of freedom for the choice of the standard table and one degree of freedom for each parameter in the link function. So $m < 10$.

The critical value of the chi-squared distribution with 9 degrees of freedom at the 5% level is 16.92 (or with 8 d.f. is 15.51, or with 7 is 14.07).

Since 11.56 < 16.92 (or 15.51, or 14.07), we do not reject the null hypothesis at 95% level of significance.

(iii) Any two from:

**Individual Standardised Deviations Test**

Under the null hypothesis that the graduated rates are the true rates underlying the observed data

we should expect individual deviations to be distributed Normal (0,1).

Only 1 in 20 of the $z_x$s should lie above 1.96 in absolute value;
OR
none should lie above 3 in absolute value;
OR
table showing split of deviations, actual versus expected as below.

| Range | −∞, −2 | −2, −1 | −1, 0 | 0, 1 | 1, 2 | 2, +∞ |
|---|---|---|---|---|---|---|
| Expected | 0 | 1.4 | 3.4 | 3.4 | 1.4 | 0 |
| Actual | 0 | 1 | 1 | 5 | 3 | 0 |

The largest deviation we have here is −1.71,

which is within the range −1.96 to 1.96,

therefore we have no reason to reject the null hypothesis at the 95% level of significance.

**Signs Test**

Under the null hypothesis that the graduated rates are the true rates underlying the observed data

the number of positive signs amongst the $z_x$ is distributed Binomial (10, ½).

We observe 8 positive signs.

The probability of observing 8 or more positive signs in 10 observations is 0.0547 OR the probability of observing exactly 8 positive signs is 0.044.

This implies that Pr[observing 8 or more] > 0.025 (a two-tailed test),

so we have insufficient evidence to reject the null hypothesis at the 95% level.

**Cumulative Deviations Test**

Under the null hypothesis that the graduated rates are the true rates underlying the observed data, the test statistic

$$\frac{\sum_x (\text{Observed deaths} - \text{Expected deaths})}{\sqrt{\sum_x \text{Expected deaths}}} \sim \text{Normal}(0,1).$$

So, calculating as follows:

| *Age x* | *Observed deaths* | *Expected deaths* | *Observed minus expected deaths* |
|---|---|---|---|
| 30 | 26 | 21.38986 | 4.6101 |
| 31 | 20 | 22.42992 | –2.4299 |
| 32 | 23 | 17.31117 | 5.6888 |
| 33 | 23 | 20.31638 | 2.6836 |
| 34 | 26 | 19.80330 | 6.1967 |
| 35 | 12 | 19.56002 | –7.5600 |
| 36 | 31 | 26.15378 | 4.8462 |
| 37 | 22 | 21.53628 | 0.4637 |
| 38 | 25 | 19.51281 | 5.4872 |
| 39 | 31 | 26.48976 | 4.5102 |
| Totals | | 214.5033 | 24.4967 |

The value of the test statistic is $\frac{24.50}{\sqrt{214.50}} = 1.6726$.

Since – 1.96 < test statistic < +1.96 ,

we have insufficient evidence to reject the null hypothesis at the 95% level.

**Grouping of Signs Test**

Under the null hypothesis that the graduated rates are the true rates underlying the observed data

$G$ = Number of groups of positive deviations = 3
$m$ = number of deviations = 10
$n_1$ = number of positive deviations = 8
$n_2$ = number of negative deviations = 2

THEN EITHER

We want $k^*$ the largest $k$ such that $\displaystyle\sum_{t=1}^{k} \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}} < 0.05$

The test fails at the 5% level if $G \le k^*$.

From the Gold Book there is no entry for $k^*$.

So we have insufficient evidence to reject the null hypothesis at the 95% level.

OR

For $t = 3$

$$\binom{n_1-1}{t-1} = \binom{7}{2} = 21 \quad \text{and} \quad \binom{n_2+1}{t} = \binom{3}{3} = 1 \quad \text{and} \quad \binom{m}{n_1} = \binom{10}{8} = 45.$$

So Pr[$t = 3$] if the null hypothesis is true is 21/45 = 0.467, which is greater than 5% .

We have insufficient evidence reject the null hypothesis at the 95% level.

(iv) The chi-squared test suggests that the graduated rates adhere satisfactorily overall to the crude rates which gave rise to the observed deaths.

The Signs Test suggests that small but consistent bias is not a problem.

The shape of the graduated rates is not significantly different from the crude rates, as evidenced by the result of the Grouping of Signs Test.

The shape of the graduated rates is not significantly different from the crude rates, as evidenced by the result of the Cumulative Deviations Test.

There are no individual ages with suspiciously large deviations between the crude rates and the graduated rates.

Therefore it would seem reasonable for the company to use the graduated rates to price life insurance policies for this particular block of businesses.

*In part (iii) there were 3 marks available for each test. ANY two of the tests described above were allowed, even if they are testing for effectively the same thing, but the test for smoothness was not given credit as the graduation had been carried out with reference to a standard table. Credit was also given for the serial correlations test, and one or two candidates attempted this (none especially successfully). Candidates who carried out more* *than two tests, were credited with the marks for their two highest-scoring. In part (iv) many candidates made vague statements that the graduation "passed all the tests". This was given only limited credit. For full credit, details of the aspects of the graduation that each test focuses on were required. Also in part (iv) some candidates (despite finding no small bias in the data) decided that the presence of eight out of ten positive deviations merited further investigation, or that since observed deaths were generally higher than expected deaths, life products should be priced cautiously. Credit was given for these sensible comments.*

## END OF EXAMINERS' REPORT