Subject CT4 — Models.
Core Technical

September 2009 Examinations

EXAMINERS REPORT

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

December 2009

Comments for individual questions are given with the solutions that follow.

Examiners' Comments

Comments on solutions presented to individual questions for the September 2009 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.

1

A uniform distribution of deaths means

EITHER

that deaths are evenly spaced between the ages *x* and *y*.

OR

that ${}_tq_x = tq_x \quad (t \le y - x)$

OR

that ${}_tp_x\mu_{x+t}$ is constant for $t \le y - x$.

It also means that the survival function decreases linearly between ages *x* and *y*. The assumption of a constant force of mortality between any two ages means

EITHER

that the hazard does not change with age over this age range.

OR

that ${}_tp_x = (p_x)^t$.

This implies that the survival function decreases exponentially between *ages* x *and* y.

Answers to this straightforward bookwork question were disappointing. Although most candidates could describe the difference between a constant force of mortality and the increasing force implied by a uniform distribution of deaths, few made correct reference to the form of the survival function. An alarming number of candidates referred to survival functions which increased with age! Credit was given for graphs which correctly depicted the shape of the survival function under the two assumptions.

2

(i) Define objectives of modelling process.

Plan the modelling process and how it will be validated.

Collect and validate the data required.

Define the form of the model.

Involve experts on the real world system/get feedback on validity.

Decide on software to be used, choose random number generator etc.
Write the computer program.

Debug the program.

Analyse the output

Test the reasonableness of the output.

Consider appropriateness of response of the model to small changes in input parameters.

Communicate and document results.

[*½ mark was awarded for each point up to a maximum of 4 marks]*

(ii) Whilst in theory all steps are still required, some may take the form of reviewing the appropriateness of existing decisions made, such as how the form of the model was determined.

Extent of work will depend on whether the existing model is to be used, adapted or superseded.

An understanding of how results compare with those previously used by the company will be required.

Process maps for the existing approach, or discussions with the people running the process about what they do, may be helpful.

The scope needs to be tightly defined up front to ensure it is clear what is expected of the consultancy.

Data sources may already be established.

[½ mark was awarded for each point up to a maximum of 2 marks]

Part (i) of this question was basic bookwork and was extremely well answered. Part (ii) required more thought, but many candidates were able to write down some relevant points.

3

(i) For each life we need

EITHER date of birth OR exact age at entry into observation OR exact age at exit from observation

Date of entry into observation

Date of exit from observation

[Alternatives were given full credit, provided the information given allowed the calculation of the date of entry into and exit from observation and the life's age]

(ii) The contribution of each life to the central exposed to risk is the number of months between STARTDATE and ENDDATE, where STARTDATE is the latest of date of 40th birthday 1 January 2008 and ENDDATE is the earliest of date of 41st birthday date of death 31 December 2008

Life	*STARTDATE*	*ENDDATE*	*number of months between STARTDATE and ENDDATE*
1	1 January 2008	1 March 2008	2
2	1 January 2008	1 May 2008	4
3	1 January 2008	1 July 2008	6
4	1 January 2008	1 October 2008	9
5	1 January 2008	1 February 2008	1
6	1 February 2008	31 December 2008	11
7	1 April 2008	31 December 2008	9
8	1 June 2008	1 November 2008	5
9	1 August 2008	31 December 2008	5
10	1 December 2008	31 December 2008	1

Summing the number of months over the 10 lives gives a total of 53 months, which is 4.42 years, which is the central exposed to risk.

(iii)

a. The total number of deaths during the period of observation is 2. So the maximum likelihood estimate of the hazard of death is 2/4.42 = 0.4528.

b. ALTERNATIVE 1

If the hazard of death at age 40 years is μ_{40}, then

$$q_{40} = 1 - p_{40} = 1 - \exp(-\mu_{40})$$

$$= 1 - \exp(-0.4528) = 1 - 0.6358 = 0.3642.$$

ALTERNATIVE 2

If the central exposed to risk is E^c_{40}, then if we work in years

$$q_{40} \approx \frac{d_{40}}{E^c_{40} + 0.5d_{40}}$$

$$= \frac{2}{4.42+1} = \frac{2}{5.42} = 0.3690.$$

This was well answered. A common error was to count 3 deaths rather than 2. Although 3 deaths are mentioned in the data given in the question, one of these occurred after the life's 41st birthday and so should not be included in the estimation of μ_{40}. Another common error was to forget that exposure ends at exact age 41 years. Each of these errors was only penalised once, so that calculations which followed through correctly in (iii) were awarded full marks for part (iii). Note also that candidates who made BOTH the above errors were only penalised for one, as if exposure is assumed to continue past exact age 41 years, it is consistent to count 3 deaths!

4

(i) The principle of correspondence states that 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*. Problems in adhering to this can arise when the deaths data and the exposed-to-risk data come from two different sources. These may classify lives differently.

(ii) Since deaths are classified by age last birthday at date of death, a central exposed to risk which corresponds to the deaths data is given by

$$E_x^c = \int_{t=0}^{t=3} P_{x,t}$$

where $P_{x,t}$ is the population aged *x* last birthday at time *t*, and *t* is measured in years since 1 January 2005. We have censuses on 30 June 2004, 30 June 2005, 30 June 2007 and 30 June 2008.

Assuming that the population varies linearly across the period between each successive census for which we have data the population aged *x* last birthday on 1 January 2005 is equal to

$$\tfrac{1}{2}(P_{x,30/6/2004} + P_{x,30/6/2005})$$

and the population aged *x* last birthday on 1 January 2008 is equal to

$$\tfrac{1}{2}(P_{x,30/6/2007} + P_{x,30/6/2008}).$$

Dividing the period of the investigation into three sub-periods

from 1 January 2005 to 30 June 2005

from 30 June 2005 to 30 June 2007

from 30 June 2007 to 1 January 2008

and applying the trapezium rule to each sub-period produces the following exposed to risk for persons aged *x* last birthday

For the sub-period between 1 January 2005 and 30 June 2005

$$\tfrac{1}{2}\left[\tfrac{1}{2}(P_{x,1/1/2005} + P_{x,30/6/2005})\right]$$
$$= \tfrac{1}{2}\left[\tfrac{1}{2}(\tfrac{1}{2}(P_{x,30/6/2004} + P_{x,30/6/2005}) + P_{x,30/6/2005})\right]$$

For the sub-period between 30 June 2005 and 30 June 2007

$$2\left[\tfrac{1}{2}(P_{x,30/6/2005} + P_{x,30/6/2007})\right]$$

For the sub-period between 30 June 2007 and 1 January 2008

$$\tfrac{1}{2}\left[\tfrac{1}{2}(P_{x,30/6/2007} + P_{x,1/1/2008})\right]$$
$$= \tfrac{1}{2}\left[\tfrac{1}{2}(P_{x,30/6/2007} + \tfrac{1}{2}(P_{x,30/6/2007} + P_{x,30/6/2008}))\right]$$

Summing these gives

$$E_x^c = \tfrac{1}{8}P_{x,30/6/2004} + \tfrac{1}{8}P_{x,30/6/2005} + \tfrac{1}{4}P_{x,30/6/2005} + P_{x,30/6/2005}$$
$$+P_{x,30/6/2007} + \tfrac{1}{4}P_{x,30/6/2007} + \tfrac{1}{8}P_{x,30/6/2007} + \tfrac{1}{8}P_{x,30/6/2008}$$

which simplifies to

$$E_x^c = \tfrac{1}{8}P_{x,30/6/2004} + \tfrac{11}{8}P_{x,30/6/2005} + \tfrac{11}{8}P_{x,30/6/2007} + \tfrac{1}{8}P_{x,30/6/2008}\,.$$

The force of mortality may be estimated using the formula

$$\mu_x = \frac{d_x}{E_x^c},$$

where d_x denotes deaths to persons aged *x* last birthday when they died.

This was very poorly answered. It was perhaps rather more difficult than some exposed-to-risk questions in previous examination papers, but nevertheless the standard of most attempts was disappointing. In part (ii) credit was given for various alternative approximations provided that they were explained clearly.

5

(i) The Markov property states that the future development of a process can be predicted from its present state alone without reference to its past history.

(ii) Formally, for times $s_1 < s_2 < ... < s_n < s < t$ and for states $x_1, x_2, ..., x_n, x$ in the state space S and all subsets A of S, the Markov property can be written

$$\Pr[X(t) \in A \mid X(s_1) = x_1, X(s_2) = x_2, ..., X(s_n) = x_n, X(s) = x] = \Pr[X_t \in A \mid X(s) = x]$$

For independent increments we can write

$$\Pr[X(t) \in A \mid X(s_1) = x_1, X(s_2) = x_2, ..., X(s_n) = x_n, X(s) = x]$$
$$= \Pr[X(t) - X(s) + x \in A \mid X(s_1) = x_1, X(s_2) = x_2, ..., X(s_n) = x_n, X(s) = x]$$
$$= \Pr[X(t) - X(s) + x \in A \mid X(s) = x]$$
$$= \Pr[X(t) \in A \mid X(s) = x]$$

(iii)

a. A Markov chain is a stochastic process with the Markov property which has a discrete time set with a discrete state space. A Markov jump process is a stochastic process with the Markov property which has a continuous time set with a discrete state space.

b. A Markov chain is irreducible if any state can be reached from any other state.

(iv)

a. A lift could not serve its purpose unless it could return to each of the floors which it serves. This means an irreducible model would be appropriate.

b. Suppose, for example, the lift is currently at the third floor, with its last two states being the fourth floor and the fifth floor. In such a case the lift is more likely to be heading downwards than upwards. So the past history is likely to provide information on the likely future movement of the lift, unless the state space is very complicated (involving a number of past floors as well as the current floor). Therefore a Markov model is unlikely to be appropriate.

This question was generally well answered, apart from section (iv)(b) in which few candidates spotted the point that the direction of travel of the lift as well as its current floor will influence its next location.

6

(i) A Poisson process is a continuous-time integer valued process

$N_t,\ t \geq 0$ with

$N_0 = 0$

independent increments

EITHER

increments follow a Poisson distribution

OR

$$P[N_t - N_s = n] = \frac{[\lambda(t-s)]^n \exp[-\lambda(t-s)]}{n!}, \quad \text{for } s < t,\ n = 0, 1, 2, ...$$

(ii) Average work created by a complaint is

60%* ½+ 30%* 1 + 10%*4 = 1 day.

Complaints arrive at a rate 1.25 per working day

So, work expected to be generated is 1.25*1*5 = 6.25 person-days.

(iii) As the time to handle complaints follows an exponential (memoryless) distribution, only need to know how many unanswered complaints there are –

but do need to know how many of each type. If cases are allocated randomly rather than in order, then the state space consists of (in terms of complaints not resolved):

r – straightforward,

s – medium,

t – complicated.

where $r = 0,1,2,3,4,5,\ldots.$

$s = 0,1,2,3,4,5,\ldots\ldots$

$t = 0,1,2,3,4,5,\ldots..$

(iv) EITHER The model will only give an approximation.

OR The model is not suitable for this purpose.

The model could not be used to do this without extending the state space to consider the time the complaint has been in the queue. There are only two employees, so holidays and sickness are important factors not taken into account.

The model assumes complaints are time-homogeneous. We do not know the nature of the business, but for some industries complaints would be seasonal e.g. holiday companies.

The model assumes that complaint arrivals are independent, but more complaints might be expected if the company has had a quality control problem at a particular time. If struggling to meet the service standard, action would be. Taken, such as overtime, or prioritising easy cases. Staff may be able to deal with complaints which are similar to other recent complaints very quickly, using standard 'template' responses.

The memoryless property is unlikely to be realistic as the work required to complete the case could be assessed and then worked through to a schedule.

The Markov jump process could be used to estimate the probability that a complaint is responded to within a given number of days of receipt.

So the model could be used to estimate the probability of a complaint not being responded to in the stated time, that is the failure to meet the service standard.

[½ mark was awarded for each point up to a maximum of 3 marks]

Answers to this question were disappointing. Most candidates were able to tackle the calculation in part (ii) but few correctly identified the state space in part (iii), and most only made a cursory attempt at part (iv).

7

(i) Two step transition matrix

$$= \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.75 & 0 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} . \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.75 & 0 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.375 & 0.375 & 0.25 \\ 0.3125 & 0.625 & 0.0625 \\ 0.3125 & 0.375 & 0.3125 \end{pmatrix}$$

(ii) $\pi = \pi \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.75 & 0 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$

$\pi_1 = 0.5\pi_1 + 0.25\pi_2 + 0.25\pi_3$

$\pi_2 = 0.25\pi_1 + 0.75\pi_2 + 0.25\pi_3$

$\pi_3 = 0.25\pi_1 + 0.5\pi_3$

and $\pi_1 + \pi_2 + \pi_3 = 1$

$\pi_1 = 2\pi_3$

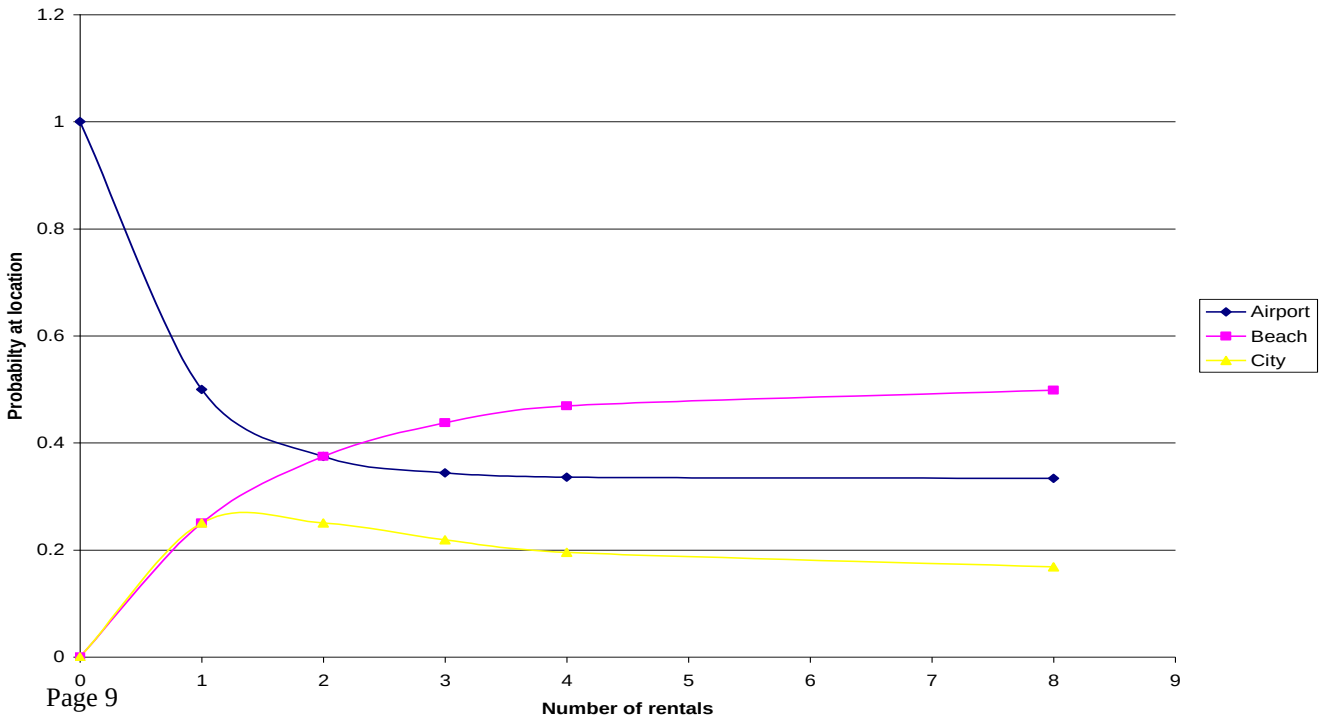
$\pi_2 = 3\pi_3$

$\pi_1 = 1/3$

$\pi_2 = 1/2$

$\pi_3 = 1/6$

(iii) The stationary distribution gives the long run probability that a particular car will be at each location. However this does not take into account the demand for hiring vehicles at each location, or the amount of space available at each location. These factors are likely to be more important in determining how many cars to base at each site.



(iv) A starts at 1, B and C at zero

Asymptote to the stationary distribution probs.

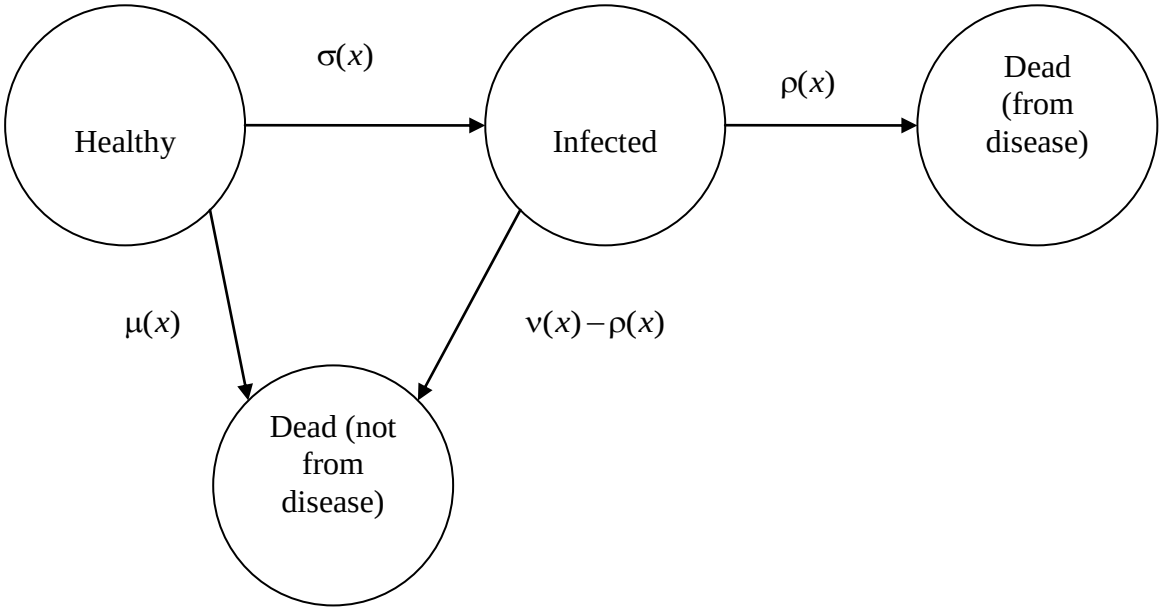
B and C same after 1 period

A and B same after 2 periods.

The calculations in parts (i) and (ii) were, as is usually the case in CT4 examinations, successfully completed by the vast majority of candidates. However only a minority made the point that, whereas the stationary distribution gives the long run probability that cars will be returned to each location, the company would be better advised to position cars at the three locations to reflect the demand for rentals. In part (iv), some candidates drew a set of histograms. Credit was given for this, provided that histograms were presented for 1 rental, 2 rentals, and the long run distribution, together with a statement that at 0 rentals the car must be at the Airport.

8

(i)



(ii) $\frac{d}{dt}P(x) = P(x)A(x)$ where with order of state space

{Healthy, Infected, Dead (not disease), Dead(from disease)}

$$A(x) = \begin{pmatrix} -\sigma(x)-\mu(x) & \sigma(x) & \mu(x) & 0 \\ 0 & -\nu(x) & \nu(x)-\rho(x) & \rho(x) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(iii)

a. $P_{HH}(x, x+t) = \exp[-\int_{w=0}^{t} (\sigma(x+w) + \mu(x+w))dw]$

b. $P_{HI}(x, x+t) = \int_{w=0}^{t} P_{HH}(x, x+w).\sigma(x+w).\exp[-\int_{u=w}^{t} \nu(x+u)du].dw$

c. EITHER

$P_{HD(\text{from disease})}(x, x+t) = \int_{w=0}^{t} P_{HI}(x, x+w).\rho(x+w).dw$

OR (backwards alternative)

$P_{HD(\text{from disease})}(x, x+t)$

$$= \int_{w=0}^{t} P_{HH}(x+w).\sigma(x+w).P_{ID(fromdisease)}(x+w, x+t).dw.$$

Now $P_{ID(fromdisease)}(x+w, x+t) = \int_{s=w}^{t} P_{II}(x+w, x+s).\rho(x+s).1.ds$

and $P_{II}(x+w, x+s) = \exp\left[-\int_{u=w}^{s} \nu(x+u)du\right]$.

So $P_{HD(\text{from disease})}(x, x+t)$

$$= \int_{w=0}^{t} P_{HH}(x+w).\sigma(x+w).\int_{s=w}^{t} \exp\left[-\int_{u=w}^{s} \nu(x+u)du\right].\rho(x+s).ds.dw$$

This question was considerably better answered than were similar questions in previous examinations. In particular, the proportion of candidates making serious attempts at part (iii) was greater than has been the case for similar questions in the past.

9

(i) Type II censoring as the study was terminated after a pre-determined number of failures. Random censoring of the device which exploded.

(ii) According to the information supplied by the sub-contractor, the Kaplan-Meier estimate of the survival function should be calculated as follows:

j	t_j	N_j	d_j	c_j	d_j/N_j	$1 - d_j/N_j$
0	0	12				
1	97	12	2	1	2/12	10/12

2	120	9	3	0	3/9	6/9
3	141	6	2	0	2/6	4/6
4	150	4	1	3	1/4	3/4

The Kaplan-Meier estimate is then

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{N_j} \right)$$

so we have

t	$\hat{S}(t)$
$0 \le t < 97$	1
$97 \le t < 120$	5/6
$120 \le t < 141$	5/9
$141 \le t < 150$	10/27
$150 \le t$	5/18 = 0.2778

(iii) Since 5/18 is not equal to 0.2727, the sub-contractor's story is internally inconsistent. The Kaplan-Meier estimate of the survival function after the failure of the 8th battery of 0.2727 would be obtained had only 11 batteries been tested at the start, and no battery being censored, as shown in the following table.

j	t_j	N_j	d_j	c_j	d_j/N_j	$1 - d_j/N_j$
0	0	11				
1	97	11	2	0	2/11	9/11
2	120	9	3	0	3/9	6/9
3	141	6	2	0	2/6	4/6
4	150	4	1	0	1/4	3/4
					+½	+½

The Kaplan-Meier estimate is then

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{N_j} \right)$$

so we have

t	$\hat{S}(t)$

$0 \le t < 97$	1
$97 \le t < 120$	9/11
$120 \le t < 141$	6/11
$141 \le t < 150$	4/11
$150 \le t$	3/11 = 0.2727

Therefore the value of $\hat{S}(150)$ reported by the sub-contractor is consistent with him having stolen the last battery.

Many candidates scored highly on this question. Credit was given in part (i) for other types of censoring provided that a sensible reason was given. In part (iii), for full credit some kind of calculation of an alternative survival function was needed, together with an explanation of why this provided evidence to support the suggestion that the sub-contractor has stolen the battery.

10

(i) The chi-squared test is for the overall fit of the graduated rates to the data

The test statistic is $\sum z_x^2$, where

$$z_x = \frac{\theta_x - E_x \overset{o}{q}_x}{\sqrt{E_x \overset{o}{q}_x (1 - \overset{o}{q}_x)}}.$$

The calculations are shown in the table below (since $\overset{o}{q}_x$ is small we use the approximation $z_x \approx \frac{\theta_x - E_x \overset{o}{q}_x}{\sqrt{E_x \overset{o}{q}_x}}$.

Age x	θ_x	$\overset{o}{q}_x$	$E_x \overset{o}{q}_x$	z_x	z_x^2
30	12	0.0091	8.645	1.141	1.302
31	14	0.0094	11.28	0.810	0.656
32	16	0.0097	11.64	1.278	1.633
33	9	0.0099	8.91	0.030	0.001
34	11	0.0106	10.60	0.123	0.015
35	15	0.0116	12.76	0.627	0.393
36	10	0.0127	10.16	-0.050	0.003
37	16	0.0138	17.25	-0.301	0.091

38	17	0.0149	20.86	-0.845	0.714
				$\sum$	4.808

The test statistic has a chi-squared distribution with degrees of freedom (d.f.) given by number of ages

– 1 (for parameter of function linking q_x^o and q_x^s)

– some d.f. for constraints imposed by choice of standard table

The critical value of the chi-squared distribution is

11.07 with 5 d.f.

12.59 with 6 d.f.

14.07 with 7 d.f.

15.51 with 8 d.f.

16.92 with 9 d.f. at the 5% level (from tables)

Since 4.808 < 11.07 (or 12.59 etc.) there is no evidence to reject the null hypothesis that the graduated rates are the true rates underlying the crude rates.

(ii) EITHER

Signs test

a. The Signs test looks for overall bias.

b. The number of positive signs among the z_x s

is distributed Binomial (9, 0.5).

We observe 6 positive signs.

The probability of obtaining 6 or more positive signs is

(from tables)

1 – 0.7461 = 0.2539.

[Alternatively, candidates could calculate the probability of obtaining exactly 6 positive signs, which is 0.1641]

Since this is greater than 0.025 (two-tailed test)

c. we cannot reject the null hypothesis and we conclude that the graduated rates are not systematically higher or lower than the crude rates.

OR

Cumulative Deviations test

a. When applied over the whole age range, the Cumulative Deviations test looks for overall bias

b. The test statistic is

$$\frac{\sum_x \left(\theta_x - E_x \overset{o}{q}_x \right)}{\sqrt{\sum_x E_x \overset{o}{q}_x}} \sim \text{Normal}(0,1)$$

Age x	θ_x	$E_x \overset{o}{q}_x$	$\theta_x - E_x \overset{o}{q}_x$
30	12	8.645	3.355
31	14	11.28	2.72
32	16	11.64	4.36
33	9	8.91	0.09
34	11	10.60	0.40
35	15	12.76	2.24
36	10	10.16	-0.16
37	16	17.25	-1.25
38	17	20.86	-3.86
	$\sum$	112.105	7.895

So the value of the test statistic is $\frac{7.895}{\sqrt{112.105}} = 0.7457$

Using a 5% level of significance, we see that

$-1.96 < 0.7457 < 1.96$

c. We accept the null hypothesis at the 5% level of significance and conclude there is no overall bias in the graduation.

Grouping of Signs test

a. The Grouping of Signs test looks for runs or clumps of deviations of the same sign OR the grouping of signs test tests for overgraduation.

b. We have:

9 ages in total

6 positive deviations

3 negative deviations

We have 1 positive run

Pr[1 positive run] is therefore equal to

$$\frac{\binom{5}{0}\binom{4}{1}}{\binom{9}{6}} = \frac{4}{\left(\frac{9.8.7}{3.2}\right)} = \frac{4}{84} = 0.0476$$

Since this is less than 0.05 (using a one-tailed

test)

c. We reject the null hypothesis that the graduated rates are the true rates underlying the crude rates (OR we conclude that the graduation is unsatisfactory OR there is evidence of over-graduation).

Individual Standardised Deviations test

a. The Individual Standardised Deviations tests looks for individual large deviations at particular ages.

b. If the graduated 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. Looking at the z_x s we see that the largest individual deviation is 1.278. Since this is less in absolute magnitude than 1.96

c. we cannot reject the null hypothesis that the graduated rates are the true rates underlying the crude rates.

Answers to this question were disappointing compared with previous years. A common error was for candidates to misread the question and to try to compare the observed number of deaths with an 'expected' number computed on the basis of the $\hat{q}_x$ *given in the question. These candidates were, in effect, examining deviations based solely on rounding! Candidates who made this error were penalised in part (i), but could gain credit for some of the alternative tests in part (ii) provided that they performed the tests correctly.*

11

(i) A proportional hazards (PH) model is a model which allows investigators to assess the impact of risk factors, or covariates, on the hazard of experiencing an event.

In a PH model the hazard is assumed to be the product of two terms, one which depends only on duration, and the other which depends only on the values of the covariates.

Under a PH model, the hazards of different lives with covariate vectors z_1 and z_2 are in the same proportion at all times:

for example in the Cox model

$$\frac{\lambda(t; z_1)}{\lambda(t; z_2)} = \frac{\exp(\beta z_1^T)}{\exp(\beta z_2^T)}.$$

(ii) 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 while an exponential term accounts for differences between individuals.

This means that 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.

(iii)

a. $\lambda(t) = \lambda_0(t)\exp(\beta_A A + \beta_E E + \beta_S S)$, where $\lambda(t)$ is the estimated hazard and $\lambda_0(t)$ is the baseline hazard.

b. A female aged exactly 16 years when she first claimed benefit who had not passed the school mathematics examination.

(iv) "The hazard of resuming work for males aged 17 years who had passed the mathematics examination was 1.5 times the hazard for males aged 16 years who had not passed the mathematics examination" implies that

$$\frac{\exp[(\beta_A * 1) + \beta_S + \beta_E]}{\exp(\beta_S)} = \exp(\beta_A + \beta_E)$$

$$= \exp(\beta_A)\exp(\beta_E) = 1.5$$

"Females who had passed the examination were twice as likely to take up a new job as were males of the same age who had failed" implies that

$$\frac{\exp(\beta_E)}{\exp(\beta_S)} = 2$$

since the age terms cancel out.

"Females aged 20 years who had passed the examination were twice as likely to resume work as were males aged 16 years who had also passed the examination" implies that

$$\frac{\exp(\beta_A * 4)}{\exp(\beta_S)} = 2.$$

Substituting from (2) into (1) gives

$$2\exp(\beta_A)\exp(\beta_S) = 1.5$$

so

$$\exp(\beta_S) = 0.75\exp(-\beta_A).$$

Substituting into (3) gives

$$\frac{\exp[\beta_A * 4)}{0.75\exp(-\beta_A)} = 2,$$

$$\exp(5\beta_A) = 1.5$$

$$\beta_A = \frac{\log_e 1.5}{5} = 0.0811$$

From (1) then, we obtain

$$\exp(\beta_E)\exp(0.0811) = 1.5$$

$$\beta_E + 0.0811 = 0.4055$$

$$\beta_E = 0.3244.$$

Finally, from (2) we obtain

$$\frac{\exp(0.3244)}{\exp(\beta_S)} = 2$$

$$0.3244 - \beta_S = \log_e 2 = 0.6931$$

$$\beta_S = -0.3688$$

This was satisfactorily answered by many candidates. Although it is still the case than only a minority of candidates seem to understand the essential feature of a proportional hazards model that the hazard can be factorised into one part depending on duration and another part depending on the values of covariates, many candidates could list some advantages of the Cox model in part (ii). In part (iii)(b) very few candidates spotted that the baseline person was aged 16 years <u>when first claiming benefit</u>. In part (iv) candidates who failed to write down the correct equations implied by the three statements in the question were given some credit for correctly solving the equations they did produce.

END OF EXAMINERS' REPORT