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 μ_x are equal when the force of mortality μ_{x+t} is constant for $0 \le t < 1$.

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

$$\exp\left(-\sum \frac{D_j}{N_j}\right) \text{ for } t_j < 9$$

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, λ_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 λ_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 μ_{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 ($i \neq j$):

μ_{AB}
μ_{AC}
μ_{BA}
μ_{BC}
μ_{CA}
μ_{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 λ_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:

$$\begin{pmatrix} -3 & 12/7 & 9/7 \\ 1/3 & -1/2 & 1/6 \\ 11/42 & 1/14 & -1/3 \end{pmatrix}$$

(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.

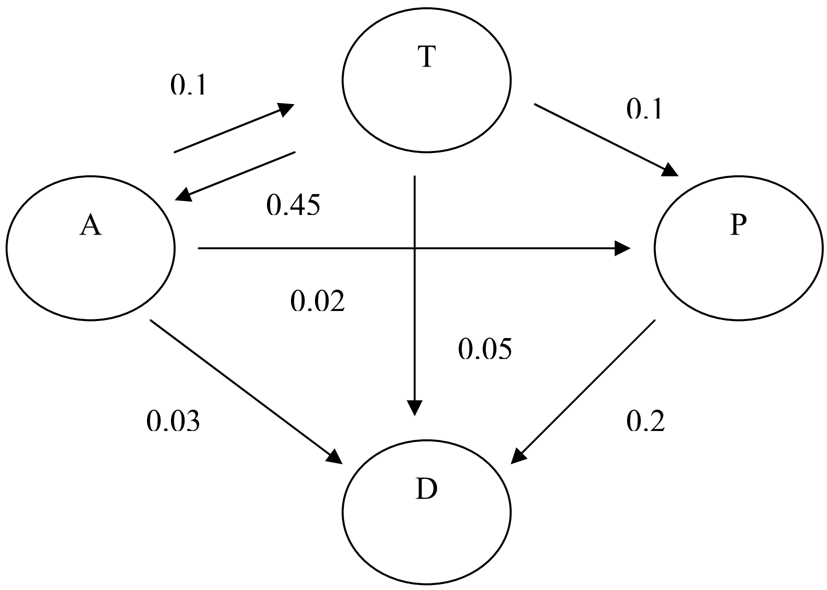
Regular exercise decreases the risk of heart disease for both males and females.

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.

$$\frac{d}{dt}(P_{\overline{AA}}(t)) = -0.15 P_{\overline{AA}}(t)$$
$$\frac{d}{dt}(\ln(P_{\overline{AA}}(t))) = -0.15$$
$$P_{\overline{AA}}(t) = \exp(-0.15t)$$

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,

the test statistic $\dfrac{\sum_x (\text{Observed deaths - Expected deaths})}{\sqrt{\sum_x \text{Expected deaths}}} \sim \text{Normal}(0,1)$

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 $t = 2$

$$\binom{n_1-1}{t-1} = \binom{5}{1} = 5 \quad \text{and} \quad \binom{n_2+1}{t} = \binom{6}{2} = 15$$

$$\text{and} \quad \binom{m}{n_1} = \binom{11}{6} = 462$$

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)
$(1-p)\pi_2 + p\pi_3 = \pi_2 \quad \text{or} \quad \pi_2 = \pi_3$ (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 $p.\pi_1 = \frac{p(1-p)}{3-p}$.

(v) Denoting R = raining, NR = not raining

From / To	*R*	*NR*
R	r	$1-r$
NR	s	$1-s$

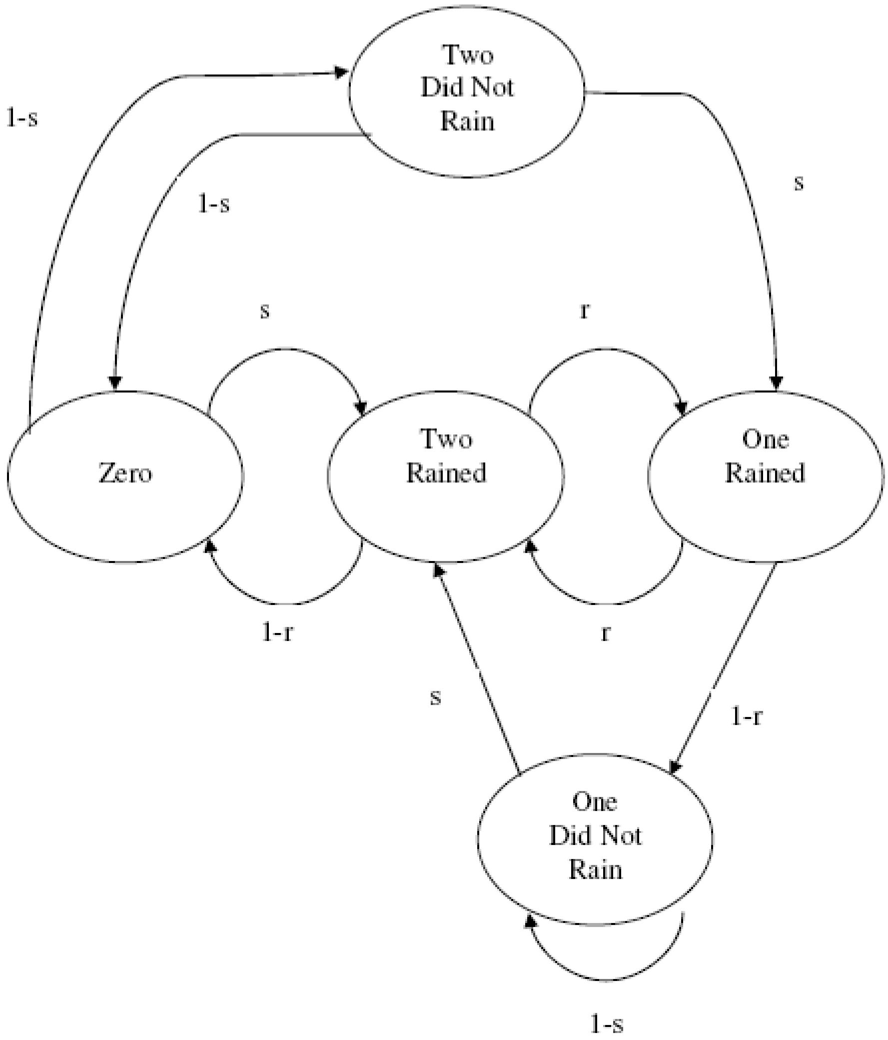
(vi) This would not satisfy the Markov property because (in states "One" and "Two") would need to know, in addition, whether it was raining or not on the last journey to determine the future evolution of the process.

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