INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2019 Examinations

Subject CS2A - Risk Modelling and Survival Analysis Core Principles

Introduction

The Examiners' Report is written by the Chief Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision

Mike Hammer
Chair of the Board of Examiners
July 2019

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

The aim of Risk Modelling and Survival Analysis (CS2) is to develop knowledge of and ability to apply statistical methods for risk modelling, time series analysis methods, stochastic processes (especially Markov chains and Markov jump processes), survival analysis (including regression methods applied to duration data) and graduation methods. It also includes a high-level introduction to machine learning. The exam paper aims at checking your understanding on both theory and application of the ideas to real data sets using R. We are not testing knowledge of the R program.

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

The overall performance is similar for both papers with average mark not far from 55 for both of them.

C. Pass Mark

The Pass Mark for this exam was 55

Q1

- All our models and analyses are based on the assumption that we can observe groups of identical lives (with respect to mortality characteristics). [½]
- Although in practice this is never completely possible, [½]
- …we can at least subdivide by characteristics which have a known impact on mortality, [½]
- …to reduce the heterogeneity of each class being investigated. [½]
- Although much heterogeneity will remain after subdividing by characteristics [½]
- Sub-division cannot be carried out unless the relevant information is collected, generally on the proposal form [½]
- A balance must be struck between obtaining more homogeneity and retaining large enough populations to make analysis possible. [½]

[3½, Max 2]

[Total 2]

Notes to Question 1:

(1) This is a relatively common question and the solution comes straight from the Core Reading.
(2) The context of the question is mortality investigations. The candidate's answer should be in this context to gain full marks
(3) Sensible examples of characteristics (e.g. sex, age etc.) are considered.

Well answered, many candidates gained marks on this question.

Q2

(i)

- Policy lasts for a fixed, and relatively short period of time [1]
- Insurance company receives a premium from the policyholder, … [1]
- … which covers claims made during the period of the contract. [1]
- At the end of the term the policy may be renewed [1]
  (at not necessarily the same premium)

[4, Max 2]

(ii)

- E.g. motor insurance, one-year term assurance. [2]

[Total 4]

Notes to Question 2:

(1) In part (ii) credit 1 mark per reasonable example
(2) In part (ii) candidates mentioning term assurance without reference to a short term receive no credit

Well answered, majority of candidates gained marks on this question.

Q3

(i)

- In supervised learning, the machine is given a specified output or aim. [1]
    - This might be the prediction of a specific numerical value
    - (e.g. a future lifetime) or the prediction of which category
    - …an individual will fall into (e.g. default on a loan or not). [½]
- In unsupervised learning the machine is set the task without a specific target to aim at. For example [1]
    - identifying clusters within a set of data without the number or nature of the clusters needing to be pre-specified). [½]

[3, Max 2]

(ii) <u>Supervised:</u>

Generalised linear models **OR** naïve Bayes classification **OR** decision trees
OR prediction of future lifetime **OR** neural networks
OR prediction of claims on certain classes of insurance **OR** defaulting of loan
OR regression models **OR** logistic regression **OR** probit models
OR discriminant analysis **OR** perceptron **OR** support vector machines [1]

<u>Unsupervised:</u>

Cluster analysis **OR** principal components analysis **OR** Apriori algorithm
OR market basket analysis **OR** text analysis **OR** neural networks [1]

[Max 2]

[Total 4]

Notes to Question 3:

(1) Part (i) asks for <u>the</u> difference. Examples in part (i) need to explain the difference to score. Note for part (i) the Core Reading states "The difference between these lies not (as one might think) in the level of involvement of the human researcher in the development of the algorithm, or in the supervision of the machine. Instead, it lies in the extent to which the machine is given an instruction as to the end-point (or target) of the analysis". Any answer which talks about human involvement in developing algorithms or supervising machines should receive no marks.

(2) In part (ii), 1 mark per reasonable example for each type is applied, practical examples are acceptable to score marks provided they are reasonable, neural networks can be either supervised or unsupervised.

Well answered, majority of candidates gained marks on this question.

Q4

(i)

- If $Y_1, \ldots, Y_j, \ldots$ is a sequence of independent and identically distributed variables then the process [1]
- $X_n = \sum_{j=1}^{n} Y_j$ [½]
- with initial condition $X_0 = 0$ [½]
- is a general random walk. [½]

[2½, Max 2]

(ii)

- Since $Y_1, \ldots, Y_j, \ldots$ is a sequence of identically distributed variables, they each have the same mean *m*. [1]
- Therefore, the mean of $X_n = \sum_{j=1}^{n} Y_j$ is *mn*. This is proportional to *n*. [1]
- Thus the mean of the process X_n is not constant, [½]
- which violates the condition for (weak) stationarity. [½]
- Hence a general random walk is not stationary. [½]

[3½, Max 3]

[Total 5]

Notes to Question 4:

(1) In part (ii) if candidates show that the variance is proportional to n using the steps similar to that outlined above for the mean then full marks was awarded.

A mixture of answers here with many scoring full marks.

Q5

(i)

- (A death rate is calculated as deaths divided by exposed to risk)
- A life alive aged *x* at time *t* should be included in the exposed to risk aged *x* at time *t* if and only if, were that life to die immediately, its death would be included in the numerator of the rate and classified as aged *x*. [1]

(ii)

- We adjust the age definition of the exposed to risk, as the deaths data carry more information. [½]
- Let $P_{x,t}$ be the population aged *x* nearest birthday at time *t*, and $P_{x,t}$* be the population aged *x* last birthday at time *t*. [½]
- Then $P_{x,t} = \frac{1}{2} P_{x-1,t}{}^{*} + \frac{1}{2} P_{x,t}{}^{*}$ [½]
- Assuming the population varies linearly over time, (we can use the trapezium rule to approximate the required exposed to risk: E_x^c): [½]

- $E_x^c = \left(\frac{1}{2}P_{x,1/1/2016} + \frac{1}{2}P_{x,1/1/2017}\right) + \left(\frac{1}{2}P_{x,1/1/2017} + \frac{1}{2}P_{x,1/1/2018}\right)$ [½]

Hence

- $E_x^c = \frac{1}{2}P_{x,1/1/2016} + P_{x,1/1/2017} + \frac{1}{2}P_{x,1/1/2018}$, [½]

  and, substituting, we have

- $E_x^c = \frac{1}{4}(P_{x-1,1/1/2016}* + P_{x,1/1/2016}*) + \frac{1}{2}(P_{x-1,1/1/2017}* + P_{x,1/1/2017}*) + \frac{1}{4}(P_{x-1,1/1/2018}* + P_{x,1/1/2018}*)$ [1]

(iii)

- If we can assume that the population at age nearest birthday varies linearly across the calendar year, the formula derived in part (ii) will produce the same exposed to risk. [1]

- This is because, whereas, on 1 January in each calendar year, there would be twice as many people in the exposed to risk aged between *x* and *x*+½ as there are aged between *x*-½ and *x*, and [½]

- in the middle of each calendar year there would be only half as many people in the exposed to risk aged between *x* and *x*+½ as there would be aged between *x*-½ and *x*. [½]

- Averaged over each calendar year, therefore, there will be an equal amount of exposure aged *x*-½ to *x* as there will be aged *x* to *x*+½. [1]

- But the uneven distribution of births across the calendar year means that the circumstances in which the assumption that the population at age nearest birthday varies linearly across the calendar year is valid are more restricted than is the case when the births are evenly distributed across the calendar year. [1]

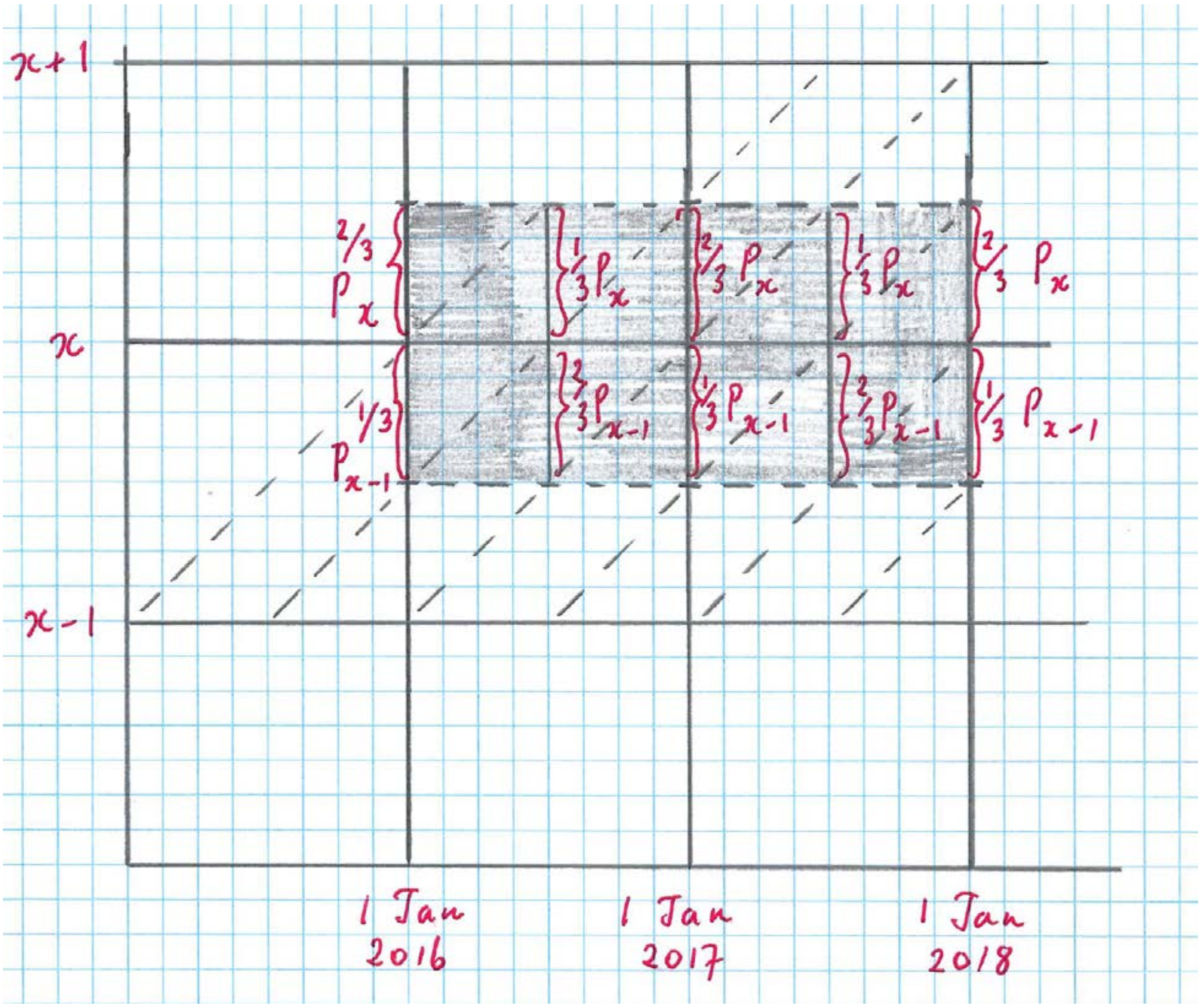
- Provided the number of births is roughly constant from year to year, the formula derived in part (ii) is still reasonable. [½]

[4½, Max 3]

[Total 8]

Explanation of the answer to CS2A April 2019 Question 5(iii)

The deaths occur in the shaded area in the Lexis diagram below. So that is the exposed to risk we need to approximate.



Let $P_{x,\tau}$ be the population aged x nearest birthday at time τ, and $P_{x,\tau}*$ be the population aged x last birthday at time τ. Assuming the trapezium rule holds (the population varies linearly between census dates), then we might approximate the required exposed to risk as follows:

$$E_x^c = \frac{1}{2}P_{x,1/1/2016} + P_{x,1/1/2017} + \frac{1}{2}P_{x,1/1/2018} \qquad (1)$$

If only one third of births take place in the first half of the calendar year, then two thirds take place in the second half of the calendar year.

Then, on 1 January in year t, the number of lives aged between x and $x+½$ will be equal to $\tfrac{2}{3}P_{x,1/1/t}*$, because these lives were born in the second half of the calendar year; and the number of lives aged between x-½ and x will be equal to ⅓ $P_{x\text{-}1,1/1/t}*$, because these lives were born in the first half of the calendar year.

Substituting into equation (1) we have:

$$E_x^c = \frac{1}{2}\left(\frac{2}{3}P_{x,1/1/2016}*+\frac{1}{3}P_{x-1,1/1/2016}*\right)+\left(\frac{2}{3}P_{x,1/1/2017}*+\frac{1}{3}P_{x-1,1/1/2017}*\right)+\frac{1}{2}\left(\frac{2}{3}P_{x,1/1/2018}*+\frac{1}{3}P_{x-1,1/1/2018}*\right) \tag{2}$$

which is different from the formula in the solution to Q5(ii).

Formula (2) assumes that the population varies linearly across the calendar year. An issue with this is that there are circumstances in which the uneven distribution of births across the calendar year renders this assumption less tenable than it was when we assumed that births were evenly distributed. To generalise about these circumstances is not so easy but we can say that:

(i) when the number of births is constant from year to year, or when the number of births oscillates regularly from year to year, the assumption that the population varies linearly over the calendar year will still hold;

(ii) when the number of births increases or decreases linearly over time, the assumption will no longer hold, even though it holds when births are evenly distributed across the calendar year;

(iii)when the number of births changes irregularly from year to year, the assumption that the population varies linearly over time will not hold, but in such a case it would often not hold (for age nearest birthday) even when births are distributed evenly across each calendar year.

If we can assume that the population varies linearly over the calendar year, then, averaged over each calendar year, therefore, the number of lives aged between x and x+½ is given by

$$\frac{1}{2}.\frac{1}{2}\left(\frac{2}{3}P_{x,1/1/t}*+\frac{1}{3}P_{x,1/7/t}*\right)+\frac{1}{2}.\frac{1}{2}\left(\frac{1}{3}P_{x,1/7/t}*+\frac{2}{3}P_{x,1/1/t+1}*\right). \tag{3}$$

But by the trapezium rule for the whole year of age:

$$P_{x,1/7/t}*=\frac{1}{2}P_{x,1/1/t}*+\frac{1}{2}P_{x,1/1/t+1}* \ ,$$

so, substituting in (2), the number of lives aged between x and x+½ is given by

$$\frac{1}{2}.\frac{1}{2}\left(\frac{2}{3}P_{x,1/1/t}*+\frac{1}{3}\left(\frac{1}{2}P_{x,1/1/t}*+\frac{1}{2}P_{x,1/1/t+1}*\right)\right)+\frac{1}{2}.\frac{1}{2}\left(\frac{1}{3}\left(\frac{1}{2}P_{x,1/1/t}*+\frac{1}{2}P_{x,1/1/t+1}*\right)+\frac{2}{3}P_{x,1/1/t+1}*\right)$$

$$=\frac{1}{4}P_{x,1/1/t}*+\frac{1}{4}P_{x,1/1/t+1}*.$$

The same reasoning may be applied to the number of lives aged between x-½ and x.

Applied to the two calendar years 2016 and 2017, this leads to the same formula as in the Marking Schedule for Q5(ii):

$$E_x^c = \frac{1}{4}\left(P_{x\text{-}1,1/1/2016}* + P_{x,1/1/2016}*\right) + \frac{1}{2}\left(P_{x\text{-}1,1/1/2017}* + P_{x,1/1/2017}*\right) + \frac{1}{4}\left(P_{x\text{-}1,1/1/2018}* + P_{x,1/1/2018}*\right).$$

We have given credit for any sensible observations or comments for this.

Notes to Question 5:

(1) In part (i), "if" is acceptable instead of "if and only if" to score full marks.

(2) Part (ii) is a derivation and so all assumptions and all terms must be defined in order to score full marks. In part (ii) the first point is not needed explicitly but can be awarded if the candidate proceeds to adjust the exposed to risk rather than the deaths data

We would recommend that credit is given for any sensible observations or comments.

Part (iii) of this question was very difficult with very few students scoring well.

Q6

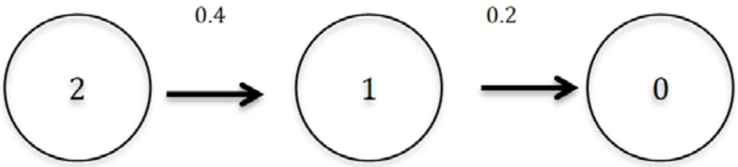
(i)

- The transition rates of moving from each state to each other state must be non-negative
- **OR** The transition rates OFF the leading diagonal must be non-negative
  OR μ_{ij} >= 0 for i ≠ j [½]
- The transition rates ON the leading diagonal must be non-positive [½]
- **OR** μ_{ii} <= 0
- The sum of each row must be zero.
  OR $\mu_{ii} = -\sum_{j \neq i} \mu_{ij}$ [½]
- Should be a finite or countable square matrix. [½]

(ii) State space {2,1,0} policies in force.

[1]

(iii)



[2]

(iv)

- $\frac{d}{dt}P_{22}(t) = -0.4P_{22}(t)$ [½]
- $\frac{d}{dt}P_{21}(t) = 0.4P_{22}(t) - 0.2P_{21}(t)$ [1]
- $\frac{d}{dt}P_{20}(t) = 0.2P_{21}(t)$ [½]

(v)

- The ongoing cost after takeover is negligible and can therefore be ignored. We therefore need to establish the point at which the cost of the insurance company maintaining the administrative system has a probability of 0.5 of exceeding the fee for migrating to the outsourcer. [½]
- $\frac{d}{dt}P_{11}(t) = -0.2P_{11}(t)$ [½]
- $P_{11}(t) = \exp(-0.2t)$ [1]

THEN EITHER:

- Need to find T such that $P_{11}(T) > 0.5$
- $\exp(-0.2T) > 0.5$
- $-0.2T > \ln 0.5$
  $T > 3.466$ [1]
- On this basis, insurance company could pay up to around 3.466*25000 which is around £86,000. [1]

OR:

- $P_{11}(T)$ follows an exponential survival model as the hazard is constant
  Expected future lifetime = 1 / 0.2 = 5 years [1]
- On this basis, insurance company could pay up to around 5 * 25000 which is £125,000. [1]

[4, Max 2]

[Total 11]

Notes to Question 6:

(1) In part (i), stating that the matrix is n x n is sufficient to score the final ½ mark.

(2) Credit for part (ii) awarded if part (iii) is correct, provided that the candidate has not given an incorrect answer for part (ii)

(3) In part (iii), full marks are awarded for a complete diagram (arrows, states and transition rates). A ½ mark is deducted for each error or omission. One mark is deducted if transition rates are parameterised. However no further marks is deducted in part (iv) if the parameterisation continues.

(4) In part (iv), one mark is deducted if the transition rates are parameterised, provided that the rates were NOT parameterised in part (iii)

(5) In part (v), full marks can be awarded for other valid solutions e.g. X% confidence interval rather than the median or mean, provided suitable justification is provided for the choice of X.

Parts (i-iii) were straightforward but answers were not as the examining would have expected. Part (iv) was more challenging and few candidates gained marks here.

Q7

(i)

- Let F be a joint distribution function with marginal cumulative distribution functions $F_1,\ldots, F_d$. [½]
- Then there exists a copula C [½]
- such that for all $x_1,\ldots, x_d \in [-\infty, \infty]$: [½]
- $F(x_1,\ldots.., x_d) = C[F_1(x_1),\ldots, F_d(x_d)]$ [½]

(ii)

- This gives the probability that RV1 is in the bottom u_1 percentile, and RV2 is in the bottom u_2 percentile, and RV3 is in the bottom u_3 percentile

[2]

(iii)

- The Gumbel generating function is defined:
  $\Psi(t) = (-\ln t)^\alpha$ [½]
- $\Psi(0) = \lim_{t\to 0}(-\ln t)^\alpha = \infty$, so the pseudo-inverse equals the normal inverse [½]
- Now invert the relationship $x = (-\ln y)^\alpha$ to find the normal inverse function:
  $y = \exp(-x^{1/\alpha})$ [½]
- So, $C(u, v, w) = \Psi^{[-1]} (\Psi(u) + \Psi(v) + \Psi(w))$ [1]
- $= \exp(- ((-\ln u)^\alpha + (-\ln v)^\alpha + (-\ln w)^\alpha)^{1/\alpha})$ [1]

[3½, Max 3]

(iv)

- exp(-((-ln 0.05)^4 + (-ln 0.075)^4 + (-ln 0.1)^4)^0.25) = 2.957% [3]

(v)

- Independent OR Independence OR Product [1]

(vi)

THEN EITHER:

- 0.05 * 0.075 * 0.1 = 0.0375% [1]

OR:

- exp (-((-ln 0.05)^1 + (-ln 0.075)^1 + (-ln 0.1)^1)^1)

= exp (ln 0.05+ ln 0.075 + ln 0.1)
= 0.05 * 0.075 * 0.1
= 0.0375% [1]

[2, Max 1]

[Total 12]

Notes to Question 7:

(1) In part (ii), the candidate has been asked to explain in words. Explanations in formulae or symbols receive no credit.

(2) In parts (iv) and (vi) the candidate has been asked to "calculate…". Full marks should be awarded for correct answers.

Many candidates struggled here more in terms of clearly expressing the key ideas, especially in part (ii).

Q8

(i)

- It ensures that the hazard is always positive. [1]
- You can ignore the shape of the baseline hazard and estimate the effects of the covariates directly from the data. [1]
- The logarithm of the hazard is linearly related to the covariates. [1]
- It is readily available in lots of software packages. [1]
- Easy to use / easy to asses impact of risk factors by assessing coefficients of covariates [1]
- Widely used and understood / easy to communicate [1]
- Hazards of different lives are in same proportion at all times [1]

[7, Max 2]

(ii)

From the second condition:

- $\exp(15\beta_A) * \exp(\beta_D) = 1.5 * \exp(15\beta_A)$
- $\beta_D = \ln(1.5)$
- $\beta_D = 0.405465$ [1½]

From the third condition:

- $\exp(18\beta_A) * \exp(\beta_D) = 2 * \exp(3\beta_A)$
- $\exp(15\beta_A) = 2 / 1.5$
- $\beta_A = 0.019179$ [1½]

From the first condition:

- $\exp(13\beta_A) * \exp(\beta_E) = 0.5 * \exp(8\beta_A) * \exp(\beta_D)$
- $\exp(\beta_E) = 0.5 * \exp(\beta_D - 5\beta_A)$
- $\beta_E = -0.383576$ [2]

(iii)

- A positive parameter increases the hazard rate and a negative one reduces it. [½]
- The larger the magnitude the greater the effect. [½]
  So the hazard increases as age increases (by about 1.9% per year of age).
- Dieting increases the hazard (by 50%).
- Doing exercises reduces the hazard (by about 32%). [2]

[3, Max. 2]

(iv)

- FlexPexApps should do a likelihood ratio test. [½]
- They should calculate the loglikelihood of the model as it stands, L_{current}. [½]

For each additional factor individually, they should

  - extend the model to include the new parameter,
  - estimate the parameter appropriate to that factor,
  - calculate the loglikelihood of the new model L_{new}
    [1½]

- Under the null hypothesis that the new parameter has no impact, i.e. the parameter associated with the new parameter = 0, [½]
- -2(L_{current} - L_{new}) has a chi-squared distribution with 1 degree of freedom. [½]
- So if -2(L_{current} - L_{new}) > 3.841 we can reject the null hypothesis [½]
- and conclude that the additional parameter has an impact. [½]

[4½, Max. 3]

[Total 12]

Notes to Question 8:

(1) In part (ii) the candidate has been asked to "calculate…".

(2) In part (iii), full marks are awarded for just for commenting on the directional impacts of each of the three factors. Quantifying the impacts is not required. If only one directional impact is correct 1 mark is awarded with the remaining two impacts receiving a ½ mark each.

Mixed performance but high average score.

Q9

(i)

- The lag polynomial is $1 - 0.9L + 0.14L^2 = (1 - 0.7L)(1 - 0.2L)$. [1½]
- Since the roots $\frac{1}{0.7}$ and $\frac{1}{0.2}$ are both greater than one in absolute value the process Y_t is stationary. [1½]

[Total 3]

(ii)

- The model is ARIMA(2,0,0). [1]

(iii)

- Since the process is stationary we know that $E(Y_t) = \mu$ where μ is some constant not depending on t. [½]
- Taking expectations on both sides of the model equation Y_t gives:

- $E(Y_t) = 0.5 + 0.9E(Y_t - 1) - 0.14E(Y_t - 2) + 0$ [1]
- and so $\mu = 0.5 + 0.9\mu - 0.14\mu$ [1]
- $\mu = \frac{0.5}{1-0.9+0.14} = \frac{0.5}{0.24} = 2.08\dot{3}$ [½]

[Total 3]

(iii)

- The auto-covariance function is not affected by the constant term of 0.5 in the equation, and this term can be ignored.
- The Yule-Walker equations are
  $\gamma_0 = 0.9\gamma_1 - 0.14\gamma_2 + \sigma^2$ (A) [½]
- $\gamma_1 = 0.9\gamma_0 - 0.14\gamma_1$ (B) [½]
- $\gamma_2 = 0.9\gamma_1 - 0.14\gamma_0$ (C) [½]
- Dividing both sides of (B) by γ_0 and noting that $\rho_S = \frac{\gamma_s}{\gamma_0}$
- $\rho_1 = 0.9 - 0.14\rho_1$ so that $\rho_1 = \frac{0.9}{1+0.14} = 0.7894737$ [1]
- Similarly, by dividing both sides of (C) by γ_0 and substituting $\rho_1$we have
- $\rho_2 = 0.9\rho_1 - 0.14 = 0.9 * 0.7894737 - 0.14 = 0.5705263$ [1]
- The value of $Var(Y_t) = \gamma_0$ can now be easily obtained by noticing that equation (A)
- $\gamma_0 = 0.9\rho_1\gamma_0 - 0.14\rho_2\gamma_0 + \sigma^2$ [1]
- And so $\gamma_0 = \frac{\sigma^2}{1-0.9\rho_1+0.14\rho_2} = \frac{\sigma^2}{1-0.9*0.7894737+0.14*0.5705263} = 2.707478\,\sigma^2$

[1½]

[Total 13]

Notes to Question 9:

(1) In parts (iii) and (iv) the candidate has been asked to "calculate…". Full marks are awarded for the correct answers.

Well answered, many candidates gained marks on this question.

Q10

(i)(a)

- If TP are the true positives, and FP are the false positives, then
- Precision = $\frac{\text{TP}}{\text{TP+FP}}$ [½]
- For the clinical procedure this is 45 / (45 + 15) = 0.75. [½]
- For the questionnaire this is 40 / (40 + 10) = 0.80. [½]

(b)

- If FN are the false negatives, then
- Recall = $\frac{\text{TP}}{\text{TP+FN}}$ [½]
- For the clinical procedure this is 45 / (45 + 5) = 0.90. [½]
- For the questionnaire this is 40 / (40 + 10) = 0.80. [½]

(c)

- The F1 score is equal to
- $$\frac{2 \text{ x precision x recall}}{\text{precision + recall}}$$ [1]
- For the clinical procedure this is 2(0.75)(0.90)/(0.75 + 0.90) = 0.818 [½]
- For the questionnaire this is 2(0.80)(0.80)/(0.80 + 0.80) = 0.800 [½]

[Total 5]

(ii)

- The F1 score may be written
- F1 score = $$\frac{2\left[\frac{\text{TP}}{\text{TP + FP}}\right]\left[\frac{\text{TP}}{\text{TP + FN}}\right]}{\left[\frac{\text{TP}}{\text{TP + FP}}\right]+\left[\frac{\text{TP}}{\text{TP + FN}}\right]} = \frac{2\left[\frac{\text{TP}}{\text{TP + FP}}\right]\left[\frac{\text{TP}}{\text{TP + FN}}\right]}{\frac{\text{TP(TP+FN)+TP(TP+FP)}}{\text{(TP + FP)(TP + FN)}}}$$
- $$= \frac{2(\text{TP})^2}{\text{TP(TP+FN)+TP(TP+FP)}} = \frac{2\text{TP}}{2\text{TP + FN + FP}} = \frac{\text{TP}}{\text{TP + 0.5FN + 0.5FP}}$$ [3]

(expresses the true positives as a proportion of the true positives plus the average of those incorrectly classified.)

(iii)

- Compared with the questionnaire, the clinical procedure is better at identifying the true positives, [½]
- but not so precise, as it classifies as positive a higher proportion of those who do not have the disease. [½]
- Whether recall or precision are chosen as measures will depend on whether it is most important to identify all the persons who have the disease, or not to unduly worry and treat people who are disease-free. [1]
- As the disease is serious it would perhaps be best to maximise the true positives and minimise the false negatives and so the clinical procedure would be preferred. [1]
- In real life, a very large proportion of those tested will not have the disease, so testing equal numbers of patients who do and do not have the disease may not be so useful. [1]
- The F1 score, however, is reasonably robust to the situation where most people do not have the disease, as its calculation does not involve the true negatives. [1]
- As the sample size is relatively small, the test should be re-performed on a larger population before drawing any firm conclusions. [1]
- The questionnaire is likely to be easier/cheaper to administer and therefore may be a good short-term substitute until the clinical procedure can be established in areas that currently have no screening in place [1]

[7, Max 4]

[Total 12]

Notes to Question 10:

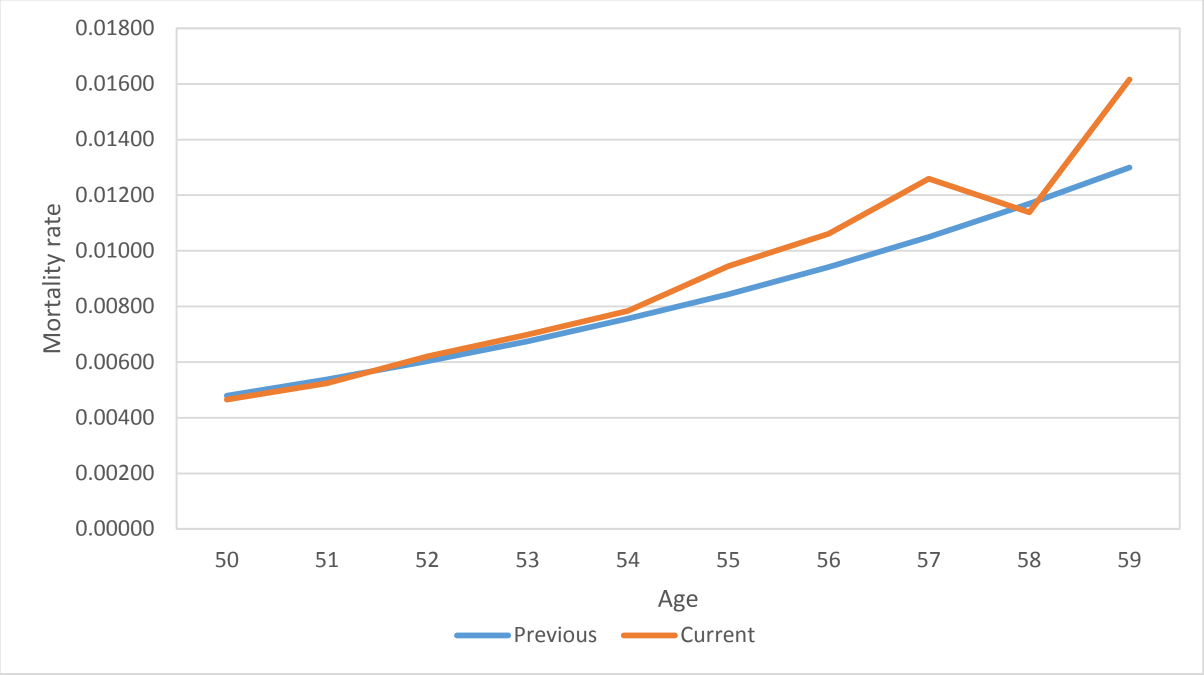
1. *In part (i) the candidate has been asked to "calculate…". Full marks are awarded for the correct answers.*
2. *In part (ii) full credit is awarded if candidates demonstrate it numerically*

This straightforward question, well answered by many candidates.

Q11

(i)

Age	*ETR*	*Actual Deaths*	*μ new*	*μ old*
50	5,368	25	0.00466	0.00479
51	4,986	26	0.00521	0.00538
52	4,832	30	0.00621	0.00603
53	5,298	37	0.00698	0.00675
54	5,741	45	0.00784	0.00756
55	4,866	46	0.00945	0.00844
56	4,901	52	0.01061	0.00942
57	5,003	63	0.01259	0.01050
58	3,952	45	0.01139	0.01169
59	2,786	45	0.01615	0.01299



[3]

(ii) The null hypothesis is that the old rates are the true rates underlying the observed data. [½]

Age	Actual Deaths	μ new	μ old	Expected Deaths	z	z^2
50	25	0.00466	0.00479	25.7127	-0.1406	0.0198
51	26	0.00521	0.00538	26.8247	-0.1592	0.0254
52	30	0.00621	0.00603	29.1370	0.1599	0.0256
53	37	0.00698	0.00675	35.7615	0.2071	0.0429
54	45	0.00784	0.00756	43.4020	0.2426	0.0588
55	46	0.00945	0.00844	41.0690	0.7694	0.5920
56	52	0.01061	0.00942	46.1674	0.8584	0.7369
57	63	0.01259	0.01050	52.5315	1.4444	2.0862
58	45	0.01139	0.01169	46.1989	-0.1764	0.0311
59	45	0.01615	0.01299	36.1901	1.4644	2.1446

[1½]

- The observed test statistic is 5.76. [½]
- There are 10 age groups and no graduation has been performed, [½]
- so there are ten degrees of freedom. [½]
- The upper tail value of the chi-squared distribution with 10 degrees of freedom at the 95% level is 18.31. [½]
- Since 5.76 < 18.31, [½]
- we have insufficient evidence to reject the null hypothesis. [½]

[Total 5]

(iii)

Signs Test.

- The null hypothesis is the same as before. [½]
- Under the null hypothesis the standard deviations are distributed
- Binomial (10, 0.5). [½]
- There are ten deviations in total of which three are negative. [½]
- The likelihood of getting three or fewer negative deviations is 0.1719. [1]
- Which exceeds 2.5% (two tailed test) [1]
- So there is insufficient evidence to reject the null hypothesis. [½]

Grouping of Signs Test.

- The null hypothesis is the same as previously stated. [½]
- G = number of groups of positive deviations = 2
- m = number of deviations = 10
- n_1 = number of positive deviations = 7

- n_2 = number of negative deviations = 3 [1]

- We want k^* the largest k such that
- $\displaystyle\sum_{t=1}^{k} \frac{\binom{n_1-1}{t-1}\binom{n_2+1}{t}}{\binom{m}{n_1}} < 0.05$. [½]
- The test fails at the 5% level if $G \le k^*$.
- From the Gold Book the value of k^* is 1. [½]
- Since, therefore, $G > k^*$ in this case [½]
- we have insufficient evidence to reject the null hypothesis at the 5% level. [½]

[7½, Max. 6]

(iv)

- In part (iii) the Signs Test indicated that there was no bias in the rates and the Grouping of Signs test indicated that the new rates were the same shape as the old rates. [½]
- The graph of the old rates in part (i) follows a roughly smooth line [½]
- whereas the new rates would but for the apparent anomaly at age 58. [½]
- If we ignore the age 58 figure, the new rates look consistently above the old ones and increasingly so with age. [½]
- This would suggest a different shape from the old rates. [½]
- Indeed the Grouping of Signs test would have failed but for the second group of positives after age 58. [½]
- The deaths at age 58 in the current investigation should be investigated. [½]
- Perhaps the excessive number of deaths was due to multiple policies held by the same policyholder. [½]

[4, Max. 3]

[Total 17]

[Paper Total 100]

Notes to Question 11:

(1) In part (i) full marks for a correct sketch of the graph. One mark for each curve and ½ mark for labelling each axis. If the sketch is wrong or is missing, then up to 1 mark is awarded for the fourth column of the table

(2) In part (ii) one mark is deducted for excessive rounding

(3) In part (ii) some justification for 10 degrees of freedom should be provided to score full marks

(4) In part (iii) no credit is given for the wrong test being applied

(5) In part (iii) approximating the Binomial distribution with a Normal distribution receives no credit as the number of age groups is not sufficiently large

A mixed performance was noted in this question especially in the later parts.

END OF EXAMINERS' REPORT