# Institute of Actuaries of India

## Subject CS2A – Risk Modelling and Survival Analysis (Paper A)

## July 2022 Examination

# INDICATIVE SOLUTION

**Introduction**
The indicative solution has been written by the Examiners with the aim of helping candidates. The solutions given are only indicative. It is realized that there could be other points as valid answers and examiner have given credit for any alternative approach or interpretation which they consider to be reasonable.

**Solution 1:**

Let T = Aggregate Total Losses, N = No. of claims and X = Amount of claims
The possible states for aggregate total losses are:
T=0; T=10000; T=20000; T=100000; T=110000; T=200000

- P(T=0) = P(N=0) = 60%
- P(T=10000) = P($X_1$=10000|N=1) = 80%*20% = 16%
- P(T=20000) = P($X_1$=10000,$X_2$=10000|N=2) = 80%*80%*20% = 12.8%
- P(T=100000) = P($X_1$=100000|N=1) = 20%*20% = 4%
- P(T=110000) = P($X_1$=100000 & $X_2$=10000|N=2) = 20%*80%*20% = 3.2%
- P(T=200000) = 1 – P(T=0) - P(T=10000) - P(T=20000) - P(T=100000) - P(T=110000) = 1 – 60% - 16% - 12.8% - 4% - 3.2% = 4%

**[6 Marks]**

**Solution 2:**

**i)** 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 50th birthday, 1 January 2019) and
ENDDATE is the earliest of (date of 51st birthday, date of death, 31 December 2019)

| *Life* | *Date of 50th birthday* | *Date of death* | *Start Date* | *End Date* | *Duration in months between Start and End of observation* |
|---|---|---|---|---|---|
| 1 | 01 February 2018 | – | 01 January 2019 | 31 January 2019 | 1 |
| 2 | 01 April 2018 | 01 October 2019 | 01 January 2019 | 31 March 2019 | 3 |
| 3 | 01 June 2018 | – | 01 January 2019 | 31 May 2019 | 5 |
| 4 | 01 September 2018 | – | 01 January 2019 | 31 August 2019 | 8 |
| 5 | 01 November 2018 | 15 March 2019 | 01 January 2019 | 15 March 2019 | 2.5 |
| 6 | 01 January 2019 | – | 01 January 2019 | 31 December 2019 | 12 |
| 7 | 01 May 2019 | 15 December 2019 | 01 May 2019 | 15 December 2019 | 7.5 |
| 8 | 01 July 2019 | 01 October 2019 | 01 July 2019 | 01 October 2019 | 3 |
| 9 | 01 August 2019 | – | 01 August 2019 | 31 December 2019 | 5 |
| 10 | 01 December 2019 | – | 01 December 2019 | 31 December 2019 | 1 |
| | | | | **Total** | **48** |

Central exposed to risk is the sum of contribution of each of the 10 lives (in number of months) to the observation i.e. 48 months or 4 years. [3]

**ii)** The total number of deaths during the period of observation is 3. So, the maximum likelihood estimate of the hazard of death is 3/4 = 0.75. [1]

**iii)**

ALTERNATIVE 1

If the hazard of death at age 50 years is $\mu_{50}$ , then

$q_{50} = 1 - p_{50} = 1 - \exp(-\mu_{50})$

$= 1 - \exp(-0.75) = 1 - 0.4724 = 0.5276.$

ALTERNATIVE 2

If the central exposed to risk is $E_{50}^{c}$, then if we work in years

$$q_{50} \approx \frac{d_{50}}{E_{50}^{c} + 0.5 * d_{50}}$$

$$= \frac{3}{4 + 0.5 * 3}$$

= 3/5.5 = 0.5454

[2]

**[6 Marks]**

**Solution 3:**

We know that ${}_t p_x = \left[\exp\left(\frac{-B}{\ln c}\right)\right]^{(c^x * (c^t - 1))}$

From the above survival probabilities,

x = 55

ln 0.9925 = $(c^{55} * (c - 1))$* (-B/ln c)

ln 0.9843 = $(c^{55} * (c^2 - 1))$* (-B/ln c)

ln 0.9843 / ln 0.9925 = $\frac{(c^2 - 1)}{(c - 1)}$

c + 1 = 2.10202

Hence, c = 1.10202

ln 0.9925 = $(1.10202^{55} * (1.10202 - 1))$* (-B/ln 1.10202)

Solving the above equation, we get

B = 3.42774*10$^{-5}$

**[4 Marks]**

**Solution 4:**

**i)**

$L = \prod_{i=1}^{10} [0.5 \propto X 0.4^{\propto} X x_i{}^{0.5} X (0.4 * x_i{}^{0.5})^{-(\propto+1)}] \; X \; [P[X > 234]]^7$

$= 0.5^{10} \; X \propto^{10} \; X 0.4^{10\propto} X \prod_{i=1}^{10} [x_i{}^{0.5}] \; X \prod_{i=1}^{10} [0.4 + x_i{}^{0.5}]^{-(\propto+1)} X \frac{0.4}{0.4+2340.5}^{7\propto}$

$\propto \propto^{10} \; X 0.4^{10\propto} X 0.40.0255^{7\propto} X \prod_{i=1}^{10} [0.4 + x_i{}^{0.5}]^{-(\propto+1)}$

⇨ $\mathit{ln}\, L \propto 10\mathit{ln}\propto + 10\propto \mathit{ln}(0.4) - (\propto + 1) \sum_{i=1}^{10} \mathit{ln}(0.4 + x_i{}^{0.5}) + 7 \propto \mathit{ln}(0.0255)$

$= 10\mathit{ln}\propto - 0.9163 X 10\propto \; -25.68 \propto \; - 26.47 \propto$

$= 10\mathit{ln}\propto - 61.31\propto$

⇨ $\frac{d\mathit{ln}L}{d\propto} = \frac{10}{\propto} - 61.31 = 0$

⇨ $\propto = 0.1631$

$$\frac{d^2 \mathit{ln}L}{d \propto^2} = \frac{-10}{\propto^2} < 0$$

[5]

**ii)** *Median = 217*

⇨ $1-(\frac{0.4}{0.4+217^{0.5}})^{\propto} = 0.5$

⇨ $(\frac{0.4}{0.4+217^{0.5}})^{\propto} = 0.5$

⇨ $(0.02643)^{\propto} = 0.5$

$\propto = 0.1907$ [3]

**iii)** The method of percentiles using median assumes is a more robust way to estimate the parameters as it assumes that the population has equivalent median to that observed in the sample. While on the other hand, the Maximum Likelihood Estimator is a more efficient way as it makes more stricter assumptions of full density. It estimates the most likely underlying parameters of the distributions. This inherent basis of the two methods leads to difference value of $' \propto '$ in the above parts. [2]

**[10 Marks]**

**Solution 5:**

**i)** Receiver operating characteristic curve. The trade-off between recall and the false positive rate can be illustrated using a receiver operating characteristic (ROC) curve. The further away from the diagonal is the ROC, the greater the area under the curve and the better the model is at correctly classifying the cases. [1]

**ii)** AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing [2]

**iii)** Precision is the percentage of cases classified as positive that are, in fact, positive. Using the abbreviations in the table this is:

Precision = $\frac{TP}{TP+FP}$

Recall is the percentage of positives that we managed to identify (correctly):

Recall = $\frac{TP}{TP+FN}$

These can be combined in a single measure known as the

F1 score = $\frac{2\ X\ Precision\ X\ Recall}{Precision+Recall}$ [3]

**[6 Marks]**

**Solution 6:**

**i)** The order s will be 3, i.e. $Y_t=\nabla_3X_t=X_t-X_{t-3}$

The characteristic polynomial will be $1-(\alpha+\beta)z+\alpha\beta z^2$ with roots $1/\alpha$ and $1/\beta$.

Hence, process is stationary for $|\alpha|<1$ and $|\beta|<1$. [1]

**ii)** The characteristic polynomial will be $1-(\alpha+\beta)z+\alpha\beta z^2$ with roots $1/\alpha$ and $1/\beta$.

Hence, process is stationary for $|\alpha|<1$ and $|\beta|<1$. [2]

**iii)** $\rho_1-(\alpha+\beta)+\alpha\beta\rho_1=0$

$\rho_2-(\alpha+\beta)\ \rho_1+\alpha\beta=0$ [2]

**iv)**

$\rho_1-(0.4+0.2)+(0.4)(0.2)\rho_1=0$

$\rho_2-(0.4+0.2)\ \rho_1+(0.4)(0.2)=0$

Solving for $\rho_1$ and $\rho_2$, we get:
$\rho_1$=1/3 and $\rho_2$=0.12

[5]

**v)** $\alpha+\beta$= 0.6

$\alpha\beta$=0.08

$Y_t = X_t - X_{t-s}$
$X_{101}$ =$Y_{101}$+ $X_{98}$

And
$X_{102}$ =$Y_{102}$+ $X_{99}$

Forecasted values:
$x^{\wedge}101 = y^{\wedge}101 + x_{98}$ and
$x^{\wedge}102 = y^{\wedge}102 + x_{99}$
where,
$y^{\wedge}101$ =0.6$y_{100}$ + 0.08$y_{100}$
=0.6($x_{100}$-$x_{97}$) + 0.08($x_{99}$-$x_{96}$)
And

$y^{\wedge}102$=0.6$y^{\wedge}_{101}$+ 0.08($x_{100}$-$x_{97}$)

[5]
**[15 Marks]**

**Solution 7:**

**i)**
Mean = 1000 = lambda /(alpha-1)
⇨ Lambda = 1000 *(alpha -1)

Variance = 1200^2 = alpha * lambda^2 / [(alpha-1)^2*(alpha-2)]
⇨ 1200^2 = 1000^2*alpha/(alpha-2)

Solving the above two equations,
Alpha = 6.55
Lambda = 5545.45

Probability that reinsurance will be involved for 1st case:
=P(X>3300)
= 1-(1-(lambda/(lambda+2200))^alpha)
= (lambda/(lambda+2200))^alpha
= (5545.45/(5545.45+3300))^6.55
= 0.047

Probability that reinsurance will be involved for 2nd case:
= P(X>2200)
= 1-(1-(lambda/(lambda+2200))^alpha)
= (lambda/(lambda+2200))^alpha
= (5545.45/(5545.45+2200))^6.55
= 0.1123 [4]

**ii)**
For 2nd arrangement:
Z(amount to reinsurer):

Z = 0 if X< 2200

= X-2200 if X > 2200

For 1st case:
E[Z] = INTEGRAL [(x-2200) * f(x)*dx] (2200, Inf)
=INTEGRAL [(xf(x)*dx] (2200, Inf) - INTEGRAL [(2200*f(x)*dx] (2200, Inf)
= A – B

A:
INTEGRAL [(6.55*x*5545.455^6.55 / (5545.45+x)^7.55dx] (2200, Inf)
= INTEGRAL [(2.1828*10^25 *x/ (5545.45+x)^7.55dx] (2200, Inf)

Integrating by parts:
INTEGRAL [p * dq/dx] = p*q – INTEGRAL [dp/dx*q]
Where, p = x
dq/dx = (x+5545.45)^-7.55
= 2.1828*[-.15267*x/(x+5545.45)^6.55] – INTEGRAL [-.15267 dx / (x+5545.45)^6.55]
= -1*.15267* INTEGRAL [du/u^6.55], where u = x + 5545.45; du/dx = 1; dx = du
= -6.0047*10^23/(x+5545.45)^5.55 – 3.3326*10^24*x/(x+5545.45)^6.55
=402.99
B:
INTEGRAL [(6.55*2200*5545.455^6.55 / (5545.45+x)^7.55dx] (2200, Inf)
= -7.33176*10^27/(x+5545.45)*6.55
=246.58
Therefore, E[Z] = 156.41

Similarly for 1st case:
E[Z] = INTEGRAL [(x-3300) * f(x)*dx] (3300, Inf)
= 229.83-154.98 [6]
=74.85

**iii)**
Expected amount paid out by reinsurance per 1 unit of premium:
2nd case: 500*0.25*156.41/3900
= 5.01
1st case: 500*0.25*74.85/3000
= 3.11

The above inference is drawn based on the consideration of the amount of reinsurance being paid out for each unit of premium pertaining to the reinsurance limits and the respective premiums given. From the insurer's point of view, the more efficient reinsurance limit would be the one where for each unit premium, more reinsurance is being paid. Accordingly, 1st case is better. [2]

**[12 Marks]**

**Solution 8:**

**i)** 11 athletes qualified during the period of observation, so the median is the number of events taken to qualify by the sixth athlete to qualify. This is 9 events. [2]

**ii)** Define *t* as the number of events which have taken place since 1 Jan 2014.
Injured and stopped participating implies recorded *after* the event number reported.

| $t_j$ | $N_j$ | $D_j$ | $C_j$ | $D_j/N_j$ | $1-D_j/N_j$ |
|---|---|---|---|---|---|
| 0 | 23 | 0 | 2 | 0 | 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-D$_j$/N$_j$**

Then the Kaplan-Meier estimate of the survival function is

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

[5]

**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 events. [2]

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

The athletes still participating at the end of 2020 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 athletes who qualified during the period. [3]

**[12 Marks]**

**Solution 9:**

Using Kendall's tau:

$\alpha$=2×0.08/(1−0.08)= 0.1739

Use the generator function for loss in 'A':

$$1/\alpha\times(0.05^{-\alpha}-1) = 1/0.1739 \times (0.05^{-0.1739}-1) = 3.931$$

Use the generator function for loss in 'B':

$$1/\alpha\times(0.40^{-\alpha}-1) = 1/0.1739 \times (0.40^{-0.1739}-1) = 0.993$$

Reverse the generator function to obtain the joint probability of loss of [1]

$(3.931 +0.993)/1/0.1739+1^{)(1/-0.1739)}=0.02852$

**[4 Marks]**

**Solution 10:**

**i)** The score currently stands at 'Tie'. Whichever team wins the next point will move into a 'Lead'. If the team in 'Lead' wins the subsequent point as well, they would win the tournament. However, if the team in 'Lead' loses the next point, the score would be back at 'Tie'.

Since the probability of moving to the next state does not depend on the history prior to entering the state, Markov property holds.

The state space is defined as follows:

| State | Description |
|---|---|
| T | Tie |
| $L_{LSG}$ | LSG Leads |
| $L_{GT}$ | GT Leads |
| $G_{LSG}$ | LSG Wins |
| $G_{GT}$ | GT Wins |

[2.5]

**ii)**

$$\text{a)}\ \begin{bmatrix} 0 & 0.6 & 0.4 & 0 & 0 \\ 0.4 & 0 & 0 & 0.6 & 0 \\ 0.6 & 0 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

[2.5]

**iii)**

After two points from the tie, the match would either be completed or be back to tie again.
The probability of returning to tie after two points is given by:
Probability of LSG winning the first point x Probability of GT winning the second point
+
Probability of GT winning the first point x Probability of LSG winning the second point
= 0.6 x 0.4 + 0.4 x 0.6 = 0.48
The number of such cycles (N) of returning to tie can be found by:
$0.48^N = 1 - 0.9$
Solving the above equation:
N= ln(1-0.9)/ln 0.48
= 3.14
Since the match can finish in cycles of two points, the required number of cycles is 4 i.e. 8 points. [4]

**iv)**

After two points:
a. GT may have won the match with a probability of 0.16 (= $0.4^2$); or
b. LSG may have won the match with a probability of 0.36 (= $0.6^2$); or
c. It may have come back to tie with a probability of 0.48 (as calculated above).

Let $LSG_T$ be the probability that LSG wins the match that is presently tied.
Let $GT_T$ be the probability that GT wins the match that is presently tied.

We have:
$GT_T = 0.16 + 0.48 \times GT_T$

Solving, $GT_T = 0.308$

Probability that LSG eventually wins the match is 0.692 ($1-GT_T$). This can be verified by:
$LSG_T = 0.36 + 0.48 \times LSG_T$

Solving, $LSG_T = 0.692$ [4]

**v)** Probability of GT winning a point is 0.4. However, in order to win the game, GT would need to win at least two consecutive points. The probability of GT winning two consecutive points is lower than the probability of winning a point. At the same time, the probability of LSG winning the tournament would be more than the probability of GT winning it as the probability of LSG winning one point is more than that of GT winning a point. [2]

**[15 Marks]**

**Solution 11:**

**i)** Solution: A [1]

**ii)** Solution: C [2]

**iii)** Solution: A [2]

**iv)** Solution: C [2]

**[7 Marks]**

**Solution 12:**

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) young population of a county in order to study accident hump

Graphical graduation, as it highly likely that there is small amount of data and it is used to study some known features in the data i.e. accident hump

c) entire population of a large developed country to prepare a new standard table

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

**[3 Marks]**

*******************