Institute of Actuaries of India

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

November 2020 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:

i) Answer – (b) [2]

ii) Answer – (b) [2]

iii) Answer – (b) [2]

iv) Answer – (a) [2.5]

v) Answer – (d) [2.5]

vi) Answer – (a) [1]

vii) Answer – (c) [3]

viii) Answer – (c) [3]

ix) Answer – (b) [5]

x) Answer – (d) [5]

xi) Answer – (i)-(C), (ii)-(A), (iii)-(D), (iv)-(B) [2]

[30 Mark]

Solution 2:

i) Answer – (c) **[2]**

ii) Roots of characteristic equation are -1+/-sq rt(6) , which are outside (-1, +1),

so {Yt} is stationary. **[1.5]**

iii)

Mean is stationary over time
(1-0.4-0.20) E [Yt] =0.025
So E [Yt]=0.025/0.4=0.0625 **[1]**

iv)

$Y_t - 0.0625 = 0.4(Y_{t-1} - 0.0625) + 0.2(Y_{t-2} - 0.0625) + Z_t$
$\rho_k = E[(Y_t - 0.0625)(Yt-k - 0.0625)] = 0.4\,\rho_{k-1} + 0.2\,\rho_{k-2}$

[2]

Put k=1 , and note that $\rho_0 = 1$ and $\rho_{-1} = \rho_1$
Therefore
$\rho_1 = 0.4 + 0.2\,\rho_1 \Rightarrow \rho_1 = 0.5$
$\rho_2 = 0.4\,\rho_1 + 0.2 = 0.4$
$\rho_3 = 0.4\,\rho_2 + 0.2\,\rho_1 = 0.26$
and so on [2]

[4]

v)

$(1-0.4B-0.2B^2)(Y_t - 0.0625) = z_t$

$Y_t - 0.0625 = (1-0.4B-0.2B^2)^{-1} Z_t$

Invert $(1-0.4B-0.2B^2)$ and multiply by Z_t to obtain equivalent moving average process. **[1.5]**

vi)

R_t follows MA(1), hence we can write

$R_t = e_t + \beta e_{t-1}$,Where $e_t \sim (0, \sigma^2)$ [0.5]

Now, $var(R_t) = var(e_t + \beta e_{t-1})$

$= \text{var}(e_t) + \beta^2 \text{ var}(e_{t-1})$ [0.5]

$= (1+\beta^2)\sigma^2$ .....(1) [0.5]

Now, $\Delta S_t = (0.8 + 0.5t + R_t) - [0.8 + 0.5(t-1) + R_{t-1}]$

$= 0.5 + (R_t - R_{t-1})$ [0.5]

Hence$var(\Delta S_t) = [cov(R_t - R_{t-1}, R_t - R_{t-1})]$ [0.5]

$= [2\gamma_R(0) - \gamma_R(-1) - \gamma_R(1)]$ .....(2) [0.5]

$$\therefore Now, \gamma_R(0) = (1+\beta^2)\sigma^2$$

And, $\gamma_R(1) = \gamma_R(-1) = \text{cov}(e_t + \beta e_{t-1}, e_t + \beta e_{t-1}) = \beta\sigma^2$ [0.5+ 0.5]

Therefore, from (2) we get,

$var(\Delta S_t) = [2(1+\beta^2)\sigma^2 - 2\beta\sigma^2] = 2\left[1 - \beta + \beta^2\right]\sigma^2$ [1]

Now, $var(\Delta S_t) - var(R_t)$

$= [2 - 2\beta + 2\beta^2]\sigma^2 - (1+\beta^2)\sigma^2$ [0.5]

$= [1 - 2\beta + 2\beta^2]\sigma^2$ [0.25]

$= (1-\beta^2)\sigma^2 > 0$ [0.5]

Hence the standard deviation of first difference of S_t is higher than that of R_t [0.25]

[6]

[16 Marks]

Solution 3:

i) The Null Hypothesis is that the company's experience is in line with population

a)

Age	Observed (O)	Expected (E) = Population mortality* Exposed to risk	(O-E)^2	(O-E)^2/E
30	5	4.688205	0.097216	0.020736
31	6	6.67408	0.454384	0.068082
32	7	6.59976	0.160192	0.024272
33	7	8.29168	1.668437	0.201218
34	9	9.04125	0.001702	0.000188
35	10	11.05434	1.111633	0.100561
36	13	12.65135	0.121555	0.009608
37	12	13.28202	1.643568	0.123744

38	13	13.64468	0.415617	0.03046
39	14	13.58506	0.172178	0.012674
Total	96	99.51243	5.846482	0.591544

Test statistic = 0.59 [4]

b) Degrees of freedom = 10 [0.5]

c) upper tail value of the Chi-square distribution with 10 degrees of freedom at 95% is 18.31 [0.5]

d) The company's experience is in line with population as there is not sufficient evidence to reject the null hypothesis. 0.59 < 18.31 [1]

ii)

No. of positive = 4
N = total no. of groups analysed = 10
P follows Binomial (10,1/2)
2*P(P<= 4) = 2*0.3770

[2]

Page 187 of tables by looking up n=10, p=0.5, x=4
p-value = 0.3770 [1]

Since p-value is greater than 5%, it is not significant. [1]

[4]

[10 Marks]

Solution 4:

i)

a) Data : RSRRSSSSRRSRSSRSRSRRSSSRRSSRRR

$p_{rr} = 6/14=3/7$

$p_{rs} = 8/14=4/7$

$p_{sr} = 8/15$

$p_{ss} = 7/15$

[2]

b)

30th June	1st Jul	2nd Jul	3rd Jul	
R	S	S	R	
	$p_{rs} = 8/14$	$p_{ss} = 7/15$	$p_{sr} = 8/15$	0.142222
R	S	R	R	
	$p_{rs} = 8/14$	$p_{sr} = 8/15$	$p_{rr} = 6/14$	0.130612
R	R	S	R	
	$p_{rr} = 6/14$	$p_{rs} = 8/14$	$p_{sr} = 8/15$	0.130612
R	R	R	R	
	$p_{rr} = 6/14$	$p_{rr} = 6/14$	$p_{rr} = 6/14$	0.078717

Total = 0.482164 [3]

ii)

a)

Here, $\mu = 3$

"Some policies " means "1 or more policies" i.e 1 minus the "zero policies" probability:

$P(X > 0) = 1 - P(x_0)$

Now, $P(X) = \frac{e^{-\mu}\mu^x}{x!}$

So, $P(x_0) = \frac{e^{-3}3^0}{0!}$ = 4.9787 x 10^{-2} [1]

Therefore the probability of 1 or more policies is given by:

Probability = $P(X \geq 0)$
= $1 - P(x_0)$
= 1- 4.9787 x 10^{-2}
= 0.95021 [1]
[2]

b)

The probability of selling 2 or more, but less than 5 policies is:

$P(2 \leq X < 5) = P(x_2) + P(x_3) + P(x_4)$ [1]

$= \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!} + \frac{e^{-3}3^4}{4!}$

= 0.61611 [1.5]
[2.5]

c)

Average number of policies sold per day: $\frac{3}{5}$ = 0.6 [1]

So on a given day, $P(X) = \frac{e^{-0.6}0.6^1}{1!}$ = 0.32929 [1.5]
[2.5]
[12 Marks]

Solution 5:

i) Answer – (a) [3]

ii)

The F1 score may be written as

$$\frac{2*[TP/(TP+FP)]*[TP/(TP+FN)]}{[TP/(TP+FP)]+[TP/(TP+FN)]}$$

Solving this, we get F1 = TP/(TP+0.5FN+0.5FP)

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

[2]

iii)

Comments -

- Compared with questionnaire, the symptom test is better at identifying the true positives
- But it is not so precise as it classifies as positive a higher proportion of those who do not have the disease.
- 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
- 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.
- 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.
- 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
- As the sample size is relatively small, the test should be re-performed on a larger population before drawing any firm conclusions.
- 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.

[3]

[8 Marks]

Solution 6:

i)

$E(X_t) = \alpha + (1+k)(1-\alpha)$

$= 1 + k(1-\alpha)$

[1]

$Var(X_t) = E(X_i^2) - E(X_i)2$

$= \alpha + (1+k)^2(1-\alpha) - (1+k(1-\alpha))^2$

$= \alpha + (1-\alpha) + 2k(1-\alpha) + k^2(1-\alpha) - 1 - 2k(1-\alpha) - k^2(1-\alpha)^2$

$= k^2(1-\alpha)(1-(1-\alpha))$

$= k^2\ \alpha(1-\alpha)$ [2]

[3]

ii)

Let Y denote the aggregate claims in a year. Then Y has a compound Poisson distribution,

$E(Y) = 500 \times E(X_i) = 500 + 500k(1-\alpha) = 500 + 400k = 633.60$

[1]

And

$Var(Y) = 500 \times E(X_i^2)$

$= 500(\alpha + (1+k)^2(1-\alpha))$

$= 500(\alpha + (1-\alpha) + 2k(1-\alpha) + k^2(1-\alpha))$

$= 500 + (1000k + 500k^2)(1-\alpha)$
$= 500 + 800k + 400k^2 = 811.82$ [1.5]

Using a normal approximation
P(Y>700)=P(Z> [700-E(Y)]/Sq rt(Var(Y)) [1]

By putting, K=0.334, we get the probability 1% (full marks who derived the 1% probability).

[1.5]
[5]
[8 Marks]

Solution 7:

i)

Lost to the investigation for each state of the model.

- Infected may or may not get tested if testing is not compulsory. Thus, may be considered in "Normal" state. Thus lost from "infected" state and in turn from "recovered" or "death" state.
- Infected may recover even before the testing is suggested by doctor. Thus, may be considered in "Normal" state. Thus lost from "infected" state and in turn from "recovered" or "death" state.
- Any other cause of death than this infection will be considered transited in "Normal" to "death" state. However, if it is not tested it will be lost from "infected" and then to "death" state.
- There will not be any lost to investigation for "Normal" state.

[2]

ii)

It is non-informative censoring for lifetime investigation.

- Infected may or may not get tested if testing is not compulsory. Thus, may be considered in "Normal" state. Thus lost from "infected" state and in turn from "recovered" or "death" state. – Non-informative censoring as information of lifetime to perform survival/ mortality analysis is not captured under appropriate state.
- Infected may recover even before the testing is suggested by doctor. Thus, may be considered in "Normal" state. Thus lost from "infected" state and in turn from "recovered" or "death" state. – Non-informative censoring as information of lifetime to perform survival/ mortality analysis is not captured under appropriate state.
- Any other cause of death than this infection will be considered transited in "Normal" to "death" state. However, if it is not tested it will be lost from "infected" and then to "death" state. – Non-informative censoring as information of lifetime to perform survival/ mortality analysis is not captured under appropriate state.
- For "Normal"state – Informative censoring as the survival expectation of this group is higher than the "infected" or "recovered" individuals.

[2]

iii)

Other types of censoring as below:

- Right censoring – since till vaccine is prepared and not till end of infection. [0.5]
- Interval censoring – since we may only know the age to the nearest birthday and not exact date of birth as we have started capturing the information from certain time point. [0.5]

[1]

iv)

List of covariates to perform survival analysis:

a) Age
b) Gender
c) Severity of symptoms
d) Number of days infection lasted
e) Co-morbid conditions such as BP, diabetes
f) Physically Active/ non-active
g) Level of fitness –no. of hours of weakly exercise/ activities
h) Profession
i) Socio – economic group / Income range
j) States
k) Metro/Non-metro city
l) Density of population

[3]

v)

The lifetime distributions for proportional hazard functions are as follows:

a) The Exponential distribution [0.25]
The hazard rate under this lifetime distribution function is **constant**. [0.25]
This constant hazard model could reflect the hazard for an individual who remains in good health. Here, the level of hazard would reflect the risk of death from unnatural causes. [0.5]

b) The Weibull distribution [0.25]
The hazard rate under this lifetime distribution function is **monotonically** decreasing or increasing. [0.25]
The decreasing hazard model could reflect hazard for patients recovering from major surgery. The level of hazard is expected to fall as the time since the operation increases. [0.5]

c) The Gompertz-Makeham formula [0.25]
The lifetime distribution function has **exponential** hazard rate. [0.25]
The exponential hazard model could reflect the hazard for leukaemia sufferers who are not responding to treatment. The severity of condition and the level of hazard increase with the survival time. **OR** Over longer period exponential hazard rate could reflect the increasing chance of death from natural causes as age increases. [0.5]

d) The log-logistic distribution (humped hazard) [0.25]
The lifetime distribution function has **humped** hazard rate. [0.25]
The humped hazard model could reflect the hazard for patients with disease that is most likely to cause death during the early stages. As the initial condition becomes more severe, the level of hazard increases. But once the patient survives the period of highest risk, the level of hazard decreases. [0.5]

To perform survival analysis of population of Originia:

a) The Exponential distribution can be used for the population which continues to be in “normal” state as they are expected to follow the natural and normal survival/ death expectations. However, it may not be applicable for all age group.

b) The Weibull distribution can be used for the population which recovers from the being infected. Their survival probability is expected to get better and better as they recover from being infected. The pandemic is fairly new and it will not be very clear whether the recovered population may or may not have any future implications/ re-occurrence/ other side effects due to getting infected once.

c) The exponential hazard function could be appropriate for age group which is most impacted by this infection. The severity of condition and the level of hazard may increase with the survival for short span of time. Also, they may be no time to observe this trend in case the infection results into deaths in really very short span of time. It does not look appropriate to apply this hazard function to any of the state of the analysis.

d) The log-logistic distribution (humped hazard) is likely to reflect the hazard infected population with most impacted age group as it is most likely to cause death during the early stages. As the initial condition becomes more severe, the level of hazard increases. But once the patient survives the period of highest risk, the level of hazard decreases.

[4]
[8]
[16 Marks]

****