Institute of Actuaries of India

Subject CS2-Paper A – Risk Modelling and Survival Analysis

June 2019 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:

Use the discharge date as the censoring date. For those who are still in the hospital, use the last day as the censoring date.

Right censoring is present since we don't know the exact future lifetime for the lives that withdrew or left the investigation on 1st Jan 2018.

Random censoring occurs since we don't have the discharge times in advance.

Type I censoring occurs since lives alive at the time 1st Jan 2018 and in hospital are certain to be censored.

Non informative censoring is also present since the exits give us no information about the future death of the lives remaining in the investigation.

[2 Marks]

Solution 2:

Let X denote your waiting time in minutes, and let N(t) be the process counting the arrivals of passenger from the moment you get in the taxi.

N(t) is a Poisson process of parameter λ = 1 passenger per minute.

Let S1 denote the first arrival time of the process.

We have E[X] = E[X|S1 ≥ 3]P{S1 ≥ 3} + E[X|S1 < 3]P{S1 < 3} [1]

E[X|S1 ≥ 3]P{S1 ≥ 3} = 3e $^{-3}$ [0.5]

E[X|S1 < 3]P{S1 < 3} = E[S1|S1 < 3]P{S1 < 3} = integrate(0,3) xf_{S1} (x)dx

= integrate(0,3) xe^{-x} dx

= $-xe^{-x}\ |^3{}_0$ + integrate(0,3) e $^{-x}$ dx = 1 − 4e $^{-3}$ [0.5]

E[X] = 1 − e $^{-3}$ = 0.95 ≈ 57sec. [1]

[3 Marks]

Solution 3:

i) The four ways of measuring the tail weight of a distribution:

- The existence of moments
- Limiting density ratios
- The hazard rate
- The mean residual life [1]

ii) $\lim_{x\to\infty} \frac{f_{\alpha=0.2}(x)}{f_{\alpha=0.3}(x)} = \lim_{x\to\infty} \left\{ \frac{0.2\lambda^{0.2}}{(\lambda+x)^{1.2}} \Big/ \frac{0.3\lambda^{0.3}}{(\lambda+x)^{1.3}} \right\}$

$= \frac{2}{3} * \frac{1}{\lambda^{0.1}} \lim_{x\to\infty} (\lambda + x)^{0.1}$

$= \infty$

Then the distribution with α = 0.2 has a much thicker tail. [2]

[3 Marks]

Solution 4:

i) The confusion matrix for Method I and II are as follows:

Method I

Predicted → Actual	**Yes**	**NO**	**Total**
Yes	65	15	80
NO	10	10	20
Total	75	25	100

Method II

Predicted → Actual	**Yes**	**NO**	**Total**
Yes	55	15	70
NO	15	15	30
Total	70	30	100

[2]

$$F_1 \text{ Score for Method I} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where Precision = True Positive / (True Positive + False Positive)
And Recall = True Positive / (True Positive + False Negative)

Therefore, F_1 Score for Method I is 84%
F_1 Score for Method II is 79% [1]

False Positive Rate = False Positive / (True Negative + False Positive)

And False Positive Rate for both Method I and Method II is 50%. [1]

F_1 Score of Method I is more than that of Method II. The false positive rate is same for both Methods. So Method I is more favourable than Method 2.

[1]

[5]

ii) s-fold cross-validation:

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

In a s-fold cross-validation, the original sample is randomly partitioned into s equal size samples. One of the s subsamples is retained as the validation data for testing the model. The remaining s-1 subsamples are used as training data. The cross-validation process is then repeated s times with each of the s subsamples used exactly once as the validation data. The s results from the folds can then be compared or averaged to produce a single prediction.

The advantage of this method is that all observations are used for both training ad validation, and each observation is used for validation exactly once. [3]

[8 Marks]

Solution 5:

i) The Chapman – Kolmogrov equations are

$$P_{ij}(s,t) = \sum_{k\in s} P_{ik}(s,u)P_{kj}(u,t)$$

To obtain the forward equations we differentiate with respect to t and evaluate at u=t;

$$\frac{\partial}{\partial t}P_{ij}(s,t) = \sum_{k\in s}\left[P_{ik}(s,u)\left(\frac{\partial}{\partial s}P_{kj}(u,t)\right)\right]_{u=t} = \sum P_{ik}(s,t)\mu_{kj}(t)$$

[0.5]

Similarly the backward equations are obtained by differentiating with respect to s and setting u=s ;

$$\frac{\partial}{\partial s}P_{ij}(s,t) = \sum_{k\in s}\left[\left(\frac{\partial}{\partial s}P_{ik}(s,u)\right)P_{kj}(u,t)\right]_{u=s} = -\sum_{k\in s}\mu_{ik}(s)P_{kj}(s,t)$$

[0.5]

We now need to explain where the minus sign in the RHS comes from.
The definition of the transition rates is such that;

$$P_{ik}(s,s+h) = \delta_{ik} + h\mu_{ik}(s) + o(h)$$

[0.5]

Or equivalently;

$$P_{ik}(s-h,s) = \delta_{ik} + h\mu_{ik}(s-h) + o(h)$$

[0.5]

Rearranging this gives:

$$\mu_{ik}(s-h) = \frac{P_{ik}(s-h,s) - \delta_{ik} - o(h)}{h}$$

Now taking the limit of both sides as h->0 and nothing that $P_{ik}(s,s)=\delta_{ik}$, we get

$$\mu_{ik}(s) = -\lim_{h\to 0}\frac{P_{ik}(s-h,s) - P_{ik}(s,s) - o(h)}{-h} = -\left[\frac{\partial}{\partial s}P_{ik}(s,t)\right]_{t=s}$$

[2]

[4]

ii) Kolmogorov's forward differential equation

The matrix form of the forward differential equation is:

$\partial/\partial t\ P(s,t) = -\ P(s,t)\ A(t)$

Since this model is time inhomogeneous and we are asked for the forward differential equation, we are differentiating with respect to t.

For this model:

$\partial/\partial t P_{PS}(s,t) = -[P_{PP}(s,t)\ 0.15t + P_{PA}(s,t)\ 0.1t - P_{PS}(s,t) 0.01t]$ [2]

and:

$\partial/\partial t\ P_{SS}(s,t) = -[P_{SS}(s,t)\ (-0.01t)] = 0.01t\ P_{SS}(s,t)$ ……………………(1)

[2]

[4]

iii) From equation 1 above,
Separating the variables gives:

$\partial/\partial t\ P_{SS}(s,t)\ /\ P_{SS}(s,t) = 0.01t$

and changing the variables from t to u:

$\partial/\partial u \ln P_{SS}(s,u) = 0.01t$

Integrating both sides with respect to u between the limits u = s and u = t, we get:

$[\ln P_{SS}(s,u)]^{t}_{s}$ = integrate (t,s) 0.01u du = $[0.005u^2]^{t}_{s}$
i.e. $\ln P_{SS}(s,s) - \ln P_{SS}(s,t) = 0.005\ (t^2 - s^2)$

However, since $P_{SS}(s,s) = 1$ and ln 1 = 0, we have,
$-\ln P_{SS}(s,t) = 0.005\ (t^2 - s^2)$

The expression above can be rearranged to give:
$P_{SS}(s,t) = e^{-0.005\ (t2 - s2)}$

Now, substituting the value of t = 5 and s = 3, we get:
$P_{SS}(s,t) = e^{-0.08} = 92.31\%$

[2]

[10 Marks]

Solution 6:

i) Let N_1, N_2, N_3 & N_4 denote the number of tax payers of States East, North, South & West respectively and X_1, X_2, X_3 & X_4 are the tax amount paid by the tax payers. Then T can be modelled using Compound Poisson random variable.

N_1~ Poisson $(25000*\frac{1}{25})$ ≡ Poisson (1000) and the $E(X_1)$ = 100 Lacs Oceania \$ and $E(X_1^2)=100^2$ Lacs Oceania \$ [0.5]

N_2, N_3 follow Poisson $(25000*\frac{1}{5})$ ≡ Poisson (5000) and $E(X_2) = E(X_3) = (10*0.5+20*0.5)=15$ Lakhs Oceania \$ and
$E(X_2^2) = E(X_3^2) = (10^2*0.5+20^2*0.5)=250$ Lacs Oceania \$.

N_4 follow Poisson (5000) and $E(X_4) = (10/2)=5.5$ Lakhs Oceania \$ and $E(X_4^2) = (11*21/6)=38.5$ Lakhs Oceania \$

[1]

Then, using the assumption that the lives are independent and the results that, for a compound Poisson random variable T, the k th central moment of T is given by $\lambda E(X^k)$, we obtain

E(T) = 1000*100 + 5000 * 15*2 + 5000 * 5.5

= 277500 Lakhs Oceania \$ [1]

$V(T) = 1000*100^2 + 5000*250*2+5000*38.5$

= 12692500 Lakhs Oceania \$ [1.5]

[4]

ii) Assuming that T ~ N(277500, 12692500), We have
P(T > 2000) = P { N(2775, 126925) > 2000 }

$= P\{ N(0,1) > \frac{2000-2775}{\sqrt{126925}} \}$

= 1- Φ(-2.175)

= 99% [2]

iii) $N_1 \sim$ Poisson $(25000*\frac{1}{25}) \equiv$ Poisson (1000) and the $E(X_1)$ = One Crore Oceania $

N_2, N_3 follow Poisson $(25000*\frac{1}{5}) \equiv$ Poisson (5000) and $E(X_2) = E(X_3)$ = 14 Lakhs Oceania $ and $E(X_2^2) = E(X_3^2)$ = 2.32 Crores Oceania $.

N_4 follow Poisson (5000) and $E(X_4)$ = 4.4 Lakhs Oceania $ and $E(X_4^2)$ = 30.8 Lakhs Oceania $ [1]
Then

Revised E(T) = 1000*100 + 5000 * 14*2 + 5000 * 4.4

= 262000 Lakhs Oceania $ [1]

Revised V(T) = $1000*100^2 + 5000*232*2+5000*30.8$

= 12474000 Lakhs Oceania $ [1.5]

Assuming that T ~ N(2620, 124740), We have
P(T > 2000) = P { N(2620, 124740) > 2000 } [0.5]

$= P\{ N(0,1) > \frac{2000-2620}{\sqrt{124740}} \}$

$= 1 - \Phi(-1.76)$

= 96% [2]

[6]

[12 Marks]

Solution 7:

Given that ${}_{30}P_{50} = 0.6$ for Male and ${}_{30}P_{50} = 0.65$ for Female.

⇨ The probability of death for the same period is
${}_{30}q_{50} = 0.4$ for Male and ${}_{30}q_{50} = 0.35$ for Female

Let X be the random variable representing the future lifetime of Male and Y be the random variable representing the future lifetime of Female. Then

$P(X \le 30) = 0.4$ and $P(Y \le 30) = 0.35$.

What we require is $P(X \le 30, Y \le 30)$.

i) The Gumbel Copula with α = 3.5

u = 0.4 and v = 0.35

u = 0.4 and v = 0.35

$$C[u, v] = \exp\left\{ - \left((-\ln u)^\alpha + (-\ln v)^\alpha \right)^{(1/\alpha)} \right\}$$

Applying the values of α, u &v , C [u, v] = 0.2996 [2]

ii) The Clayton Copula with α = 3.5

$C[u, v] = (u^{-\alpha} + v^{-\alpha} - 1)^{(-1/\alpha)}$

Applying the values of α, u &v , C [u, v] = 0.30595 [1]

iii) The Frank copula with α = 3.5

$$C[u, v] = (-1/\alpha) \ln \left\{ 1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{(e^{-\alpha}-1)} \right\}$$

Applying the values of α, u &v , C [u, v] = 0.2273 [1]

iv) Clayton Copulas gives the highest probability of claims because it exhibits lower tail dependence. This means that if one life does not survive for long then there is high probability that the other life will also not survive for long.

If the deaths are independent then the probability of paying the benefit is 0.14.

However, two lives covered are related, so we would expect the probability of paying the benefit to be higher than under the assumption of independence. Hence Clayton copula is more appropriate.

[2]

[6 Marks]

<u>Solution 8:</u>

i) The hazard function for getting married is given by:
$\lambda(t,Z) = \lambda_0(t)\exp[0.3 Z_1 + 0.2 Z_2 + 0.3 Z_3 + 0.5 Z_4 - 0.1 Z_5 + 0.7 Z_6 + 0.5 Z_7 - 0.4 Z_8]$

Where
$\lambda_0(t)$ = baseline hazard at time t since looking for the life partner.

$Z = (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8)$
Z_1 = 1 if female, 0 if not.
Z_2 = 1 if location = Non Metro, 0 if not
Z_3 = 1 if profession = Service, 0 if not
Z_4 = 1 if profession = Business, 0 if not
Z_5 = 1 if profession = Social Service, 0 if not
Z_6 = 1 if Age Band = 20-25, 0 if not
Z_7 = 1 if Age Band = 25-30, 0 if not.
Z_8 = 1 if Age Band = 35-40, 0 if not. [2]

ii) People most likely to stay single with the lowest hazard function.

The probability that a person who has been looking for a life partner for one year will stay single for next 2 years is:

$\exp(-\int_1^3 \lambda(t,Z)\, dt)$ [1]

If the person is a female, profession as a social service and aged 37, the probability is:
$P_F = \exp[-e^{0.3Z_1 - 0.1Z_5 - 0.4Z_8} \int_1^3 \lambda_0(t)dt]$
$P_F = \exp[e^{-0.2} \int_1^3 \lambda_0(t)dt]$,
Let $A = \exp[\int_1^3 \lambda_0(t)dt]$

$P_F = Ae^{-0.2} = 0.3$
$A = 0.2298$ [1]

If the person is a male, working as a businessman and aged 24, the probability is:
$P_M = \exp[-e^{0.2Z_2+0.5Z_4+0.7Z_6}$ integral$|_1^3\lambda_0(t)dt]$
Let A = exp [integral$|_1^3\lambda_0(t)dt$]
$P_M = A\,e^{1.4} = 0.00257$ [1]

[3]

iii) In graduation there is tradeoff between smoothness & adherence to data. A satisfactory graduation must provide a good balance between the two.

'Under Graduation' occurs when too much emphasis is given to goodness of fit. The 'under graduated' rates adhere more closely to crude rates at the cost of smoothness.

'Over Graduation' is the reverse when too much emphasis is given to smoothness at the cost of adherence to data (crude rates).

[2]

iv) **Chi Square Test**

The null hypothesis is:
H_0: the graduated rates are the true underlying rates of getting married for the population.
We calculate the individual standardised deviations at each age using the formula:
Zx = (Qx – ExUx) / (ExUx)^(1/2)

The ISDs are:
0.218, -1.054, -1.647, 0.636, -1.226, -0.121, -0.015, -0.252, -1.606, 1.062, -0.930, -0.475, -0.558, 2.121, 0.068, -0.343, -0.447, 2.469, 1.262, 0.477, -0.149

The test statistic for the chi-squared test (based on unrounded Zx values) is:
Sum of $Z_x^2 = 23.72$

We now compare this with χ^2 distribution. We were given data from 21 ages. Since the graduation was carried out using a formula with 4 parameters, we lose 4 degrees of freedom. So we are left with 17 degrees of freedom.

From the tables, we see that the upper 5% point of χ^2_{17} is 27.59. As the value of the test statistic is lower than this, we do not reject the null hypothesis and conclude that the graduated rates do provide a good fit to the data.

[4]

Signs Test

This is a simple two tailed test for overall bias.

There should be roughly equal numbers of positive and negative ISDs. Under the null hypothesis, the number of positive deviations has a *Binomial*(21,0.5) distribution.

We have 8 positives and 13 negatives, which is fine.

So we do not reject the null hypothesis and we conclude that there is no evidence of overall bias in the graduated rates. [2]

[6]

[13 Marks]

Solution 9:

a) Uniform Distribution of Deaths

If deaths are uniformly distributed between the ages of x and y, then the number of lives in the population decreases linearly between the ages of x and y.
So the monthly mortality assumption shall be 24/12 = 2 per 1000. [1.5]

The survival function is linearly decreasing function of t. [0.5]

b) Constant Force of Mortality

This assumption says that μ_{x+t} is equal to some constant μ for all t between 0 and y-x.
In general:

${}_tp_x$ = exp (-integration$|_0^t \mu_{x+s} ds$)
Under the constant force assumption, this simplifies to:
${}_tp_x = e^{-t\mu}$
${}_tp_x$ = exp(-2/1000) = 0.998002. [1.5]

The survival function is an exponentially decreasing function of t. [0.5]

[4 Marks]

Solution 10:

i) In terms of the backwards shift operator we have
$(1 + 2\alpha B - \alpha^2 B^2)Y = Z.$
We must find the values of α such that the roots of the polynomial
$1 + 2\alpha x - \alpha^2 x^2$ lie outside the unit circle.

$The\ roots\ are\ \frac{1}{\alpha}(1 \pm \sqrt{2}), so\ we\ require\ \frac{(\sqrt{2}+1)}{\alpha} > 1\ and\ \frac{(\sqrt{2}-1)}{\alpha} > 1, in\ other\ words$
$\alpha < \sqrt{2} - 1.$

[2]

ii) $Y_t = -2\alpha Y_{t-1} + \alpha^2 Y_{t-2} + Z_t$

$$COV(Y_t, Y_t) = \gamma_0 = -2\alpha\,\gamma_1 + \alpha^2\gamma_2 + \sigma^2$$
$$COV(Y_t, Y_{t-1}) = \gamma_1 = -2\alpha\,\gamma_0 + \alpha^2\gamma_1$$
$$COV(Y_t, Y_{t-2}) = \gamma_2 = -2\alpha\,\gamma_1 + \alpha^2\gamma_0$$

$$From\ above; = \gamma_1 = \frac{-2\alpha\,\gamma_0}{1-\alpha^2}$$

[2]

Solving above equations

$$\gamma_0 = \frac{\sigma^2(1-\alpha^2)}{(1+\alpha^2)(1-6\alpha^2+\alpha^4)}$$
$$\gamma_1 = \frac{-2\alpha\sigma^2}{(1+\alpha^2)(1-6\alpha^2+\alpha^4)}$$
$$\gamma_2 = \frac{\sigma^2(5\alpha^2-\alpha^4)}{(1+\alpha^2)(1-6\alpha^2+\alpha^4)}$$

[3]

iii) Exponential smoothing is simple to apply and does not suffer from problems of over-fitting. If the data appear fairly stationary but are not especially well fitted by any of the Box-Jenkins methods, exponential smoothing is likely to produce more reliable results. More advanced versions of exponential smoothing can cope with varying trends and multiplicative variation. [2]

[9 Marks]

Solution 11:

i) $\gamma_k = Cov(X_t, X_{t-k}) \quad = \frac{1}{(m+1)^2} Cov(\sum_{r=0}^{m} e_{t-r}, \sum_{r=0}^{m} e_{t-k-r})$

$$= \frac{1}{(m+1)^2} \sum_{r=0}^{m} \sum_{r=0}^{m} Cov(e_{t-r,} e_{t-k-r})$$

Clearly if k > m, all terms are zero, so that $\gamma_k = 0$.

$For\ 0 \le k \le m, there\ are\ exactly\ (m-k+1)\ non-zero\ terms, and\ each\ of\ these\ covariance\ terms\ equals\ \sigma_e^2. Thus$

$$\gamma_k = \begin{cases} 0 & k > m \\ \dfrac{m-k+1}{(m+1)^2}\sigma_e^2 & 0 \le k \le m \end{cases}$$

The autocorrelation function is

$$\rho_k = \begin{cases} 1 & k > 0 \\ \dfrac{m-k+1}{(m+1)} & 0 \le k \le m \\ 0 & k > m \end{cases}$$

[4]

ii) For the process to be invertible, we require that roots of the characteristic equation should be greater than 1 in absolute value.
We can rewrite the MA model with the aid of the backward shift operator B as follows:

$$X_t - \mu = \frac{1}{3}(1 + B + B^2)e_t$$

$Root\ of\ chracteristics\ equations\ \ (1 + B + B^2) = 0\ are$

$B = -\frac{1}{2} + \frac{1}{2}i\sqrt{3}\ and\ B = -\frac{1}{2} - \frac{1}{2}i\sqrt{3}$

In both cases $|B|$ = 1. Thus, the process in not invertible. [2]

iii) The original data are clearly subject to seasonal variation, and the size of the seasonal fluctuations is increasing in line with the value of the underlying quantity. This suggests that the seasonal variation is multiplicative rather than additive, in which case taking the logarithm is the sensible thing to do. In addition to this, a look at the plot of y_t against time confirms that the variation is much more regular.

[2]

iv)

i) AR(2) [1]

ii) MA(1) [1]

iii) AR(1) [1]

[3]

[11 Marks]

Solution 12:

i) If each room is represented by the state, then the transition matrix P for this Markov chain is as follows:

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$ [2]

ii) The chain is irreducible, because it is possible to go from any state to any other state. However, it is not aperiodic, because for any even n ,$P^n_{6,1}$ will be zero and for any odd n $P^n_{6,5}$ will also be zero . This means that there is no power of P that would have all its entries strictly positive.

[2]

iii) For P to be stationary,

πP = P

Perform matrix multiplication and show that π P is equal to P [3]

iv) We find from π that the mean recurrence time (i.e. the expected time to return) for the room 1 is 1/π(1)=12 [1]

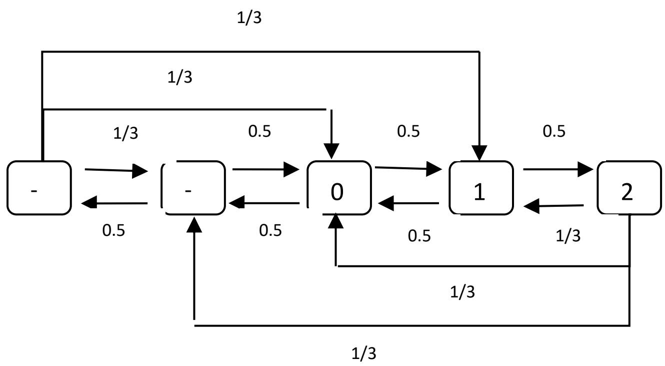
v) Let, ψ(i) = E(number of steps to reach state 5 | X_0 = i).
We have
ψ(5) = 0
ψ(6) = 1 + (1/2)ψ(5) + (1/2)ψ(4)
ψ(4) = 1 + (1/2)ψ(6) + (1/2)ψ(3)
ψ(3) = 1 + (1/4)ψ(1) + (1/4)ψ(2) + (1/4)ψ(4) + (1/4)ψ(5)
ψ(1) = 1 + ψ(3)
ψ(2) = 1 + ψ(3). [1.5]
We solve and find ψ(1) = 7. [1.5]

[3]

[11 Marks]

Solution 13:

i) Transition Diagram



[1.5]

Transition Matrix

	-2	-1	0	1	2
-2	0	1/3	1/3	1/3	0
-1	½	0	1/2	0	0
0	0	1/2	0	1/2	0
1	0	0	1/2	0	½
2	0	1/3	1/3	1/3	0

[1.5]
[3]

ii) $[\pi_{-2}, \pi_{-1}, \pi, \pi_1, \pi_2] = [\pi_{-2}, \pi_{-1}, \pi, \pi_1, \pi_2]$ P

From the symmetry $\pi_{-2} = \pi_2$, $\pi_{-1} = \pi_1$, we have [0.5]

$\pi_0 = 1/3\pi_2 + ½\, \pi_1 + ½\, \pi_1 + 1/3\pi_2$
$\pi_1 = 1/3\pi_2 + ½\, \pi_0 + 1/3\, \pi_2$
$\pi_2 = 1/2\pi_1$
$\pi_0 + 2\pi_1 + 2\pi_2 = 1$ [1]

From there we can get $\pi_0 = 4/3\, \pi_1$, and $\pi_1 = 3/13$ so
π = [3/26, 3/13, 4/13, 3/13, 3/26] [1.5]
[3]

iii) This chain is irreducible since it is possible to move from each state to any other.

A periodic chain is one in which a state can only be revisited at multiples if some fixed number d>1. State "-2" is aperiodic as it can be revisited after any number of steps. Also, since this chain is irreducible, all the states have the same periodicity. So the chain is aperiodic.

[2]
[8 Marks]

****