INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

November 2026

Subject CS1 – Actuarial Statistics (Paper A)

Time allowed: 3 Hours 15 Minutes

Total Marks: 100

Q. 1) A regional railway operator has installed automatic ticket-vending machines at a suburban station. Each time a passenger uses a machine, the transaction fails due to a card-reader fault with probability 0.08, independently each time.

Over one morning, 15 transactions are attempted at a machine. What is the probability that at the most one of these transactions fails?

A. 0.482
B. 0.660
C. 0.771
D. 0.286
E. 0.374. **[2]**

Q. 2) A food-delivery platform models the number of orders received by a restaurant during a one-hour period using generalized linear models with a log link function. Following two models have been fitted:

Model 1: $\log(\mu) = 1.0 + 0.04x_1 - 0.15x_2$

Model 2: $\log(\mu) = 0.8 + 0.05x_1 - 0.10x_2 + 0.02\, x_1\, x_2$

where:
- x_1 is the number of active delivery riders available during the hour and
- $x_2 = 1$ if the restaurant is located outside the city centre, and $x_2 = 0$ otherwise

Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) have been calculated for both the models to test the model fitting. They are as follows:

Model	AIC	BIC
Model 1	512.40	525.10
Model 2	510.80	527.60

Which of the following statements is correct?

A. Model 1 is preferred by both AIC as well as BIC
B. Model 2 is preferred by both AIC as well as BIC
C. Model 1 is preferred by BIC whereas Model 2 is preferred by AIC
D. Model 2 is preferred by BIC whereas Model 1 is preferred by AIC
E. None of the above. **[2]**

Q. 3) Ravi's rooftop solar inverter typically runs for an average of 500 hours before a fault occurs.

It has already been running without any fault for 200 hours. What is the probability that it will continue running for at least another 100 hours without a fault?

A. 0.819
B. 0.181
C. 0.368
D. 0.905
E. 0.632. **[2]**

Q. 4) A music-streaming company examines whether the number of songs added to a user's personal playlist is related to the user's monthly listening time. A simple linear regression model is fitted based on a total of 100 pairs of observations.

For this model,

- Total Sum of Squares (SS_{TOT}) = 1,250 and
- Residual Sum of Squares (SS_{RES})= 375.

Which of the following is correct?

A. $R^2 = 0.70$ and Adjusted $R^2 \leq R^2$
B. $R^2 = 0.70$ and Adjusted $R^2 \geq R^2$
C. $R^2 = 0.30$ and Adjusted $R^2 = R^2$
D. $R^2 = 0.70$ and Adjusted $R^2 < R^2$
E. $R^2 = 0.68$ and Adjusted $R^2 > R^2$. **[2]**

Q. 5) A wellness platform records the duration of guided meditation sessions. The duration of a session is modelled by an exponential distribution with mean 15 minutes. A random sample of 100 sessions is selected.

Using central limit theorem, what is the approximate probability that the average session duration is less than 14 minutes?

A. 0.748
B. 0.500
C. 0.409
D. 0.391
E. 0.252. **[2]**

Q. 6) An insurer is reviewing the time taken to settle travel insurance claims.

For a random sample of 25 claims from a population where the population standard deviation is known, the upper bound of a two-sided 95% confidence interval for the claim settlement time is found to be twice that of the lower bound.

Which of the following statements is correct?

A. Population standard deviation σ is equal to the sample mean $\overline{X}$
B. Population standard deviation σ is lower than the sample mean $\overline{X}$
C. Population standard deviation σ is higher than the sample mean $\overline{X}$
D. Population standard deviation σ is equal to zero
E. Sample mean $\overline{X}$ is equal to zero. **[2]**

Q. 7) In a stock exchange, following variables are required to be predicted using generalised linear models:

X: represents percentage return of a stock which can take positive and negative values
Y: represents the probability that a stock will have positive return on a given trading day
Z: represents the number of times a circuit breaker is breached on a given trading day
W: represents stock market volatility where large values of the linear predictor are associated with smaller fitted values of W

The stock exchange wants to choose appropriate link functions for modelling each of the above variables.

Which of the following pairs are correctly matched?

	X	Y	Z	W
A.	$\mu = \eta$	$\ln(\mu / (1-\mu)) = \eta$	$\ln(\mu) = \eta$	$1 / \mu = \eta$
B.	$\ln(\mu) = \eta$	$\mu = \eta$	$1 / \mu = \eta$	$\ln(\mu / (1-\mu)) = \eta$
C.	$\mu = \eta$	$\ln(\mu) = \eta$	$\ln(\mu / (1-\mu)) = \eta$	$1 / \mu = \eta$
D.	$1 / \mu = \eta$	$\ln(\mu / (1-\mu)) = \eta$	$\mu = \eta$	$\ln(\mu) = \eta$
E.	$\ln(\mu / (1-\mu)) = \eta$	$\mu = \eta$	$\ln(\mu) = \eta$	$1 / \mu = \eta$

[2]

Q. 8) A municipal corporation is testing whether a redesigned grievance-triage process reduces the proportion of citizen complaints that remain unresolved on its public portal.

The corporation tests the following hypotheses:

H_0: the redesigned process does not reduce the proportion of unresolved complaints,

v.

H_1: the redesigned process reduces the proportion.

Which of the following statements is correct?

A. Type I error occurs when H_0 is not rejected when the redesigned process does not really reduce the proportion of unresolved complaints.
B. Type II error occurs when H_0 is rejected when the redesigned process really reduces the proportion of unresolved complaints.
C. Type I error occurs when H_0 is rejected when the redesigned process does not really reduce the proportion of unresolved complaints.
D. Type II error occurs when H_0 is not rejected when the redesigned process does not really reduce the proportion of unresolved complaints.
E. Both options (C) and (D) are correct. **[2]**

Q. 9) A consumer research company is comparing the variability of customer satisfaction scores reported for two different online retail platforms. The scores for each platform are assumed to follow normal distributions.

A random sample of 10 customers from Platform A gives a sample variance of 25, while a random sample of 13 customers from Platform B gives a sample variance of 10.

For an F-test comparing the two population variances, with the larger sample variance placed in the numerator, what is the value of the test statistic and its degrees of freedom?

A. F-statistic = 0.4 over (12, 9) degrees of freedom
B. F-statistic = 2.5 over (10, 13) degrees of freedom
C. F-statistic = 0.4 over (9, 12) degrees of freedom
D. F-statistic = 2.5 over (12, 9) degrees of freedom
E. F-statistic = 2.5 over (9, 12) degrees of freedom. **[2]**

Q. 10) A streaming platform uses the following multiple linear regression model to predict the number of minutes, Y, that a user spends watching content in a week:
$Y = 18 + 0.6X_1 - 2.5X_2 + 0.04X_3$
where:

- X_1 = number of personalised recommendations shown to the user,
- X_2 = age of the user's account, in years,
- X_3 = number of days during the week on which the user logged in.

Two users, A and B, have the same values of X_2 and X_3. User A is shown 10 more personalised recommendations than User B. Based on the fitted model, how many minutes

in a week will User A spend more than User B for watching content on the streaming platform?

A. 18
B. 9
C. 3
D. 6
E. None of the above. **[2]**

Q. 11) A researcher is studying the daily proportion of Ganesh mandals in the city which have breached the judicial directives relating to use of loudspeakers beyond 55 decibels during the ten-day Ganesh festival celebrations.

For the ten days of the festival, following daily proportions were observed:

(0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.70)

The researcher decides to use non-parametric bootstrap sampling to assess the sampling variability of the estimated daily proportion. Which of the following are valid bootstrap samples?

A. (0.20, 0.30, 0.30, 0.40, 0.50, 0.55, 0.55, 0.60, 0.25, 0.70)
B. (0.25, 0.25, 0.35, 0.40, 0.45, 0.45, 0.50, 0.60, 0.60, 0.70)
C. (0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.70)
D. (0.25, 0.30, 0.35, 0.35, 0.40, 0.45, 0.50, 0.50, 0.60, 0.70)
E. All of the above. **[2]**

Q. 12) A historian is studying the duration, in years, of ancient dynasties in a particular region.

Suppose the duration (X) of a dynasty is modelled by a Uniform (0, θ) distribution, where $\theta > 0$.

For a random sample of five dynasties, the recorded durations in years are:

1, 3, 7, 9, 10

What is the method of moments estimate of θ based on the first moment?

A. 6
B. 8
C. 12
D. 18
E. 24. **[2]**

Q. 13) A marital counselling application records a "Relationship Strength Score" (RSS) for married couples. A generalised linear model is fitted using the following four explanatory variables:

- A: Age group: three categories (young, middle-aged, older)
- D: Duration of marriage in years
- S: Sessions frequency measured as number of counselling sessions per month
- C: Communication style (either direct or indirect)

Considering young couples with indirect communication style as the baseline, following two models have been proposed:

Model 1: RSS ~ A + D + S + C

Model 2: RSS ~ A + D + S + C + D * C + S * C

How many parameters are to be estimated for each of these models?

A. Model 1 – 5 parameters, Model 2 – 7 parameters
B. Model 1 – 6 parameters, Model 2 – 8 parameters
C. Model 1 – 6 parameters, Model 2 – 7 parameters
D. Model 1 – 7 parameters, Model 2 – 9 parameters
E. Model 1 – 5 parameters, Model 2 – 8 parameters. **[2]**

Q. 14) An insurance company is comparing the frequency of motor claims for two groups of policyholders over one year. For drivers using a telematics-based policy, 90 claims were reported over 3,000 policy-years of exposure. For drivers using a standard policy, 64 claims were reported over 2,000 policy-years of exposure.

Assuming the two claim counts are independent Poisson variables with parameters λ_T and λ_S respectively, what is an approximate 95% two-sided confidence interval for the difference in claim rates, $(\lambda_T - \lambda_S)$? Use normal approximation.

A. (−0.0120, 0.0080)
B. (−0.0100, 0.0060)
C. (−0.0020, 0.0100)
D. (−0.0200, 0.0160)
E. (0.0020, 0.0120). **[2]**

Q. 15) A city public-works department is estimating the average number of pothole reports received per week for a district's road network. The number of pothole reports is modelled using a Poisson distribution with mean parameter λ. Before observing new data, the department uses a Gamma (4, 2) prior for λ, where the second parameter is the rate parameter. Data from 10 districts, each observed over one week, give a total of 55 pothole reports.

Which of the following options represents the correct posterior distribution?

A. Gamma (4, 57)
B. Gamma (55, 10)
C. Gamma (14, 59)
D. Gamma (59, 12)
E. Gamma (4, 2). **[2]**

Q. 16) Air pollution is a major environmental and public-health concern in many large cities. In particular, nitrogen dioxide (NO_2) concentrations are often higher near busy roads because of emissions from road traffic.

The government of a large city has introduced several measures to reduce roadside NO_2 concentrations. Three of the main measures are:

- **Strategy A:** optimisation of traffic-signal timings to reduce vehicle idling at major junctions.
- **Strategy B:** introduction of a low-emission zone restricting the most polluting vehicles from entering selected central areas.
- **Strategy C:** expansion of dedicated bus services to encourage commuters to use public transport.

After six months, monitoring stations operating under each strategy classify the change in peak-hour NO_2 concentration as either large reduction, moderate reduction, or small/no reduction.

Government strategy	Large reduction	Moderate reduction	Small/no reduction	Total
Strategy A	8	10	6	24
Strategy B	14	7	3	24
Strategy C	9	9	6	24

The government wishes to determine whether the observed effectiveness of the three strategies differs.

i) State the null and alternative hypotheses for a χ^2 test of independence between the government strategy and the category of reduction in NO_2 concentration. **[1]**

ii) Calculate the expected frequency for each cell of the contingency table. **[2]**

iii) Perform the χ^2 test at the 5% level of significance and determine whether there is evidence of an association between the government strategy and the category of reduction in NO_2 concentration. **[3]**

A separate study was carried out at 20 monitoring stations using Strategies A and B. The reduction in peak-hour NO_2 concentration, measured in $\mu g/m^3$, was recorded for each station. The summary statistics were:

Government strategy	Number of stations	Mean reduction	Standard deviation
Strategy A	20	4.8	2.4
Strategy B	20	6.1	2.7

Assume that the two samples are independent and that the population variances may be regarded as equal. A 95% confidence interval for the difference in mean reduction, $\mu_A - \mu_B$, is required.

iv) Construct the 95% confidence interval for the difference between the mean NO_2 reductions under Strategies A and B. **[3]**

v) Using the confidence interval determined in part (iv) together with your conclusion from part (iii), comment on whether the government has statistical evidence to favour Strategy B over Strategy A as it shows a higher mean reduction by 1.3 $\mu g/m^3$. **[1]**

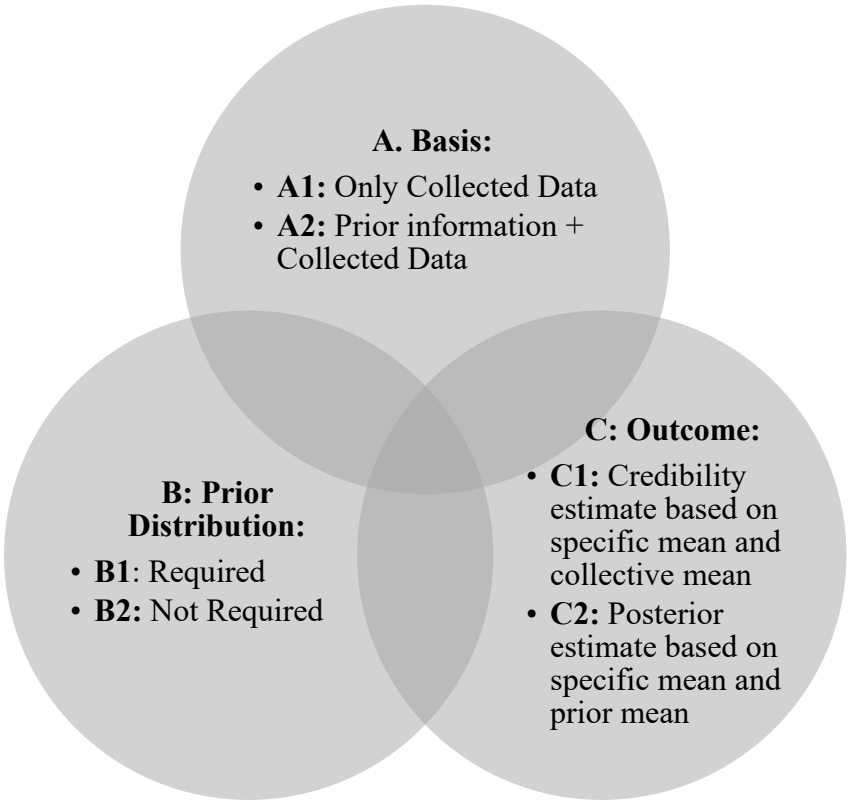
[10 Marks]

Q. 17) A general insurance company provides home insurance to customers in several regions. The insurer monitors the number of claims arising from policies written in each region. Since the number of policies written varies considerably between regions and from year to year,

the insurer wants to allow for differences in the volume of business when estimating future claim frequencies.
For each region and year, let:

- Y_{ij} = number of claims arising in region i during year j,
- P_{ij} = number of policies in force in region i during year j, and
- X_{ij} = observed claim frequency = Y_{ij} / P_{ij}.

First, it is decided to compare empirical credibility theory with Bayesian credibility theory. Following infographics can be constructed for the same.



i) Using the above infographics, complete the table clearly indicating "Yes" or "No":

Particulars	A1	A2	B1	B2	C1	C2
Empirical Credibility	Yes	____	____	Yes	____	____
Bayesian Credibility	No	____	____	No	____	____

[2]

The insurer has collected the following data for three regions over four years.

	Year 1	Year 2	Year 3	Year 4
Region A — claims	32	46	36	46
Region A — policies ('000)	10	10	10	10
Region B — claims	84	116	92	108
Region B — policies ('000)	20	20	20	20
Region C — claims	156	204	168	192
Region C — policies ('000)	30	30	30	30

In the absence of an appropriate distributional assumption, it has been decided that empirical credibility is more suitable for the purpose of estimation. Empirical Bayes Credibility Theory Model 2 (EBCT Model 2) would be used by the insurer to estimate expected number of claims.

For the purposes of applying EBCT Model 2, following R codes were written by two student actuaries Aman and Nikhil who are preparing for CS1.

R Code written by Aman:

```
claims <- data.frame(
  Year1 = c(32, 84, 156),
  Year2 = c(46, 116, 204),
  Year3 = c(36, 92, 168),
  Year4 = c(46, 108, 192)
)

policies <- data.frame(
  Year1 = c(10, 20, 30),
  Year2 = c(10, 20, 30),
  Year3 = c(10, 20, 30),
  Year4 = c(10, 20, 30)
)

n <- ncol(claims)

N <- nrow(claims)

X <- claims / policies

Xibar <- rowSums(claims) / rowSums(policies)

Pibar <- rowSums(policies)

Pbar <- sum(Pibar)

Pstar <- sum(Pibar * (1 - Pibar / Pbar)) / (N * n - 1)

m <- sum(claims) / Pbar

s <- mean(rowSums(policies * (X - Xibar)^2) / (n - 1))

v <- (sum(rowSums(policies * (X - m)^2)) / (n * N - 1) - s) / Pstar
```

R Code written by Nikhil:

```
claims <- data.frame(
  RegionA = c(32, 46, 36, 46),
  RegionB = c(84, 116, 92, 108),
  RegionC = c(156, 204, 168, 192)
)

policies <- data.frame(
  RegionA = c(10, 10, 10, 10),
  RegionB = c(20, 20, 20, 20),
  RegionC = c(30, 30, 30, 30)
```

```
)
n <- ncol(claims)
N <- nrow(claims)
X <- claims / policies
Xibar <- rowSums(claims) / rowSums(policies)
Pibar <- rowSums(policies)
Pbar <- sum(Pibar)
Pstar <- sum(Pibar * (1 - Pibar / Pbar)) / (N * n - 1)
m <- sum(claims) / Pbar
s <- mean(rowSums(claims * (X - Xibar)^2) / (n - 1))
v <- (sum(rowSums(claims * (X - m)^2)) / (n * N - 1) - s) / Pstar
```

Following quantities were estimated by Aman and Nikhil based on the R codes:

	E[m(θ)]	**E[s²(θ)]**	**Var[m(θ)]**
Aman	5.3333	10.5778	0.7648
Nikhil	5.3333	54.3741	1.9043

ii) Referring to the appropriate formulae from Actuarial Formulae and Tables, comment with reasons which of the two codes is correct. Hence determine the correct estimates for E[m(θ)], E[s²(θ)] and Var[m(θ)] under assumptions of EBCT Model 2. **[3]**

iii) Calculate the credibility factor for Region B and hence obtain the credibility estimate of its claim frequency for the coming year. Use this estimate to calculate the expected number of claims for Region B, assuming that 22,000 policies will be in force. **[3]**

iv) What would be the impact on the credibility factor for Region B in the following independent situations? You are NOT required to perform any additional calculations but provide qualitative justifications in support of your answer.

- Number of in-force policies across all regions are doubled (assume E[s²(θ)] / Var[m(θ)] is also double the original value).

- Expected variance within each region is expected to increase leading to increase in E[s²(θ)].

[2]
[10 Marks]

Q. 18) A public health analytics company was investigating whether neighbourhoods with higher levels of daily physical activity also tend to have lower average resting heart rates.

Information was collected from nine neighbourhoods.

For each neighbourhood,

- X is the average number of minutes of moderate-to-vigorous physical activity per person per day, and
- Y is the average resting heart rate, measured in beats per minute.

The CEO of the company has asked the project team to present the correlation between the two quantities in a week's time. However, the original data files have been corrupt and cannot be retrieved. Based on the working folder, following summary quantities are available:

$\Sigma x_i = 315$,
$\Sigma y_i = 615$,
$\Sigma x_i^2 = 12{,}525$,
$\Sigma y_i^2 = 42{,}515$,

A simple linear regression line fitted to the data as per the working folder is as follows:

$$Y = 87.70 - 0.5533 * X$$

i) Calculate the Karl Pearson's coefficient of correlation between daily physical activity and resting heart rate based on the available information.

Hint: [$r = S_{xy} / sqrt\ (S_{xx} * S_{yy}) = S_{xy} / S_{xx} * sqrt\ (S_{xx} / S_{yy})$.] **[3]**

Based on the notepad of one of the team leads working on the project, ranks for the nine neighbourhoods according to their values of X and Y are available. However, the team lead is of the view that this is an older version and may not be treated as final.

The ranks as per the notepad are as follows:

Neighbourhood	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Rank of X (R_X):	1	3	2	5	4	7	6	9	8
Rank of Y (R_Y):	9	7	8	5	6	3	4	1	2

ii) Calculate Spearman Rank's coefficient of correlation between daily physical activity and resting heart rate using the ranks as per the notepad. **[3]**

The project manager has also asked the team to present the sensitivity of the Karl Pearson's correlation coefficient numbers to changes in the values of the underlying variables.

Following independent situations are being considered for this purpose:

- Average resting heart rate is considered to be half of the observed values
- Average time spent on physical activity is considered to be twice the observed values
- Instead of considering average total time spent on physical activity, average time spent on physical activity in excess of 1 hour is being considered and compared with average resting heart rate
- Instead of considering average total time spent on physical activity, average time in a day which is not spent on physical activity is being considered and compared with average resting heart rate

iii) Without performing any additional calculations, quantify the effect of each of the above independent situations on the value of Karl Pearson's correlation coefficient. Justify your answers. **[4]**

[10 Marks]

Q. 19) A financial reporting actuary working for a multinational corporation models the annual outcome (profit or loss), X, in crores arising from a new product launch. X is a normally distributed variable with $\mu = \sigma^2 = 1$ and the probability density function of X is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-1)^2}{2}\right], -\infty \leq x \leq \infty.$$

She wishes to investigate the distributional properties of annual outcomes using cumulant generating functions.

i) State the cumulant generating function $C_x(t)$. **[2]**

ii) Using the cumulant generating function from part (i), calculate the mean and variance of X. **[2]**

It is expected that there will be three product launches during the year with the distribution of their outcomes corresponding to random variable X.

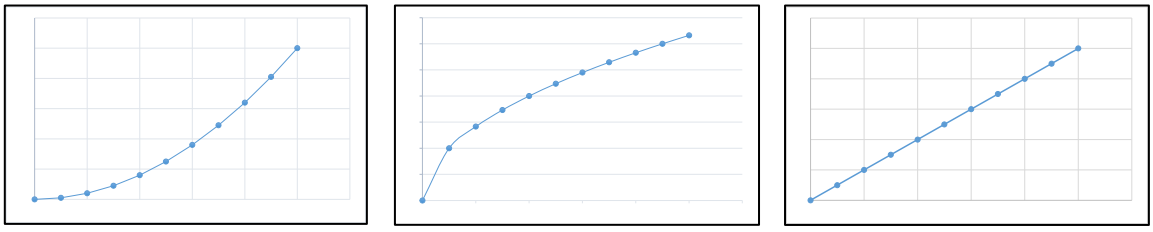
Let W represent the total outcome (profit / loss) from all three product launches during the year.

iii) Using cumulant generating functions, show that the distribution of W is symmetrical about the mean.

Hint: [*A distribution is said to be symmetrical about the mean when its skewness is equal to zero.*] **[3]**

iv) Which of the following graphs correctly represents the cumulant generating function of W i.e. $C_W(t)$? Justify your answer.

Graph A (Convex Curve)	**Graph B** (Concave Curve)	**Graph C** (Straight Line)



Hint: [*Slope of the $C_W(t)$ curve is represented by the first derivative of $C_W(t)$ i.e. $C'_W(t)$ and the change in the slope is represented by the second derivative of $C_W(t)$ i.e. $C''_W(t)$. Both of them decide the shape of the CGF curve.*] **[3]**

[10 Marks]

Q. 20) A health insurer is analyzing the time, measured in days, between a policyholder submitting a request for specialist treatment and the treatment being approved. Let X denote this time for a randomly selected policyholder.

The insurer models X using an exponential distribution with unknown mean θ, so that

$$f(x; \theta) = (1/\theta) \exp(-x/\theta), \quad x > 0, \theta > 0.$$

A random sample $X_1, X_2, ..., X_n$ is observed where each X_i is independent of the other.

i) Derive the log-likelihood function for θ and hence show that the maximum likelihood estimator $\hat{\theta}_{MLE} = \overline{X}$.

Also verify that the estimator maximises the likelihood function. **[4]**

ii) Show that the Cramer Rao lower bound (CRLB) for $\theta = \theta^2 / n$.

You can use the expression CRLB = $- 1 / E (d^2/d\theta^2 \log L(\theta))$ without deriving it. **[2]**

Based on industry data, θ was considered to be 18.4 days with a CRLB variance of 63 days^2.

Chief actuary of the health insurer was sceptical about the industry averages and asked the actuarial team to collect data of approval times over a period of six months and verify the authenticity of the industry data.

Following data was collected by the actuarial team:

Month 1: Sample 1: $\overline{X_1}$ = 12.16

3.2	8.7	12.4	5.6	21.3
17.8	9.1	14.6	2.8	18.5
11.2	6.4	24.7	13.5	7.9
19.6	4.3	10.8	16.2	22.1
5.1	13.7	8.2	17.4	9.0

Month 2: Sample 2: $\overline{X_2}$ = 18.40

12.6	20.3	15.8	18.1	16.7
17.5	21.4	14.2	19.6	15.8
22.7	16.1	19.3	17.8	23.5
13.9	18.6	21.8	15.4	20.7
16.9	19.8	14.7	22.1	24.7

Month 3: Sample 3: $\overline{X_3}$ = 7.30

1.4	6.8	12.5	3.1	1.7
4.6	8.3	5.8	1.6	7.2
1.8	2.4	4.9	1.6	2.4
2.6	1.9	8.6	4.7	15.4
12.5	20.8	8.9	27.6	13.4

Month 4: Sample 4: $\overline{X_4}$ = 18.67

5.8	14.3	26.7	18.2	9.4
32.5	11.7	21.6	7.3	16.8
38.4	13.5	19.7	24.8	6.2
15.9	29.1	10.6	22.4	17.3
34.6	8.7	20.5	12.9	27.8

Month 5: Sample 5: $\overline{X_5}$ = 24.97

9.7	31.4	18.6	42.3	6.8
25.7	13.2	57.1	21.5	34.8
16.9	28.6	8.4	45.2	19.7
11.3	37.5	23.8	14.6	52.4
27.1	7.9	33.6	20.4	15.8

Month 6: Sample 6: $\overline{X_6}$ = 29.28

13.4	37.7	4.6	26.5	62.2
28.8	9.1	50.6	23.3	31.9
6.7	43.5	16.4	69.3	11.8
28.2	20.6	55.8	2.5	34.4
24.7	47.1	8.9	41.3	32.8

After collecting the above data, one of the student actuaries in the team is of the view that, except sample 2, all other samples have means which are different from the industry average of $\theta = 18.4$. Hence, he considers that $\hat{\theta}_{MLE}$ is not an unbiased estimator of θ.

iii) Clearly point out the fallacy in the view of the student actuary. Support your answer with suitable numerical calculations.

Hint: [*Bias* $(\hat{\theta}_{MLE}) = E(\hat{\theta}_{MLE}) - \theta$.] **[2]**

iv) Show that $\hat{\theta}_{MLE}$ is also an asymptotically efficient estimator of θ as its variance is very close to the CRLB of 63 days^2. **[2]**

[10 Marks]

Q. 21) An electricity distributor is investigating the relationship between the average daily electricity demand and the number of commercial customers served in two regions.

For each day, let:

X = number of commercial customers served, and

Y = average daily electricity demand, in MWh (megawatt-hours).

The distributor assumes the simple linear regression model $Y_i = \alpha + \beta X_i + \varepsilon_i$, where the errors ε_i are independent with mean zero and common variance σ^2.

Data were collected separately for Region A and Region B. The following summary statistics have been calculated within each region:

Particulars	Region A	Region B
Number of observations	8	12
$\bar{x}$	20	30
$\bar{y}$	60	80
$\Sigma(X_i - \bar{x})^2$	280	420
$\Sigma(X_i - \bar{x})(Y_i - \bar{y})$	700	840
Regression line	Y = 10 + 2.5X	Y = 20 + 2X

The distributor now wishes to fit a single regression model using all 20 observations combined. The following pooled summary statistics have already been calculated for the combined data set:

$$\Sigma(X_i - \bar{x})^2 = 1180 \qquad \Sigma(X_i - \bar{x})(Y_i - \bar{y}) = 2500 \qquad \Sigma(Y_i - \bar{y})^2 = 6200$$

i) Calculate the overall values of $\bar{x}$ and $\bar{y}$ for the combined data set. **[1]**

ii) Determine the single regression line for the combined data set. **[2]**

iii) Construct a 95% two-sided confidence interval for the slope coefficient β of the combined regression line, and comment on whether it is significantly different from zero. **[4]**

The single regression line as determined above is fraught with limitations as it does not take into consideration the region to which the observations relate. An alternative approach is being explored to deal with these limitations. Under this approach, following alternative regression line is being proposed:

$$Y_i = \alpha + \beta X_i + \gamma * R_i + \lambda * (R_i * X_i) + \varepsilon_i$$

where –

- baseline intercept coefficient is α
- baseline slope coefficient is β
- R is an indicator variable which takes the value 0 for Region A and 1 for Region B
- incremental intercept coefficient for Region B is γ
- incremental slope coefficient for Region B is λ

The above regression line is expected to be consistent with the separate regression lines for regions A and B.

iv) Determine the values of the coefficients α, β, γ and λ in the above regression equation.

Hint: [*Refer to the intercept and slope coefficients of the separate regression lines for regions A and B as given in the tabular data.*] **[3]**

[10 Marks]

Q. 22) A data analyst working for MediLink, a telemedicine provider, is investigating the number of follow-up consultations required by patients during the 12 months following their initial virtual consultation.

For patient i, let Y_i denote the number of follow-up consultations and let x_i denote the patient's age.

The analyst considers modelling Y_i using a generalised linear model.

Following is the probability mass function of the response variable expressed in the form of exponential family of distributions:

$$P(Y_i = y_i) = \exp(y_i \theta_i - b_i(\theta_i) + c_i(y_i))$$

where -

1. $y_i = 0, 1, 2, \ldots$
2. $\theta_i = \log(\mu_i)$, $b_i(\theta_i) = \exp(\theta_i)$, $c_i(y_i) = -\log(y_i!)$, $a(\varphi) = 1$.

i) Show that the distribution of Y_i is a Poisson distribution with parameter μ_i.

Using the formulae for $E(Y_i)$ and $Var(Y_i)$ for distributions belonging to exponential family, also show that the mean and variance of Y_i are equal to μ_i. **[3]**

The analyst fits the following model:

$$\log(\mu_i) = \alpha + \beta x_i.$$

The fitted model gives

$$\hat{\alpha} = 0.40, \quad \hat{\beta} = 0.025.$$

Consider the following two patients:

Patient	Age, x_i	Observed number of follow-up consultations, y_i
A	40	6
B	60	8

ii) Calculate the fitted mean number of follow-up consultations for each patient. **[2]**

For a Poisson observation with observed value y and fitted mean $\hat{\mu}$, the contribution to the deviance is:

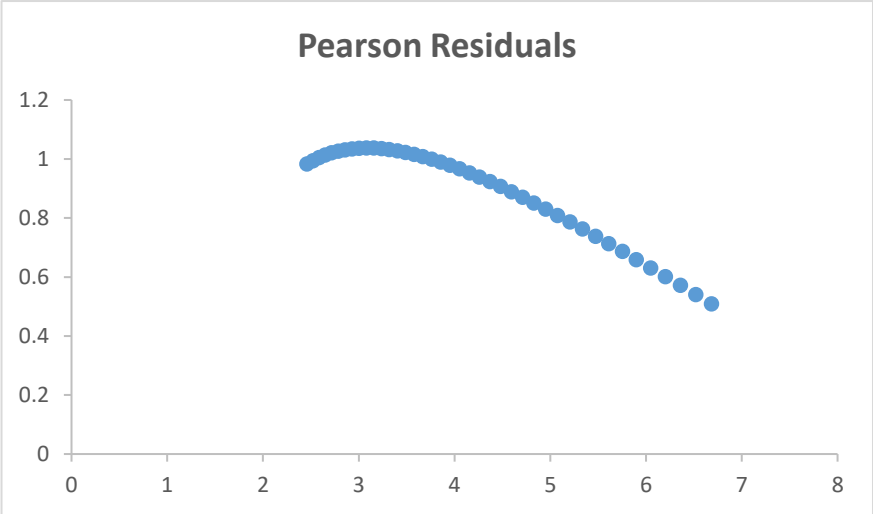
$$d^2 = 2[y \log(y/\hat{\mu}) - (y - \hat{\mu})].$$

Take the deviance residual to have the same sign as $(y - \hat{\mu})$.

iii) Calculate the Pearson residual and deviance residual for Patient A. **[3]**

iv) Without carrying out detailed calculations for Patient B, determine whether its Pearson and deviance residuals will be positive or negative. Explain your reasoning. **[1]**

The analyst finds that, across the full dataset from age 20 to age 60, when Pearson residuals are plotted against the predicted values, it gives a positively skewed graph.



He has concluded that as the plot of Pearson residuals is positively skewed, the Poisson model is not a good fit for the data.

v) Is it pre-mature to conclude that the model is not a good fit solely based on Pearson residuals? What other checks can the analyst perform before reaching a conclusion about the fit of the model? **[1]**

[10 Marks]

****