INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

November 2025

CS1 - Actuarial Statistics (Paper A)

Time allowed: 3 Hours 15 Minutes

Total Marks: 100

Q.1) A random variable X is normally distributed with mean 5 and standard deviation 2.

a) Calculate the probability that $X^2 > 9$. (2)

b) Profit is given by $a + bX$. The expected profit is 10 and the 80th percentile of the profit is 16. Calculate the 90th percentile of the insurance company's profit in the year. (2)

c) $Y = e^X$. What is $P(Y > 70)$

a) 0.355
b) 0.500
c) 0.646
d) 0.750
e) 0.495

(2)

[6]

Q.2) Let $X_1, X_2, X_3, \dots, X_{30}$ be independent and identically distributed random variables from a Poisson distribution with mean $\lambda = 2$. Let $\bar{X} = \frac{1}{30}\sum_{i=1}^{30} X_i$ be the sample mean.

i) Find the moment generating function of a single X_i.

a) e^{2t}
b) $e^{2(e^t-1)}$
c) $\frac{1}{1-2t}$
d) e^{t^2}
e) $(1-2t)^{-1}$

(2)

ii) Find the moment generating function of the sum $S = \sum_{i=1}^{30} X_i$.

a) $e^{30(e^t-1)}$
b) $e^{60(e^t-1)}$
c) $e^{(e^t-1)}$
d) $e^{30(2^t-1)}$
e) $e^{30(t-1)}$

(2)

iii) Use the Central Limit Theorem to approximate $P(\bar{X} < 2.2)$. (2)

[6]

Q.3) A company's server monitoring system tracks events using Poisson processes.

i) The number of software crashes experienced by a startup in one year follows a Poisson distribution with mean 2.

a) Calculate the expected number of crashes in a year, given that there are at least two crashes. (2)

b) Calculate the variance of the number of software crashes in a year given that at least one software crash occurs. (2)

ii) Security alerts are generated at a Poisson rate of $\lambda = 4$ per hour.
Calculate the probability that the fourth alert arrives after more than one hour. (2)

iii) IT support tickets arrive according to a Poisson distribution with $\lambda = 6$ per hour. Tickets are categorized as priority: 30% are Low-priority (A), 45% are Medium-priority (B), and 25% are High-priority (C).
Calculate the probability that at least two Medium-priority tickets arrive in the first 2 hours. (2)

iv) IT support ticket arrives according to a Poisson distribution with $\lambda = 6$ per hour. If a ticket arrives after 20 minutes, then it is labelled as Low priority. Calculate the probability that we observe two low-priority tickets in one hour. (3)

[11]

Q.4) A factory monitors the lifetime (in hours) of a certain type of machine part. The lifetime of each part follows a gamma distribution with shape parameter $\alpha = 3$ and scale parameter θ. A random sample of size n is taken to test whether the average lifetime has increased.

We wish to test the hypotheses:

$H_0: \theta = \frac{1}{4}$ and $H_1: \theta > \frac{1}{4}$

The rejection rule is, reject H_0 if $\bar{X} >= 14.5$ where $\bar{X}$ is the sample mean.

i) Find the minimum sample size required so that the probability of a Type I error is at most 4% using a normal approximation. (3)

ii) Suppose instead that the lifetime is normally distributed with known variance 9, and a sample of 10 parts is taken. We test

$$H_0: \mu \leq 50 \; and \; H_1: \mu > 50$$

at the 5% significance level. If the true mean lifetime is 52 hours, find the probability of a Type II error. (3)

iii) A machine part's lifetime (in hours) is known to follow an **Exponential distribution** with mean θ

A reliability engineer wishes to test whether the average lifetime has increased.

$$H_o: \theta = 20 \; and \; H_1: \theta = 30$$

Only one machine part is tested, and its lifetime X is observed. The engineer decides to reject H_o if X>c, where c is the critical value. If the type-ii error is set to 0.75, **calculate c** and the **Type I** error probability. (3)

[9]

Q.5) A streaming service surveyed users from two age groups (Under 25 and Over 25) about their preference for a new recommendation algorithm (Like New vs. Prefer Old). The results are:

	Under 25	Over 25
Like New	650	720
Like Old	150	180

i) Perform a χ^2 test on this contingency table to test the hypothesis that there is no association between age group and algorithm preference.

H_o: There is no association between age group and product preference.
H_1:There is an association between age group and product preference. (5)

A behavioral model proposes that the user population is split in the proportions below, where $p(0 < p < 1)$ is a parameter relating to the preference for the old algorithm.

	Under 25	Over 25
Like New	$\frac{1-p}{2}$	$\frac{1-p^2}{2}$
Like Old	$\frac{p}{2}$	$\frac{p^2}{2}$

The maximum likelihood estimate of *p* from the data is 0.212.

ii) Test the goodness of fit of this model to the data. (5)

[10]

Q.6) A university wants to estimate the proportion p of students who use the library weekly. The prior belief is $p \sim Beta(2,5)$. A survey of $n = 100$ students shows that 60 use the library weekly.

i) Find the posterior distribution of p.

a. $\text{Beta}(62,45)$
b. $\text{Beta}(60,40)$
c. $\text{Beta}(61,43)$
d. $\text{Beta}(60,45)$
e. $\text{Beta}(65,45)$

(2)

ii) Calculate the posterior mean and variance. (2)

iii) Give a 95% credible interval for p using the normal approximation to the Beta distribution.

a. $(0.451,0.820)$
b. $(0.486,0.672)$
c. $(0.5,0.6)$
d. $(0.5,0.672)$
e. $(0.420,0.625)$

(2)

[6]

Q.7) A financial analyst is studying the relationship between company profit (Y in ₹ crore) and research expenditure (X in ₹ crore). Data from 7 companies:

Company	Research(X)	Profit(Y)
1	2	15.5
2	3.5	22.8
3	1.5	12.2
4	4	25.1
5	2.5	18.3
6	3	20.7
7	5	30.5

i) Calculate the fitted regression line: $\hat{y} = \alpha + \beta x$. (3)

ii) The analyst wants to test whether there is a significant linear relationship between research expenditure and profit. Which of the following hypotheses should be tested?

a. $H_0: \beta = 0$ vs $H_1: \beta \neq 0$
b. $H_0: \alpha = 0$ vs $H_1: \alpha \neq 0$

c. H_0: ρ = 0 vs H_1: ρ ≠ 0
d. H_0: β = 1 vs H_1: β ≠ 1
e. H_0: β = 1 vs H_1: β ≠ 0

(1)
[4]

Q.8) A data scientist is building GLMs for different types of data in a manufacturing quality-control system.

i) The occurrence of a defect in a component follows a Bernoulli distribution with probability p:

$$f(y;p) = p^y(1-p)^{(1-y)}, y \in \{0,1\}$$

where p is the probability of success.

Identify the canonical parameter θ, the function $b(\theta)$, the dispersion parameter ϕ, and the functions $a(\phi)$and $c(y,\phi)$. (2)

a) $\theta = p, b(\theta) = \theta(1-\theta), \phi = 1, a(\phi) = 1, c(y,\phi) = 0$
b) $\theta = log\left(\frac{p}{1-p}\right), b(\theta) = log(1+e^{\theta}), \phi = 1, a(\phi) = 1, c(y,\phi) = 0$
c) $\theta = p, b(\theta) = -log(1-\theta), \phi = 1, a(\phi) = 1, c(y,\phi) = log(1-p)$
d) $\theta = log(p), b(\theta) = e^{\theta}, \phi = 1, a(\phi) = 1, c(y,\phi) = 0$
e) $\theta = \frac{p}{1-p}, b(\theta) = log(1+\theta), \phi = 1, a(\phi) = 1, c(y,\phi) = 0$

ii) The lifetime of components (in hours) follows a Gamma distribution with known shape parameter k and scale parameter β, with PDF:

$$f(y;k;\beta) = \frac{1}{\Gamma(k)\beta^k} y^{k-1} e^{-\frac{y}{\beta}}$$

Choose the correct canonical link function for a Gamma GLM (with known shape k):

a) $g(\mu) = \mu$
b) $g(\mu) = log(\mu)$
c) $g(\mu) = \frac{1}{\mu}$
d) $g(\mu) = \mu^2$
e) $g(\mu) = \mu$

(2)

iii) Determine which of the following statements about GLMs are true:

I. The highest possible deviance can only be attained by the saturated model.
II. For all models in the exponential family, deviance follows a χ^2 distribution.

III. For a Bernoulli GLM, using the log link function ensures that the predicted probabilities are always between 0 and 1.

a) I and II
b) I and III
c) III
d) I, II and III
e) None of the above

(2)

[6]

Q.9) An analyst is studying factors affecting urban housing prices using regression models on a dataset of properties.

i) Simple Linear Regression for Land Size vs. Price

For a simple linear regression model $y = \beta_o + \beta_1 x + \epsilon$, you are given the following summary statistics from a sample of 25 observations:

- The unbiased sample covariance between x and y is 88.5.
- The sum of squared deviations of x is $\sum_{i=1}^{25}(x_i - \bar{x})^2 = 1250$
- The sample mean of y is 50.
- The sample mean of x is 10.

Calculate the ordinary least squares estimates of the slope ($\widehat{\beta_1}$) and the intercept ($\widehat{\beta_o}$). (2)

ii) Regression Goodness-of-Fit Metrics

For another simple linear regression (number of bedrooms vs. price) on 20 properties:

- The total sum of squares (SST) is 1450.
- The error sum of squares (SSE) is 522.

Calculate:

a) Calculate the degrees of freedom. (1)

b) The regression sum of squares (SSR). (1)

c) The coefficient of determination R^2. (2)

d) The value of the unbiased estimate of the error variance (σ^2 or MSE). (2)

iii) Multiple Regression Model Comparison

The analyst builds two multiple regression models to predict house prices $(n = 50)$

- **Model 1:** Uses 4 predictor variables.
  - $R^2 = 0.712$
  - Residual Standard Error = 35,000
- **Model 2:** Uses 7 predictor variables.
  - $R^2 = 0.725$

a) Calculate the Adjusted R^2 for Model 1. (2)

b) The analyst states, "Model 2 is definitively a better model than Model 1 because it has a higher R^2"

Critique this statement by discussing the limitations of R^2 for model comparison and explaining how Adjusted R^2 provides a better metric. Your answer should include the necessary calculations to *support* your critique, even if you cannot fully compute Adjusted R^2 for Model 2. (3)

iv) A modeler uses a GLM to model house price across location and property type. He has the following information about the model:

- House price follows a Gamma distribution.
- Log is the selected link function.
- Urban is the base level for Location.
- Apartment is the base level for Property type.
- The modeler achieves the following model output:

Variable	$\hat{\beta}$
Intercept	12.62
Location – Urban	0
Location – Rural	-0.25
Property type – Apartment	0
Property type - House	0.4

a) Predicted (expected) house price for an **urban house** and a **rural apartment.** (2)

A) $300,000, $300,000
B) $447,540, $348,550
C) $451,350, $235,625
D) $348,550, $233,640

b) Suppose interaction terms were included in the model and the GLM produces the following predicted values for house price.

	Urban	Rural
House	300,000	260,000
Apartment	447,540	420,000

Calculate the estimated beta for the interaction of Location Rural and Property type House. (3)

[18]

Q.10) A national supermarket chain has launched a new customer loyalty card program. The data science team has collected three months of transaction data for 5,000 customers, measured across 10 standardized variables. The correlation matrix of these variables was found to have several high correlations, so Principal Component Analysis (PCA) was performed on the standardized data. Partial R output is shown below:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.8741	1.6327	1.2214	0.9452	0.8884
Proportion of variance	0.3512	0.2666	0.1492	0.0893	0.0789
Cumulative proportion	0.3512	0.6178	0.7670	0.8563	0.9352

i) Calculate the variance for the third principal component (PC3). (1)

ii) List two advantages of PCA. (2)

iii) Using the **Kaiser's criterion** determine how many principal components should be retained. Show your calculations. (2)

iv) Rank Correlation Analysis of Customer Spending Patterns.

The team investigates whether customers' spending ranks in the first month and predicts their spending ranks in the third month. Data for 8 randomly selected loyalty card customers:

Customer	Month-1 spending rank(X)	Month-3 spending rank(Y)
A	3	2
B	1	1
C	7	9
D	5	5
E	4	6
F	10	8
G	9	10
H	2	4

Calculate **Spearman's rank correlation coefficient** (ρ) for this data. (4)

[9]

Q.11) An automotive manufacturer uses Bayesian methods to model various risks across production quality, warranty claims, and component reliability.

i) The quality control department models the number of defective components in production batches. Each batch contains m components and X_i is number of defective items in batch i, following a binomial distribution with unknown defect rate θ is unknown. Let $X = (X_1, X_2, \cdots, X_n)$ denote n batches. With n batches sampled, and a beta distribution with parameters α and β prior on θ, determine the posterior distribution of θ . (2)

a) $Beta(\alpha + \sum_{i=1}^{n} X_i, \beta + nm - \sum_{i=1}^{n} X_i)$

b) $Beta(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i)$

c) $Beta(\alpha + n, \beta + nm - \sum_{i=1}^{n} X_i)$

d) $Beta(\alpha + \sum_{i=1}^{n} X_i, \beta + m - \sum_{i=1}^{n} X_i)$

e) $Beta(\alpha + nm, \beta + \sum_{i=1}^{n} X_i)$

ii) Derive the **maximum likelihood estimate** of θ, denoted by $g(x)$. (2)

a) $\hat{\theta} = \frac{\sum x_i}{mn}$

b) $\hat{\theta} = \frac{\sum x_i}{n}$

c) $\hat{\theta} = \frac{\sum x_i}{m}$

d) $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{m}$

e) $\hat{\theta} = \frac{\sum x_i}{\beta + mn}$

iii) Under quadratic loss, the Bayes estimate of θ can be written in the credibility form

$$\theta = Z * g(x) + (1 - Z) * \mu,$$

where μ is the prior mean. Determine the credibility weight Z. (2)

a) $Z = \frac{mn}{2\beta + mn}$

b) $Z = \frac{mn}{\beta + mn}$

c) $Z = \frac{n}{2\beta + n}$

d) $Z = \frac{\sum x_i}{2\beta + \sum x_i}$

e) $Z = \frac{2\beta}{2\beta + mn}$

iv) Each production batch contains $m = 8$ components, and observations from $n = 5$ batches are as follows:

$$x_1 = 2, \qquad x_2 = 1, \qquad x_3 = 0, \qquad x_4 = 3, \qquad x_5 = 2$$

Calculate the Bayesian estimate of the defect probability θ and the corresponding credibility factor Z when $\beta = 6$. Verify that the credibility form also results in the same values. (3)

[9]

Q.12) Define the following terms:

i) Neyman-pearson lemma (2)

ii) Type I and Type II error (2)

iii) The size of a test (1)

iv) The power of a test (1)

[6]

****