Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper A)

November 2023 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) Correct Answer is **Option B**

We know that Cx(t) = ln (Mx(t)). Hence, $M_x(t) = e^{Cx(t)}$. [1]

ii)

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \frac{t^4}{4!}E(X^4) + \cdots$$
$$M'_X(t) = E(X) + \frac{2t}{2!}E(X^2) + \frac{3t^2}{3!}E(X^3) + \frac{4t^3}{4!}E(X^4) + \cdots$$
$$= E(X) + tE(X^2) + \frac{t^2}{2!}E(X^3) + \frac{t^3}{3!}E(X^4) + \cdots$$
$$M'_X(0) = E(X)$$

$$M''_X(t) = E(X^2) + tE(X^3) + \frac{t^2}{2!}E(X^4) + \cdots$$
$$M''_X(0) = E(X^2)$$
$$var(X) = E(X^2) - [E(X)]^2 = M''_X(0) - [M'_X(0)]^2$$ [3]

iii) Correct Answer is **Option C**

$$C_X(t) = ln\, M_x(t)$$
$$C'_X(t) = \frac{1}{M_x(t)}M'_X(t)$$
$$C''_X(t) = -\frac{1}{[M_x(t)]^2}M'_X(t)M'_X(t) + \frac{1}{M_x(t)}M''_x(t)$$
$$= \frac{M''_x(t)}{M_x(t)} - \left(\frac{M'_X(t)}{M_x(t)}\right)^2$$
$$C''_X(0) = \frac{M''_x(0)}{M_x(0)} - \left(\frac{M'_X(0)}{M_x(0)}\right)^2 = M''_x(0) - [M'_X(0)]^2 = E(X^2) - [E(X)]^2$$
$$C''_X(0) = var(X)$$ [2]

[6 Marks]

Solution 2:

i) Correct Answer is **Option C**

Type I error represents False Positives i.e. where actually innocent individuals (as decided by the court later) are found to be guilty based on the lie-detector test. This is thus represented by Area B in the matrix.

Type II error represents False Negatives i.e. where a person who is actually guilty of the crime (as decided by the court later) is considered to be innocent based on the lie-detector test. This is thus represented by Area C in the matrix.

Kindly note that here Positive means "being found guilty" and Negative means "being innocent". [2]

ii) a) Probability of Type I Error
= Probability (False Positive)
= Probability (Person is Innocent but has been identified as guilty by the lie-detector)
= P(I | LG)

b) Probability of Type II Error

= Probability (False Negative)
= Probability (Person is Guilty but has been identified as innocent by the lie-detector)
= P(G | LI)

[2]

iii) Correct Answer is **Option A**

Probability that the lie-detector correctly identifies the perpetration / non-perpetration of the crime is given by (A+D) / (A + B + C + D)
= (356 + 428) / 1000
= 0.784 [1]

iv) Probability of Type I Error
= P(I | LG)
= B / (B+D)
= 111/(111+428)
= 0.206

Probability of Type II Error
= P(G | LI)
= C / (A+C)
= 105 / (356+105)
= 0.228

[3]

[8 Marks]

Solution 3:

i) The three components of a GLM are:

(1) A distribution for the response variable – This belongs to the exponential family.
(2) A linear predictor η – This is a linear function of the covariates.
(3) A link function g – This connects the mean response to the linear predictor, g(μ) = η

[3]

ii) Correct Answer is **Option C**

The number of parameters (standalone) for each factor (excluding passes) is:

- *Age – 2 (including intercept)*
- *Experience – 2 (including intercept)*
- *Duration – 2 (including intercept)*

Exam passes: This is effectively 13 yes / no factors. Each of these 13 factors would contribute 2 parameters on standalones basis (including intercept).

The main effects are therefore going to contribute: 2 + 13(2-1) + (2 – 1) + (2 – 1) = 17 parameters.* [2]

iii) Correct Answer is **Option D**

*The interactions will contribute (2-1) * (2-1) = 1 parameter.* [1]

iv) Scaled Deviance = 2 (ln L_S – ln L_M)
15 = 2 (16 – ln L_M)
ln L_M = (32 – 15) / 2 = 8.5

AIC
= -2 * ln L_M + 2 * number of parameters
= -2 *8.5 + 2 * (17+1)

= -17 + 2 * 18
= -17 + 36
= 19 [4]

[10 Marks]

Solution 4:

i) **Based on sample mean:**

$X\_bar = \sum x / n = 5.13/10 = 0.513$

Mean of population = (b+a)/2 = θ / 2

Equating population mean with sample mean,
X_bar = θ / 2
0.513 = θ / 2
$\hat{\theta}_{MOM}$ = 0.513*2 = 1.026

Based on sample variance:

S^2
$= 1/(n-1) * (\sum x^2 - n * (\sum x / n)^2)$
$= 1/9 * (2.8761 - 10 * 0.513^2)$
= 1/9 * 0.24441
= 0.027157

Variance of the population
$= (b - a)^2 / 12$
$= \theta^2 / 12$

Equating population variance with sample variance,
$S^2 = \theta^2 / 12$
$\hat{\theta}_{MOM}$
$= (0.027157*12)^{0.5}$
= 0.5709 .... as θ > 0. [3]

ii) Correct Answer is **Option B**

X is uniformly distributed over the interval [0, θ]. So, X can take values which lie between 0 and θ and not beyond θ.

$L(\theta, x) = 1 / \theta^{10}$ for all $x_i \le \theta$ i.e. $\theta > \max(x_i)$
$L(\theta, x) = 0$ otherwise. [1]

iii) $L(\theta, x) = 1 / \theta^{10}$
log L = - 10 * log θ

Differentiating both sides,
d/dθ (log L)
= d/dθ (-10 * log θ)
= -10/ θ

Required equation to be solved to get MLE is:
d/dθ (log L) = 0
-10/ θ = 0

Kindly note that when we try to equate it with 0, we will get that θ tends to infinity. So, we won't get a finite value which maximizes the likelihood function.

Similarly, when we take a second derivative to check maxima,

$d^2/d\theta^2$ (LogL) = $10 / \theta^2 >$ --------- This is indicative of minima and not maxima
Hence, the method of differentiation does not work in this case. [3]

iv) $L(\theta, x) = 1 / \theta^{10}$ for $\theta > \max(x_i)$
$L(\theta, x) = 0$ otherwise.

Hence, till the point of $\theta = \max(x_i)$ i.e. for all values of θ which are lower than $\max(x_i)$, the value of the likelihood function is equal to 0 as can be seen in the graph.

At $\theta = \max(x_i)$, the likelihood function sees a sudden spike and for all values of θ which are greater than $\max(x_i)$, the likelihood function keeps on declining.

So, it is evident from the graph that the likelihood function is maximized at $\theta = \max(x_i)$. So, $\hat{\theta}_{MLE} = \max(x_i)$ [2]

v) E(Z)
= integral ($10 * z * z^9 / \theta^{10}\,dz)_0^\theta$
= $10 / \theta^{10} * (z^{11} / 11)_0^\theta$

= $10 / \theta^{10} * (\theta^{11} / 11)$
= $10/11 * \theta$

Bias ($\hat{\theta}_{MLE}$)
= $E(\hat{\theta}_{MLE}) - \theta$
= $E(Z) - \theta$
=$10/11 * \theta - \theta$
= $-1/11 * \theta$
= $- 11^{-1} * \theta$ [2]

vi) Correct Answer is **Option C**

$MSE\ (\hat{\theta}_{MLE})$
$= E\ (\hat{\theta}_{MLE} - \theta)^2$
$= Var\ (\hat{\theta}_{MLE}) + Bias^2(\hat{\theta}_{MLE})$ [1]

[12 Marks]

Solution 5:

i) Correct Answer is **Option D**

Since 1) from Column A relates to returns over a period of time, it is longitudinal data (Option ii from Column B).

Since 2) from Column A relates to stocks listed during the middle of the year (hence data prior to their listing would not be publicly available). Hence, it is censored data (Option iv from Column B)

Since 3) from Column A relates to value of LENSEX at a point of time, it is cross-sectional data (Option iii from Column B)

Since 4) from Column A relates to stocks on MSE during a truncated week (which has less than 5 working days due to presence of public holidays), it is truncated data (Option I from Column B) [2]

ii) Probability that the sample returns for MSE's MIFTY are more volatile as compared to the population

$= P(S^2_x > \sigma^2)$

$= P(S^2_x / \sigma^2 > 1)$

$= P((10-1) * S^2_x / \sigma^2 > 9)$

$= P(9 * S^2_x / \sigma^2 > 9)$

$(9 * S^2_x / \sigma^2) \sim \chi^2_9$

Probability that the sample returns for MSE's MIFTY are more volatile as compared to the population

$= P(\chi^2_9 > 9)$

$= 1 - P(\chi^2_9 \leq 9)$

$= 1 - 0.5627$ ……………….. taken from Actuarial Tables

$= 0.4373 \sim 0.4$ (as required) [4]

iii) Sample mean $\overline{X}$ and sample variance S^2_x are independent on each other. This is true because the distribution of the population is normal distribution. For non-normal distributions, it may not be true. [1]

iv) Conditional probability that the sample tends to overestimate the average annual return of MIFTY as compared to the population

$= P(\overline{X} > \mu \mid S^2_x > \sigma^2)$

$= P(\overline{X} > \mu)$ ……………… $\overline{X}$ and S^2 are independent on each other

$= P(\overline{X} - \mu > 0)$

$= P((\overline{X} - \mu) / (\sigma / \sqrt{n}) > 0)$

$= P(Z > 0)$ …………………. as $\overline{X} \sim N(\mu, \sigma^2 / n)$

$= 1 - P(Z \leq 0)$

$= 1 - 0.5 = 0.5$ [3]

v) Probability that the sample standard deviation of the returns of MIFTY is lower than the sample standard deviation of the returns of LENSEX

$= P(S_x \leq S_y)$

$= P(S^2_x \leq S^2_y)$

$= P(S^2_x / S^2_y \leq 1)$

$= P((S^2_x / S^2_y) * 1.5 \leq 1.5)$

$= P((S^2_x / 16) / (S^2_y / 24)) \leq 1.5)$

$= P((S^2_x / \sigma^2) / (S^2_y / \beta^2)) \leq 1.5)$

$= P(F_{9,9} \leq 1.5)$

$= P(F_{9,9} > 2/3)$

As per Actuarial Tables, $P(F_{9,9} > 2.440) = 10\%$, $P(F_{9,9} > 3.179) = 5\%$

Hence $P(F_{9,9} > 2/3) > 10\%$ [4]

vi) Correct Answer is **Option A**

As mentioned in part (vi), principal components are un-correlated combinations of the original random variables, so in this variance / covariance matrix as well, the value of covariances for any two distinct PCs will be equal to 0. So, 870 values will be 0.

[1]

Only the 30 values in the diagonal will represent the variance of each PC. So, the maximum number of non-zero values in the variance / covariance matrix will be 30.

[15 Marks]

Solution 6:

i) $L(\mu)$

$= 1/\mu * e^{-x1/\mu} * \ldots\ldots\ldots\ldots\ldots * 1/\mu * e^{-xn/\mu}$

$= e^{-\sum x/\mu} / \mu^{n}$ [2]

ii) The posterior PDF is proportional to the prior PDF multiplied by the likelihood function.

$f_{post}(\mu) \propto (e^{-\theta/\mu} / \mu^{\alpha+1}) * (e^{-\sum x/\mu} / \mu^{n})$

$= e^{-(\theta + \sum x)/\mu} / \mu^{n+\alpha-1}$

This has the same form as the prior distribution, but with different parameters. So we have the same distribution, but with parameters:

$\alpha^* = n + \alpha$

$\theta^* = \theta + \sum x$ [3]

iii) Using the formula for mean given in the question,

$E(\mu \mid \underline{x})$

$= \theta^* / (\alpha^* - 1)$

$= \frac{\theta + \sum x}{n+\alpha-1}$

This is the Bayesian estimate of μ under squared error loss. [2]

iv) Correct Answer is **Option B**

Option A is incorrect as the prior and posterior distributions need not be exactly identical for conjugate priors. Even if the prior and posterior distributions belong to the same family, even then the two are considered as conjugate priors.

Option B is correct as conjugate priors make Bayesian calculations simpler.

Option C is incorrect as the distribution of random variable X need not be identical with the posterior distribution, for conjugate priors.

Option D is incorrect, as a prior uniform distribution does not necessarily lead to a posterior uniform distribution. [1]

v) Splitting the formula for posterior mean into two parts, we see that:

$E(\mu \mid \underline{x})$

$= \frac{\theta + \sum x}{n+\alpha-1}$

$= \frac{\theta}{n+\alpha-1} + \frac{\sum x}{n+\alpha-1}$

$= \frac{n}{n+\alpha-1} \times \frac{\sum x}{n} + \frac{\alpha - 1}{n+\alpha-1} \times \frac{\theta}{\alpha-1}$

This is a weighted average of the maximum likelihood estimate of μ (which is the sample mean $\frac{\sum x}{n}$) and the mean of the prior distribution $\frac{\theta}{\alpha-1}$. So it is a credibility estimate.

The credibility factor is:

$$Z = \frac{n}{n+\alpha-1}$$ [3]

vi) Prior mean $= 40 / (1.5 - 1) = 80$

Sample Mean $= 9000/100 = 90$

Using the given figures, the Bayesian estimate of μ is:

$= \frac{\theta + \sum x}{n+\alpha-1}$
$= (40 + 9000) / (100 + 1.5 - 1)$
$= 89.95$

The value of the credibility factor is:

Z
$= \frac{n}{n+\alpha-1}$
$= 100 / (100 + 1.5 - 1)$
$= 0.9950$ [3]

vii) Correct Answer is **Option A**

Since the value of the credibility factor is close to 1, the posterior estimate is closer to the sample mean (direct data) as compared to the prior mean (collateral data). Sample mean in this case is 90 and prior mean is 80. It is evident that the posterior estimate 89.95 is closer to the sample mean. [1]

[15 Marks]

Solution 7:

i) Reading the formulae from the table,

E(Y) = α/λ = θυ

Var(Y) = α/ λ^2 = $\theta\upsilon^2$

Mean = θυ = 4 * 25 = INR 100 lakh

Standard deviation = $\sqrt{\theta\upsilon^2}$ = INR 50 lakh [2]

ii) From the table, the 99.5th percentile value of Gamma(4,1/25) is INR 274.44 lakhs. This is the capital requirement. [1]

iii) Correct Answer is **Option C**

Independent and identically distributed needs the following two aspects –

- *Variables have identical distributions (belonging to the same family of distributions is not sufficient, the distributions have to be identical)*
- *Variables are not dependent (uncorrelated variables are not sufficient, we know that when two variables are independent they are necessarily uncorrelated, but when two variables are uncorrelated they are not necessarily independent of each other)* [1]

iv) Correct Answer is **Option C**

For N, a discrete distribution would be appropriate, as the number of claims would be a non-negative whole number.

For X, a continuous distribution would be appropriate, as it is a monetary amount, which is a continuous quantity.
(Although it may be argued that money is not infinitely subdivisible, modelling it as a discrete quantity serves no purpose – e.g. if the 99.99th percentile of X is 1,000,000; modelling a range of 0 to 1,000,000 in steps of say 0.01 would get us no significant increase in accuracy.) [1]

v) $E(X) = \alpha\beta$

$E(Y) = E[E(Y|N)] = E[N\ E(X)] = E[N\ \alpha\beta] = E(N)\ \alpha\beta = \mu\alpha\beta$

$Var(X) = \alpha\beta^2$

$Var(Y) = E[Var(Y|N)] + var[E(Y|N)]$
$E[Var\ (Y|N)] = E[N\ Var(X)] = E(N)\ \alpha\beta^2 = \mu\alpha\beta^2$

$Var[E(Y|N)] = Var[N\ E(X)] = var(N)\ \alpha^2\beta^2 = \mu\alpha^2\beta^2$

Thus $Var\ (Y) = \mu\alpha\beta^2 + \mu\alpha^2\beta^2 = \mu\alpha\beta^2\ (1+\alpha)$

$E(Y) = \mu\alpha\beta = 10 * 2/3 * 15$ lakh = INR 100 lakh

$Var(Y) = E(Y)\ (1 + \alpha)\ \beta = 100$ lakh $* 5/3 * 15$ lakh $= 25 * 10^{12}$ [7]

vi) First, we need to simulate a value for N, which follows Poisson (10).
First random variate is 0.19.
From Tables, P(N = 6) < 0.19 < P(N=7). Thus the simulated value of N is 7.

We now need to simulate 7 claims.

The next 7 random variates and the corresponding values from the Gamma (2/3, 1/15) distribution are:

Claim No	Variate	Corresponding X value
1	0.95	34.64
2	0.70	11.48
3	0.80	16.46
4	0.20	1.21
5	0.10	0.41
6	0.20	1.21
7	0.50	5.65

Adding up the X values, the simulated value of Y is 71.06 lakh rupees. [4]

[16 Marks]

Solution 8:

i) $\bar{x} = 13.8$
$\bar{y} = 7.58$

[2]

ii)

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 230.8$$

$$S_{yy} = \sum y_i^2 - n\bar{y}^2 = 172.828$$

$$S_{xy} = \sum x_i\ y_i - n\overline{xy} = 196.78$$

[3]

iii)

$$\beta = \frac{S_{xy}}{S_{xx}} = 0.85$$

$$\alpha = \bar{y} - \beta\bar{x} = -4.186$$

So the fitted regression line is:

$\hat{y} = -4.186 + 0.85\text{x}$. [3]

iv) A 99% confidence interval for the slope parameter β is given by:

$$\widehat{\beta} \pm t_{n-2;0.005}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}$$

From our data,

$\hat{\sigma}^2$

$= 1 / (\text{n} - 2) * (\text{S}_{\text{yy}} - \text{S}^2_{\text{xy}} / \text{S}_{\text{xx}})$

$= 1 / 3 * (172.83 - 196.78^2 / 230.8)$

$= 1.6845$

So the 99% confidence interval for the slope parameter β is given by:

$0.85 \pm 5.841\sqrt{\frac{1.6845}{230.8}}$

$= 0.85 \pm 0.499$

$= (0.351, 1.349)$ [4]

v) Correct Answer is **Option A**

The residual values don't seem to have any relationship with x and seem to be distributed approximately normally around the origin. As such, the distribution is as expected. [1]

vi) Correct Answer is **Option A**

Proportion of variance explained by the model is given by R^2.

$R^2 = S^2_{xy} / (S_{xx} * S_{yy}) = 200^2 / (190*250) = 0.84$ [1]

vii)

$$Adjusted\ R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)$$

For one variable model,

$$Adjusted\ R^2 = 1 - \left(\frac{6-1}{6-1-\ 1}\right)(1-0.84) = 0.80$$

For two variable model,

$$Adjusted\ R^2 = 1 - \left(\frac{6-1}{6-2-\ 1}\right)(1-0.87) = 0.78$$

As the adjusted R^2 for the one variable model is higher, it can be considered the better model. [4]

[18 Marks]

****