# INSTITUTE OF ACTUARIES OF INDIA

# Subject CS1A – Actuarial Statistics (Paper A)

# May 2024 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:** Correct Answer is **Option D.**

*Let X denote the number of likes received in a minute.*
*Since, the number of likes received per second is 0.4, X ~ Poi(24).*

*P(X=40)*
*= $24^{40}$ / 40! * exp(-24)*
*= 0.075%* **[2]**

**Solution 2:** Correct Answer is **Option B.**

*If the number of likes follow a Poisson process with mean of 0.4 likes per second, the waiting time between two consecutive likes is exponentially distributed with parameter = 0.4.*

*Mean waiting time*
*= Mean(Exp(0.4))*
*= 1/0.4*
*= 2.5 seconds.* **[2]**

**Solution 3:** Correct Answer is **Option C.**

*Numero Cero has waited for 10 seconds already, and we has to wait for 10 more seconds. Using the memoryless property of an exponential distribution, waiting for 10 more seconds given that there is already a waiting of 10 seconds, is equivalent to waiting for 10 seconds.*

*P(W > 10)*
*= exp(-0.4*10)*
*= 1.832%* **[2]**

**Solution 4:** Correct Answer is **Option A.**

*f(x) = exp(-x)*
*f(y) = ½ *exp(-½y)*

*As X and Y are independent variables,*
*f (x, y) = f(x) * f(y)*

*f (x, y) = exp(-x) × ½ exp(-½y) for ∞ > x, y > 0;*
*= 0 otherwise* **[2]**

**Solution 5:** Correct Answer is **Option B.**

*$M_x(t) = E[exp(tx)]$*
*$M_y(s) = E[exp(sy)]$*

*$M_{x,y}(t, s)$*
*= E [ exp(xt + ys) ]*
*= E[exp(xt)] * E[exp(ys)]*
*= $M_x(t) * M_y(t)$* **[2]**

**Solution 6:** Correct Answer is **Option C.**

$\hat{\lambda}_{MOM} = \sum y / \sum x = 55 / 480 = 0.115$ **[1]**

**Solution 7:** Correct Answer is **Option D.**

*Expression to be minimized*
$=\sum(y - E(y))^2$
$=\sum(y - \lambda x)^2$

$$\frac{d}{d\lambda}\sum(y - \lambda x)^2$$

$= -2\sum xy\ (1) + 2\lambda(\sum x^2)$

*We equate this with 0 to find the value of least square estimator of* $\lambda$

$\lambda_{LSE} = \sum xy / \sum x^2$

$\frac{d^2}{d\lambda^2}\sum(y - \lambda x)^2 = 2(\sum x^2) > 0$ ------- *this is indicative of minima* **[3]**

**Solution 8:** Correct Answer is **Option A.**

*L(λ)*
$= e^{-\lambda\sum x} \times \prod(\lambda x)^y \times 1/\prod y!$
$= e^{-\lambda\sum x} \times \prod(\lambda x)^y \times$ *constant* **[2]**

**Solution 9:** Correct Answer is **Option B.**

*log L(λ)*
$= -\lambda\sum x + \sum log\lambda^y +$ *constant*
$= -\lambda\sum x + log\lambda\sum y +$ *constant*

*d/dλ (log L(λ))*
$= -\sum x + \sum y\ (1/\lambda)$

*Equating d/dλ (log L(λ)) = 0*
$\sum y\ (1/\lambda) = \sum x$
$\lambda = \sum y / \sum x$

*d²/dλ² (log L(λ))*
$= \sum y\ (-1/\lambda^2)$
$= -\sum y / \lambda^2$ ---------- *this is indicative of maxima*

*Thus,* $\hat{\lambda}_{MLE} = \frac{\sum y}{\sum x}$. *From part (i), we know that* $\hat{\lambda}_{MOM} = \frac{\sum y}{\sum x}$ *but from part Q8, we know* $\lambda_{LSE} = \sum xy / \sum x^2$. *Hence* $\hat{\lambda}_{MLE} = \hat{\lambda}_{MOM}$ **[3]**

**Solution 10:** Correct Answer is **Option D.**

$\lambda_{LSE}$
$= \sum xy / \sum x^2$
$= 5100/46850$
$= 0.1089$ **[1]**

**Solution 11:** Correct Answer is **Option D.** **[3]**

*$E(s^2(\theta))$*
*= Mean (Variances for individual categories)*
*= (39.3 + 17.3 + 215.5 + 132.7) / 4*
*= 101.20*

*$Var(m(\theta))$*
*= Variance(Means for individual categories) – 1/n * $E(s^2(\theta))$*
*= 110.57 – 1/5 * 101.20*
*= 90.33*

*n = 5*

*$Z_A$*
*= n / (n + $E(s^2(\theta))$ / $Var(m(\theta))$)*
*= 5 / (5 + 101.20 / 90.33)*
*= 0.8169*

**Solution 12:** Correct Answer is **Option B.**

*Credibility Premium for adults category*
*= Z * (Category Mean) + (1 – Z) * (Overall Mean)*
*= 0.8169 * (68.40) + (1 – 0.8169) * (50.4 + 68.4 + 74.0 + 58.2)/4*
*= 67.3655*
*= 67.37* **[2]**

**Solution 13:** Correct Answer is **Option D.**

*Option A is true as teenagers and adults show a positive correlation of 0.63.*

*Option B is true as seniors and adults show a correlation of 0.02 which is very close to 0.*

*Option C is true as non-divorcee adults and divorcee adults have a negative correlation of -0.51*

*Option D is false as the correlation between divorcees and seniors is 0.33. This is much higher than 0.13 i.e. the correlation between seniors and teenagers.* **[1]**

**Solution 14:** Correct Answer is **Option C.**

*Value of t test statistic*
*= $r\sqrt{(n-2)} / (\sqrt{1-r^2})$*
*= $0.12 * \sqrt{3} / (\sqrt{1-0.12^2})$*
*= 0.209359*
*= 0.21* **[2]**

**Solution 15:** Correct Answer is **Option C.**

*Degrees of freedom for the t distribution = 5 – 2 = 3.*

*From Tables, $P(t_3 > 0.2767) = 40\%$.*

*p-value*
*= $P(t_3 > 0.21)$*
*> 40%* **[2]**

*As this p-value is much higher than 5%, we do not have strong evidence to reject the null hypothesis and hence conclude that $\rho = 0$. So, dating behaviour of teenagers and divorcee adults is uncorrelated.*

**Solution 16:** Correct Answer is **Option A.**

*Minimum of two pairs of ($X_i$, $\hat{Y}_i$, $e_i$) would be required to derive the fitted regression equation of Y on X. This can be done by solving the two simultaneous equations in X and $\hat{Y}$ and thus the values of $\hat{\alpha}$ and $\hat{\beta}$ can be found out.* **[2]**

**Solution 17:** Correct Answer is **Option D.**

*Let us take any two pairs and solve the equations simultaneously to arrive at the Y on X regression equation. Let us consider points A (10,40) and I (30, 140).*

$40 = \alpha + 10\beta$
$140 = \alpha + 30\beta$

$100 = 20\ \beta$
$\beta = 5$

$\alpha = 40 - 10*5 = -10$

*So, fitted equation of Y on X is: Y = -10 + 5X.* **[2]**

**Solution 18:** Correct Answer is **Option D.**

*Since the regression line of Y on X passes through the point $(\bar{X}, \bar{Y})$, among these 10 points A to J, only that point which passes through the regression line can be a possible candidate for the mean of X and Y. Only Point H passes through the regression line as seen in the plot and hence it is possible that $\bar{X} = 45.00$ and $\bar{Y} = 215.00$.* **[2]**

**Solution 19:** Correct Answer is **Option C.**

*Adjusted $R^2$*
*$= 1 - (n - 1) / (n - k - 1) * (1 - R^2)$*
*$= 1 - (100 - 1) / (100 - 1 - 1) * (1 - 0.70)$*
*$= 0.696939$*
*$= 69.69\%$* **[2]**

**Solution 20:** Correct Answer is **Option B.**

*We need to compare the observed value of F-statistic with $F_{1,98}$ at 1% level of significance. From tables, $F_{1,60} = 7.077$ and $F_{1,120} = 6.851$. So, value of $F_{1,98}$ lies between 6.851 and 7.077. Since, the observed value 229 is much higher than $F_{1,98}$ we have sufficient evidence to reject the null hypothesis (i.e. $\beta = 0$). Hence, we conclude that $\beta \neq 0$.* **[2]**

**Solution 21:** **i)** E(Y | X = "H")
= 1 * (2/9)/(3/9) + 0 * (1/9)/(3/9)
= 2/3 (1)

**ii)** $E^2$(Y | X = "H")
= $1^2$ * (2/9)/(3/9) + $0^2$ * (1/9)/(3/9) (1)

= 2/3

Var(Y | X = "H")
= $E^2$(Y | X = "H") – [E(Y | X = "H")]$^2$
= 2/3 – (2/3)$^2$
=6/9 – 4/9
= 2/9 (1)
**(2)**

**iii)** Marginal distribution of X:

P(X = 1) = 2/9 + 1/9 = 1/3
P(X = 0) = 4/9 + 2/9 = 2/3 (1)

Marginal distribution of Y:

P(Y = 1) = 2/9 + 4/9 = 2/3
P(Y = 0) = 1/9 + 2/9 = 1/3 (1)
**(2)**

**iv)**
E(Y) = 1 * 2/3 + 0 * 1/3 = 2/3 (0.5)

$E^2$(Y) = $1^2$ * 2/3 + $0^2$ * 1/3 = 2/3

Var(Y)
= $E^2$(Y) – [E(Y)]$^2$
= 2/3 – (2/3)$^2$
= 2/9 (0.5)
**(1)**

**v)**
As P(X = 0) > P(X = 1) (i.e. 2/3 > 1/3), first coin is biased in the favour of Tails (T).
As P(Y = 1) > P(Y = 0) (i.e. 2/3 > 1/3), second coin is biased in the favour of Head (H) (1)

**vi)**
P(X = 1) * P(Y = 1) = 1/3 * 2/3 = 2/9 = P(X = 1,Y = 1) = P(X = "H", Y = "H")
P(X = 0) * P(Y = 1) = 2/3 * 2/3 = 4/9 = P(X = 0,Y = 1) = P(X = "T", Y = "H")
P(X = 1) * P(Y = 0) = 1/3 * 1/3 = 1/9 = P(X = 1, Y = 0) = P(X = "H", Y = "T")
P(X = 0) * P(Y = 0) = 1/3 * 2/3 = 2/9 = P(X = 0, X = 0) = P(X = "T", Y = "T")

Hence random variables X and Y can be considered independent of each other. (2)

As X and Y are independent, the conditional expectation of Y will be equal to the unconditional expectation of Y and the conditional variance of Y will be equal to the unconditional variance of Y.

E(Y | X = "H") = E(Y)
Var(Y | X = "H") = Var(Y)

From parts (i), (ii) and (iv), it can be validated that:
E(Y | X = "H") = E(Y) = 2/3
Var(Y | X = "H") = Var(Y) = 2/9 (1)
**(3)**
**[10]**

**Solution 22:** **i)** X_bar ~ N(20, 5^2 / 5)

P(X_bar < 15)
= P(Z < (15 – 20) / (5 / sqrt(5))
= P(Z < -2.236)
= 1 – P(Z > 2.236)
= 0.0127 (2)

**ii)** For a normal population,

(n-1) * S^2 / sigma^2 ~ chi_square with (n-1) degrees of freedom.

P(S > 6.65)
= P(S^2 > 6.65^2)
= P(4 * S^2 / 5^2 > (4 * 6.65^2) / 25)
= P(chi_square(4) > 7.0756)
= 0.1320 (2)

**iii)** S^2 and X_bar are independent if we are sampling from a normal distribution. (1)

P(X_bar < 15 and S > 6.65)
= P(X_bar < 15 and S^2 > 6.65^2)
= 0.0127 * 0.1320
= 0.001676 (1)
**(2)**

**iv)** E( (n – 1) *S^2 / sigma^2) ) = (n – 1)

E(S^2) = (n – 1) / (n – 1) * sigma^2 = sigma^2 = 25 (1)

Var ((n – 1) *S^2 / sigma^2) ) = 2 (n – 1)

Var (S^2)
= 2 (n – 1) / (n – 1)^2 * sigma^4
= 2 * sigma^4 / (n – 1)
= 2 * 625 / 4
= 312.50 (2)

If the value of n is increased to 100 instead of 5, E(S^2) will remain unchanged at 25, but Var(S^2) will be reduced from 312.50 to 12.6263 (1250/99). This means that for a higher sample size i.e. as n tends to infinity, the sample variance tends to be closer and closer to the population variance (sigma^2) as the variance of the sample variance i.e. Var(S^2) becomes smaller and smaller. (1)
**(4)**
**[10]**

**Solution 23:** **i)** From the scatter plots it can be observed that –

- the relationship between Y and X is non-linear (shape of a curve)
- the relationship between W=log(Y) and X seems to be linear

Hence, logarithmic transformation is justified in this case in order to fit a linear regression model to the data. (2)

**ii)**
S(xw) (1)

= Sum(xw) – (Sum(x) * Sum(w))/n
= -1246.7879 – (588 * - 17.0568)/8
= 6.8869

S(xx)
= Sum(x^2) – ((Sum(x))^2) / n
= 43260 – ((588)^2) / 8
= 42 (1)

beta_hat
= S(xw) / S(xx)
= 6.8869 / 42
= 0.1640 (1)

x_bar = 588/8 = 73.50
w_bar = -17.0568/8 = -2.1321 (1)

The regression line of W on X passes through the point (x_bar, w_bar)

alpha_hat
= w_bar – beta_hat * x_bar
= -2.1321 – 0.1640 * 73.50
= -14.1842 (1)

Least squares fit regression line of W on X is –
w_hat = -14.1842 + 0.1640 * x (1)
**(6)**

**iii)** For the year 2018,

w_hat
= -14.1842 + 0.1640 * 72
= -2.3762 (1)

w = log(y)
y_hat
= exp(w_hat)
= exp (-2.3762)

y_hat
= 0.0929 (0.5)

n_hat
= E * y_hat
= 454 * 0.0929
= 42.1766 (0.5)
**(2)**
**[10]**

**Solution 24:** **i)**

**Step1:** Prior density of μ

f(μ) = 1 / (**10** – **0**) = 1/10 = constant. **So, A = 10 and B = 0**

(5)

**Step 2:** Likelihood function of X ignoring any coefficient of proportionality:

$L(x, \mu) = \text{constant} * e^{\left\{\frac{-1}{2\mathbf{sigma\text{\^{}}2}} \times \sum(x - \mathbf{mu})^2\right\}}$. **So, C = sigma^2 and D = mu**

**Step 3:** Arriving at the posterior density of μ:

$f(\mu \mid x)$

$= f(\mu) * L(x, \mu)$

$= \text{constant} * e^{\left\{\frac{-1}{2\mathbf{sigma\text{\^{}}2}} \times \sum(x - \mathbf{mu})^2\right\}}$

$= \text{constant} * e^{\left\{\frac{-1}{2\mathbf{sigma\text{\^{}}2}} \times (\sum x^2 - 2\mu \sum x + n\mu^2)\right\}}$

$= \text{constant} * e^{\left\{\frac{-1}{2\mathbf{sigma\text{\^{}}2}} \times (-2\mu \sum x + \mathbf{n} * \mathbf{mu\text{\^{}}2})\right\}}$ **So, E = n * mu^2**

$= \text{constant} * e^{\left\{\frac{-1}{\frac{2\sigma^2}{n}} \times (-2\mu\bar{x} + \mu^2)\right\}}$ **So, F = n**

$= \text{constant} * e^{\left\{\frac{-1}{\frac{2\sigma^2}{n}} \times (\mu^2 - 2\mu\bar{x} + \mathbf{x\_bar\text{\^{}}2})\right\}}$ **So G = x_bar^2**

$= \text{constant} * e^{\left\{\frac{-1}{2(\mathbf{sigma\text{\^{}}2} / \mathbf{n})} \times (\mu - \mathbf{x\_bar})^2\right\}}$ **So H = sigma^2 / n and I = x_bar**

Hence, the posterior density of μ is as follows:

$\mu \sim N(5, \sigma^2 / \mathbf{n})$. **So J = n = 150**

**ii)** For a normal distribution, mean = median = mode.
For mu, mean = median = mode = 5 (1)

So, Bayesian Estimate under –
(a) Squared error loss (mean) = 5
(b) All-or-nothing loss (mode) = 5
(c) Absolute error loss (median) = 5 (1)
**(2)**

**iii)** From tables we know that P(-1.96 < Z < 1.96) = 0.95

95% equal tailed credible interval for mu
= (5 – 1.96 * sqrt(25/150), 5 + 1.96 * sqrt(25/150))
= (4.1998, 5.8002) (2)

**iv)** For symmetrical distributions like normal distribution, equal tailed credible interval is equal to highest posterior density interval as the minimum density of any point within this interval is higher than any density outside this interval.

So, highest posterior density interval is (4.1998, 5.8002) (1)
**[10]**

**Solution 25:** **i)** Examples of non-parametric approaches to hypothesis testing (any **two** names are sufficient):

- Permutations approach
- Chi-square goodness of fit tests
- Contingency tables
- Fisher's Exact Test (1)

**ii)** (1)

H0: Passing actuarial examinations is independent of region
H1: Passing actuarial examinations is not independent of region

**iii)**

XP(o) = Candidates from Region X who passed = 2050 * 52% = 1066
XF(o) = Candidates from Region X who failed = 2050 – 1066 = 984
YP(o) = Candidates from Region Y who passed = 800 * 40% = 320
YF(o) = Candidates from Region Y who failed = 800 – 320 = 480. (1)

**iv)**

Total passed candidates (P) = 1066+320 = 1386
Total failed candidates (F) = 984 + 480 = 1464
Total Candidates from Region X (B) = 2050
Total Candidates from Region Y (G) = 800
Total Candidates (T) = 2850

XP(e) = Candidates from Region X expected to pass = 2050 * 1386 / 2850 = 996.9474 = 997
XF(e) = Candidates from Region X expected to fail = 2050 * 1464 / 2850 = 1053.053 = 1053
YP(e) = Candidates from Region Y expected to pass = 800 * 1386 / 2850 = 389.0526 = 389
YF(e) = Candidates from Region Y expected to fail = 800 * 1464 / 2850 = 410.9474 = 411 (2)

**v)**

Test statistic for chi-square test
= (1066 – 997)^2 / 997 + (984 – 1053)^2 / 1053 + (320 – 389)^2 / 389 + (480 – 411)^2 / 411
= 33.1197 (1)

Number of degrees of freedom
= (2-1) * (2-1) = 1

We are carrying out a one-sided test. The upper 0.5% point of a chi_square distribution with 1 degree of freedom is 7.879. (1)

As the observed value of the test statistic (33.1197) is in excess of 7.879, we have sufficient evidence to reject the null hypothesis at 0.5% level of significance. Therefore, it is reasonable to conclude that passing actuarial examinations is dependent on region to which the candidate belongs (candidates from Region X seem to be better at passing these examinations as compared to candidates from Region Y) (1)
**(3)**

**vi)** The observation raised by Mrs. Numara is valid as it is clear from the table that in every subject the pass percentage of Region Y is higher than that of Region X.

However, in simpler subjects like CS1, CM1, CB1, CB2 where the average pass percentage is higher, the number of candidates from Region Y appearing for these papers is relatively very low. Whereas for difficult subjects like CM2 and CS2 where the average pass percentage is very low, the number of candidates from Region Y appearing for these papers is relatively very high.

For Region X, it is exactly the opposite. Candidates from Region X seem to have attempted simpler papers in large numbers and their participation in difficult papers is relatively low. (2)

So, in case of Region X despite of an average performance across papers, due to higher number of candidates attempting simpler papers, the overall pass percentage of Region X tends to be higher.

On the contrary in case of Region Y, despite of performing above average across all papers, due to lower number of candidates attempting simpler papers, the overall pass percentage of Region Y tends to be lower.

However, there is no error in the test performed using contingency tables. With more insights into the underlying data, a better picture has emerged.

**[10]**

**Solution 26:** **i)** For checking whether a parameter is significant (i.e. significantly different from zero), as a general rule we use –

mod(beta) > 2 * se(beta) (0.5)

mod(beta_w_i=0) = 0.036
2 * se(beta_w_i=0) = 0.046 > 0.036
beta_w_i=0 is not significant (0.5)

mod(beta_w_i=1) = 0.100
2 * se(beta_w_i=1) = 0.160 > 0.100
beta_w_i=1 is not significant (0.5)

mod(beta_k) = 0.003
2 * se(beta_k) = 0.002 < 0.003
beta_k is significant (0.5)
**(2)**

**ii)** In case of a weekend travel over a distance of 300 kms
g(pc)
$= \alpha + \beta_{w\ i=1} + \beta_k * kms$
= -0.372 – 0.100 – 0.003*300
= -1.372 (1)

From Tables, canonical link function for a binomial model is given by:
g(pc) = log (pc / (1-pc))
-1.372 = log (pc / (1-pc))
exp(-1.372) = pc / (1-pc)
0.2536 = pc / (1-pc)
0.2536 = (1+0.2536) pc
pc = 0.20 as required (2)
**(3)**

**iii)**
We are using a binomial distribution with –
n = 4

success = travelling by train (due to confirmation of railway ticket)
p = pc = 0.20

failure = travelling by private bus (due to non-confirmation of railway ticket)
q = 1 – p = 0.80 (1)

Required probability
= P(X>=2)
= 1 – P(X<= 1)
= 1 – P(X=1) – P(X=0)
= 1 – 0.4096 – 0.4096
= 0.1808 (2)
**(3)**

**iv)** If we add an interaction term (wknd * kms), the revised linear predictor would be as follows:

g(pc) = alpha + beta_w_i=0,1 + beta_k * kms + beta_wk_i=0,1 * kms (1)

where:

1. beta_w_i=0 is used in case of a weekday and beta_w_i=1 is used in case of a weekend
2. beta_wk_i=0 is used in case of a weekday and beta_wk_i=1 is used in case of a weekend (1)

**(2)**
**[10]**

**********