Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper A)

July 2022 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)

1. Developing objectives to be met by the results of the data analysis
2. Identifying the data items required for the analysis
3. Collecting the data
4. Processing and formatting the data
5. Cleaning the data
6. Exploratory data analysis
7. Modelling
8. Communicating the results
9. Monitoring the process (updating the data & repeating the process if required)

[4]

ii) B [1]

Explanation:
In classical statistics, μ is a fixed quantity, and therefore cannot have a probability distribution associated with it.

However, in Bayesian statistics, μ is a random variable, and therefore statements about its probability can be made

iii) A – III
B – IV
C – II
D – I

[2]
[7 Marks]

Solution 2:

i) The number of claims incurred by each policyholder follows the poisson distribution with mean 0.03. Therefore X, the number of claims for the 100 policyholders follows the Poi(3), X ~ Poi(3).

Since the poisson distribution only takes integer value P(X<6) = P(X<=5)
Using the poisson cumulative probability tables gives 0.91608

[2]

ii) Counting the numbers of trials up to and including the 4th success. This describes the variable (X) is Type 1 negative binomial distribution with k= 4 and p = 0.4

P(X=x) = $\binom{x-1}{3} 0.4^4 0.6^{x-4}$ x = 4,5,6,….

So P(X <7) = P(X=4) + P(X=5) + P(X=6)

P(X=4) = $\binom{3}{3} 0.4^4$ = 0.0256

Now using the iterative formula P(X=x) = $\frac{x-1}{x-4}$q P(X=x-1)

P(X=5) = $\frac{4}{1}$ ×0.6 × 0.0256 = 0.06144
P(X=6) = $\frac{5}{2}$ ×0.6 × 0.06144 = 0.09216

Hence, P(X <7) = 0.0256 + 0.06144 + 0.09216 = 0.1792

[2]

iii) Here the variable(X) is binomial distribution with n = 1000 and p = 0.015
Since n is large and p is small, hence poisson approximation can be used

Bin(1000,0.015) ~ Poi(15) (approximately)

Using the cumulative Poisson table gives

P(X <10) = P(X <=9) = 0.06985 [2]

[6 Marks]

Solution 3:

i) Correct Answer (B) [2]

$E(e^{tX}) = \frac{1}{\mu}(\frac{1}{\mu} - t)^{-1} \int_0^{\infty} e^{-z} . dz$

$E(e^{tX}) = (1 - t\mu)^{-1}$

ii) Total time Y of time periods of N policies will be $Y = X_1 + X_2 + \ldots\ldots + X_N$

MGF of Y is given by $E(e^{tY}) = E(e^{t\sum X_i}) = \prod_1^N E(e^{tX})$

$M_Y(t) = (1 - t\mu)^{-N}$

[3]

iii) The $M_Y(t)$ is of the form of MGF for Gamma distribution
Thus, the distribution is Gamma(N, $\frac{1}{\mu}$) [1]

[6 Marks]

Solution 4: Let X be the amount of fixed benefit health insurance claims and Y the amount of indemnity based health insurance claim.

Then:
$X \sim N(900, 100^2)$ and $Y \sim N(1400, 300^2)$

We require

$P((Y_1+Y_2+Y_3) > (X_1+X_2+X_3+X_4) + 900)$
$= P((Y_1+Y_2+Y_3) - (X_1+X_2+X_3+X_4) > 900)$

So we need the distribution of $(Y_1+Y_2+Y_3) - (X_1+X_2+X_3+X_4)$:

$(Y_1+Y_2+Y_3) - (X_1+X_2+X_3+X_4) \sim N(3\times1400 - 4\times900, 3\times300^2+4\times100^2)$

i.e $(Y_1+Y_2+Y_3) - (X_1+X_2+X_3+X_4) \sim N(600,310000)$

Therefore

$P((Y_1+Y_2+Y_3) - (X_1+X_2+X_3+X_4) > 900)$

$= P(Z > \frac{900-600}{\sqrt{310000}}) = P(Z > 0.54) = 1 - P(Z < 0.54) = 1 - 0.70540 = 0.2946$

[4 Marks]

Solution 5:

i) A group of random variables is said to be independent and identically distributed if the variables are independent of each other and follow the same probability distribution

[1]

ii)

a) As there are 5 coin tosses and the probability of each coin toss being either heads or tails is 0.5, the probability of this exact outcome is 0.5^5 = 0.03125 [1]

b) p-value is the probability of an observation at least as "extreme" as the actual observation. Under the null hypothesis, the expected number of heads is 2.5, while the actual number of heads is 4 (> 2.5). Thus, we need to calculate the probability of 4 or 5 heads.
Let the number of heads be X. Then X ~ Bin (5, 0.5)
Prob (X >= 4) = 1 – Prob (X <= 3) = 1 – 0.8125 *(from tables)* = 0.19 approx
As the p-value is 0.19 > 0.05, the null hypothesis cannot be rejected at 5% significance level
[3]

iii)

a) Likelihood can be calculated as:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\ \theta)$$

which yields

$$L(\theta) = \ \theta^4\ (1-\theta)$$

[1]

b) C [3]
Explanation:
Differentiating the log likelihood,

$$\frac{\partial}{\partial\theta}\log L(\theta) = \ \frac{\partial}{\partial\theta}\ [4\log\theta + \log(1-\theta)] = \frac{4}{\theta} - \ \frac{1}{1-\theta}$$

Equating to 0,

$$\frac{4}{\theta} - \ \frac{1}{1-\theta} = 0\ \Rightarrow\ \theta = 4 - 4\theta\ \Rightarrow\ \theta = \ \frac{4}{5} = 0.8$$

Checking for maximum:

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta) = \ \frac{\partial}{\partial\theta}\left[\frac{4}{\theta} + \ \frac{1}{1-\theta}\right] = \ -\frac{4}{\theta^2} - \ \frac{1}{(1-\theta)^2}$$

Substituting θ = 0.8, this works out to -31.25, which is negative. Thus, θ = 0.8 represents the maximum.
The MLE of θ is therefore 0.8.

c) Prior expected value of θ = 0.2*(0.1 + 0.3 + 0.5 + 0.7 + 0.9) = 0.5 [1]

d) As the prior distribution is uniform across nearly the entire possible range of θ (0 to 1), it indicates that we have no knowledge (or very little knowledge) about the value of θ.
[1]

e) C [2]
Explanation:
The posterior expected value would lie somewhere between the prior expected value and the MLE (observed test statistic). Here, the prior EV is 0.5 and the MLE is 0.8.
Thus, the posterior EV would lie between 0.5 and 0.8 – i.e., it would be greater than 0.5.

f) C [2]
Explanation:
The posterior EV would lie between the prior EV and the MLE.
The prior EV is 0.5.
The MLE is simply the proportion of coin tosses that result in "heads" – in this case, also 0.5 (8 tosses, 4 heads, 4 tails).
Since both the prior EV and the MLE are 0.5, the posterior EV must also be exactly 0.5.

g) B [1]

h) D [1]

iv)

a) No – For the chi-squared test, values less than 5 for any expected value are generally not considered. If we try to form a contingency table (as is done in the next sub-question) based on 8 coin tosses, the expected value in each cell would be less than 5

[1]

b) Number of degrees of freedom = (rows – 1) * (columns – 1) = 1 * 1 = 1
Expected values in each cell would be 5 [= row total * column total / table total]

Thus, the squared difference of the actual value in each cell with the expected value is (observed value – 5)^2, i.e. 4, 4, 4, 4.
The χ^2 statistic is therefore 4 * 4/5 = 3.2
For 1 df, the 5% value of χ^2 is 3.841, which is higher than the figure of 3.2 calculated above.

Thus, there is insufficient evidence to reject the null hypothesis (i.e., that each coin toss is independent of the preceding toss) at the 5% level.

[3]
[21 Marks]

Solution 6:

i) The marginal density is
$f_Y(y) = 3\int_0^\infty e^{-x}e^{-3y}dx = 3e^{-3y}\int_0^\infty e^{-x}dx = 3e^{-3y}$ [1]

ii) The conditional probability $P(Y \le y \mid Y > 4)$ is $F_Y(y)$

$P(Y \le y \mid Y > 4) = \frac{P(Y \le y, Y > 4)}{P(Y > 4)} = \frac{P(4 < Y \le y)}{P(Y > 4)} = \frac{P(4 < Y \le y)}{P(Y > 4)} = \frac{FY(y) - FY(4)}{P(Y > 4)}$, $y > 4$
Therefore

$f(y \mid Y > 4) = \frac{fY(y)}{P(Y > 4)} = \frac{3e^{-3y}}{e^{-12}} = 3e^{12-3y}$, $y > 4$ [2]

iii) The correct option is (C) [2]
The conditional expectation is given as

$E[Y \mid Y > 4] = \int_4^\infty y f(y \mid Y > 4)dy = \int_4^\infty 3ye^{12-3y}dy$

By taking t = y-4,

$E[Y \mid Y > 4] = \int_0^\infty 3(t+4)e^{-3t}dt = \int_0^\infty 3te^{-3t}dt + \int_0^\infty 12e^{-3t}dt$

[5 Marks]

Solution 7:

i) The graph appears to show an approximately linear relationship. However, it does appear to have a slight curve and this would warrant closer inspection of the model to see if it is appropriate to the data. [1]

ii) Least squares estimates:
Obtaining the estimates of α and β with $Y = \ln \mu_X$

$S_{XX} = \sum X^2 - n\bar{X}^2 = 15540 - 10(\frac{390}{10})^2 = 330$

$S_{XY} = \sum XY - n\bar{X}\bar{Y} = -2726.66 - 10(\frac{390}{10})(\frac{-70.47}{10}) = 21.67$

$\hat{\beta} = \frac{SXY}{SXX} = \frac{21.67}{330} = 0.0657$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = (\frac{-70.47}{10}) - 0.0657 \times \frac{390}{10} = -9.61$

Therefore, we obtain

B = $e^{\alpha} = 0.000067$

C = $e^{\beta} = 1.07$

[3]

iii) $r = \frac{S_{XY}}{\sqrt{S_{XX}\ S_{YY}}} = .990645$

The correlation coefficient shows a strong positive relationship between the variables force of mortality and age. The positive value of the regression slope parameter $\hat{\beta}$ also suggest the positive correlation between the variables.

[2]

iv) The coefficient of determination is given by

$R^2 = \frac{S_{XY}^2}{S_{XX}S_{YY}} = \frac{21.67^2}{330 \times 1.45} = 98.14\%$

Where $S_{YY} = \sum Y^2 - n\bar{Y}^2 = 1.45$

This says that 98.14% of the variation in the data can be explained by the model and hence indicates an extremely good fit of the model

[2]

v) The completed table of residuals using $\hat{e}_i = y_i - \hat{y}_i$ is:

Age,X	30	32	34	36	38	40	42	44	46	48
Residual, $\hat{e}_i$	0.079	0.028	-0.004	-0.045	-0.087	-0.058	0.001	0.009	0.048	0.036

Age 34: -7.38 – (-9.61 + 0.0657×34) = -0.004
Age 42: -6.85 – (-9.61 + 0.0657×42) = 0.001
Age 48: -6.42 – (-9.61 + 0.0657×48) = -0.036

The residuals should be pattern less when plotted against X, however it is clear to see that some pattern exists – this indicates that the linear model may not be a good fit. [3]

vi) The variance of mean predicted response is:

$\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{XX}}\right\}\hat{\sigma}^2 = \left\{\frac{1}{10} + \frac{(45-39)^2}{330}\right\} \times 0.0034 = 0.00071$

Where $\hat{\sigma}^2 = \frac{1}{8}(1.45 - \frac{21.67^2}{330}) = 0.0034$

The estimate is $Y = \ln \mu_{45} = -9.61 + 0.067 \times 45 = -6.65$

Using the t_8 distributions , a 95% confidence interval for $Y = \ln \mu_{45}$ is

$-6.65 \pm 2.306\sqrt{0.00071} = (-6.71, -6.59)$

The corresponding 95% confidence interval for μ_{45} is (0.001219, 0.001374)

[4]

vii) The width of the interval is only affected by the variance of the mean predicted response. Which depends on the value of $(X_0 - \bar{X})^2$. This term will now be smaller as the new X_0 = 41 value is closer to $\bar{X}$than X_0 = 45. Therefore the interval will be narrower. [2]

[17 Marks]

Solution 8:

i)

Period 1:

E(X) = (10+30)/2 = 20

S_{XX} = (10-20)^2 + (30-20)^2 = (-10)^2 + (10)^2 = 200

E(Y) = (10+20)/2 = 15

S_{YY} = (10-15)^2 + (20-15)^2 = (-5)^2 + 5^2 = 50

S_{XY} = (-10 * -5) + (10 * 5) = 100

Correlation = S_{XY} / √ (S_{XX} * S_{XY}) = 100 / √ (200 * 50) = 1

Period 2:

E(X) = (10+30)/2 = 20

S_{XX} = 200, as above

E(Y) = (10+15)/2 = 12.5

S_{YY} = (10-12.5)^2 + (15-12.5)^2 = (-2.5)^2 + 2.5^2 = 12.5

S_{XY} = (-10 * -2.5) + (10 * 2.5) = 50

Correlation = 50 / √ (200 * 12.5) = 1

[5]

ii) In the period 1 base, the figures of X & Y are: (10, 10); (30, 20); (20, 4); (60, 6).

E(X) = (10 + 30 + 20 + 60) / 4 = 30

S_{XX} = (-20)^2 + 0^ 2 + (-10)^2 + 30^2 = 1400

E(Y) = (10 + 20 + 4 + 6)/4 = 10

S_{YY} = 0^2 + 10^2 + (-6)^2 + (-4)^2 = 152

S_{XY} = (-20 * 0) + (0 * 10) + (-10 * -6) + (30 * -4) = -60

Correlation = -60 / √(1400*152) = -0.13

[4]

iii) In part (a), the correlation between X and Y was calculated separately for each period, and they appeared to be perfectly positively correlated.

However, on combining the periods in part (b), X and Y turn out to be (weakly) negatively correlated. Thus, while the friend's assumption of strong positive correlation may be valid for some periods, overall, the correlation between the two indices / industries X & Y appears to be very weak and negative. As such, the portfolio is diversified.

[2]

[11 Marks]

Solution 9:

i) Sample mean = (16.4 + 17.3 + 16.7) / 3 = 16.8
Prior mean = A/B = 15/1 = 15 *(formula from Tables)*
Credibility factor Z = 3/(3+1) = 0.75
Credibility estimate = Z * 16.8 + (1-Z) * 15 = 16.35

[3]

ii) The variance of the gamma distribution is mean / B. Reducing the B parameter while keeping the mean constant increases the variance, reflecting greater uncertainty. [1]

iii) Revised credibility factor Z = 3/3.2 = 0.94 (approx.)
Revised credibility estimate = 0.94 * 16.8 + 0.06* 15 = 16.69 [1]

[5 Marks]

Solution 10:

i) B

Explanation:
Likelihood is the probability of the exact outcome observed. As x policies have resulted in a claim and 1-x have not, and the policies are independent, the probability is given by the product: $q^x(1-q)^{n-x}$

The ${}^nC_x$ factor is not relevant here as for each policy, we know whether there has been a claim or not. [1]

ii) Let X be the number of claims. Then X ~ Bin (n, q), with mean nq and variance nq(1-q).

By central limit theorem, $\hat{q}$ = X/n approximately follows $N(\mu, S)$, where:

μ = x/n = 3 / 10,000 = 0.3 per mille = 3 * 10^-4

S = $\hat{q} * (1-\hat{q})$
= 3 * 10^-4 * 0.9997 = 3 * 10^-4 (approx.) = 0.3 per mille

$1.96 * \sqrt{\frac{S}{n}} = 1.96 * \sqrt{\frac{3*10^{-4}}{10,000}}$= 0.34 per mille (approx.)

The 95% confidence interval is $\hat{q} \pm 1.96 * \sqrt{\frac{S}{n}}$

Plugging in the values calculated above, and noting that $\hat{q}$ can't be negative,

95% confidence interval: (0 per mille, 0.64 per mille) [4]

iii) As 0.2 per mille falls within the 95% confidence interval, there is insufficient evidence to reject the null hypothesis q at p = 5%. [1]

[6 Marks]

Solution 11:

i) The PF of Z is

$$f(z) = \binom{n}{z}\mu^z(1-\mu)^{(n-z)}$$

The PF function of Y can be obtained by replacing z with ny :

$$f(y) = \binom{n}{ny}\mu^{ny}(1-\mu)^{(n-ny)}$$

This can be written as :

$$f(y) = \exp\left\{\ln\binom{n}{ny} + ny\ln\mu + n\ln(1-\mu) - ny\ln(1-\mu)\right\}$$
$$= \exp\left\{ny\ln\left(\frac{\mu}{1-\mu}\right) + n\ln(1-\mu) + \ln\binom{n}{ny}\right\}$$
$$= \exp\left\{\frac{y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu)}{1/n} + \ln\binom{n}{ny}\right\}$$

Comparing this to the generalized form of exponential family of distributions:

$\theta = \ln\left(\frac{\mu}{1-\mu}\right)$. Rearranging this gives $\mu = \frac{e^\theta}{1+e^\theta}$

b(θ) = - ln$(1-\mu)$ = - ln(1-$\frac{e^{\theta}}{1+e^{\theta}}$) = - ln($\frac{1}{1+e^{\theta}}$) = ln $(1+e^{\theta})$

$\varphi = n$,

a(φ) = $\frac{1}{\varphi}$

c(y, φ) = $\ln\binom{n}{ny}$ = $\ln\binom{\varphi}{\varphi y}$]

[4]

ii) Using the properties of exponential distributions

E(Y) = $b^{/}(\theta)$ = $\frac{d}{d\theta}$(ln $(1+e^{\theta})$) = $\frac{e^{\theta}}{1+e^{\theta}}$ = μ

V(Y) = a(φ) $b^{//}(\theta)$ = $\frac{e^{\theta}(1+e^{\theta})-e^{\theta}e^{\theta}}{n(1+e^{\theta})^2}$ = $\frac{e^{\theta}}{n(1+e^{\theta})^2}$

Substituting $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$

(

V(Y) = $\frac{\frac{\mu}{1-\mu}}{n(1+\frac{\mu}{1-\mu})^2}$ = $\frac{\mu}{n(1-\mu)}(1-\mu)^2$ = $\mu(1-\mu)$/n

[3]

iii) Using the model output, we can see that

$\beta_1 > 2 \times standard\ error(\beta)$

i.e 0.5459 > 2 × 0.08352 = 0.16704

Since
$\beta_1 > 2 \times standard\ error(\beta)$, it can be concluded that the parameter β_1 for the variable "no. of assignment" is significant in the model.

[2]

iv) Using binomial canonical link function,

$\eta(\mu)$ = $\ln\left(\frac{\mu}{1-\mu}\right)$ = $\alpha_i + \beta_1 N + \beta_2$S

So for α_Y = - 1.501 , β_1 = 0.5459, β_2 = 0.0251 and N = 4, S = 65

$\ln\left(\frac{\mu}{1-\mu}\right)$ = - 1.501 + 0.5459 × 4 + 0.0251 × 65 = 2.3141

μ = $(1+e^{-2.3141})^{-1}$ = 91%

Hence probability of passing students in the given scenario is 91%

[3]

[12 Marks]

****