Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper A)

September 2021 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) a) Descriptive analysis: It involves summarising the data or presenting it in a format which highlights any patterns or trends i.e. producing summary statistics like measures of central tendency and dispersion. It describes a data set rather than giving any specific conclusions. (1)

Example (any one point should also fetch marks):

- Calculating mean and standard deviation of number of motor claims in a day
- Plotting graphs on the average rainfall every month to illustrate the months with heaviest rainfall (1)

b) Inferential analysis: This involves estimating the summary parameters of a population based on the sample data set under consideration and testing hypotheses. (1)

Example (any one point should also fetch marks):

- Any example of hypothesis testing (people visit malls more on weekend than on a weekday
- Rate of health claims in India is same as health claims made in Tier I cities (1)

c) Predictive analysis: This extends the principle behind inferential analysis in order for the user to analyse the past data and make predictions about the future event. (1)

Example (any one point should also fetch marks):

- Predicting the number of lapses that will happen in the future years' basis the number of lapses in the last 1 year
- Forecasting the number of customers who would move to using electric vehicles (1)

[6]

ii) C (1)

[7 Marks]

Solution 2:

Let X_i (i = 1,2,3) denote the number of hospitalisations in the month of October, November and December respectively. (0.5)

From the information provided, $X_1 \sim Poi\ (2)$, $X_2 \sim Poi\ (3)$ and $X_3 \sim Poi\ (1)$ (0.5)

Let the total hospitalisation over this period be denoted by X where $X = X_1 + X_2 + X_3$ (1)

Since all X_i's are independent

$X \sim Poi\ (2 + 3 + 1)$ (1)

Thus, $P\ [X < 5] = \sum (6^x e^{-6})/x!$ (summation over x = 0 to 4) (1)

$= 0.0248 + 0.0149 + 0.0446 + 0.0892 + 0.1339$

$= 0.2851$ (1)

[5 Marks]

Solution 3:

i) **a)** Let X_i represent each motor claim amount for i = 1 to 10 (1)

Moment generating function for exponential distribution, $M_x(t) = (1 - t/\lambda)^{-1}$

Hence, for $Y = \sum X_i$ (i = 1 to 10) (1)

$M_Y(t) = (M_X(t))$ ^ 10

$= (1 - t/\lambda)^{-10}$ (0.5)

which is the moment generating function of gamma distribution with α = 10 and λ = 1.25 (0.5)

[3]

b) MGF of 2.5Y is $E[e^{(2.5t)Y}]$ (0.5)

$= M_Y[2.5t]$ (0.5)

$= (1 - 2t)^{-10}$ (0.5)

$= (1 - t/0.5)^{-10}$ (0.5)

which is the moment generating function of gamma (10, 0.5) (0.5)

i.e. χ^2_{20} distribution (0.5)

[3]

ii) B (2)

iii) From i.a. above, Y ~ gamma (10, 1.25) (0.5)
Therefore Y has mean 10/1.25 = 8 and variance = $10/(1.25)^2 = 6.4$ (1)
Applying central limit theorem Y ~ N (8,6.4) (0.5)
Thus, P[Y>10] = P [Z > (10 – 8)/(√6.4) = 0.791] (1)
= 1 – 0.786 (0.5)
= 0.214 (0.5)
[4]

iv) n is not large enough for the central limit theorem to be used, but the approximation is still close to the true probability (1)
[13 Marks]

Solution 4:

E **[4 Marks]**

Solution 5:

i) f(x,y) = (1/27)*(2x + y) where x = 0,1,2 and y = 0,1,2
Joint probability distribution of X, Y i.e f(x,y) is given by the table
f(x,y = 0,0) = 0
f(x,y = 0,1) = 1/27
f(x,y = 0,2) = 2/27
f(x,y = 1,0) = 2/27
f(x,y = 1,1) = 3/27

f(x,y = 1,2) = 4/27
f(x,y = 2,0) = 4/27
f(x,y = 2,1) = 5/27
f(x,y = 2,2) = 6/27

$f_Y(0)$ = 6/27
$f_Y(1)$ = 9/27
$f_Y(2)$ = 12/27 **[3]**

ii) C (2)

[5 Marks]

Solution 6:

i) a) An estimator is said to be consistent when
- mean square error tends to zero
- as 'n' tends to infinity
- where 'n' is sample size (1)

b) A good estimator is one that
- has small mean square error
- is unbiased and
- is consistent (1)

[2]

ii) **a)** probability of finding 2 senior grade employees having ESOPs is 0.3637 (i.e., W3)
W1 and W2 are more extreme scenarios than W3.
Hence, p-value of finding 2 senior grade employees is P(W1) + P(W2) + P(W3)
=0.0606+0.1212+0.3637 =0.5455
(Or p-value of W3 can be found as 1- P(W4) = 1-0.4545 =0.5455) (2)

b) Required probability is 3C3 * 8C2 / 11C5
=1*(8*7/1*2) / ((11*10*9*8*7/1*2*3*4*5)
=4*7/(11*7*3*2)
=2/33
=0.0606 (2)

iii) Option A
(12+18)/ (100+3*80) = 30 / 340 = 0.088 (3)

iv) If 'm' is the mean of the lognormal distribution then by invariance property,
'm cap' = e^('mu cap' + ½ * 'sigma square cap')
=e(1.25) = 3.49

If 'var' is the variance of the lognormal distribution then by invariance property,
'var cap' = e^(2*'mu cap' + 'sigma square cap') * (e^('sigma square cap') -1)
=3.49^2 *(e(0.5)-1)

=12.1825*0.6487
=7.903 (3)

v)

a) CRLB = -1/E[second derivative of log likelihood with respect to lambda]
= 1/E[n/lambda^2]
=lambda^2/n
=0.01/20
=0.0005 (2)

b) 'lambda cap' ~ N(lambda, CRLB) approximately. Hence confidence interval is given by
('lambda cap' -1.96 * sqrt(CRLB), lambda cap' + 1.96 * sqrt (CRLB)
=(0.1-1.96*sqrt(0.0005), 0.1+1.96*sqrt(0.0005))
=(0.056173,0.143827) (2)

c) Using, 2*lambda*n*X bar ~ Chi square distribution with 2*n degrees of freedom
40*lambda*X bar ~ chi square distribution with 40 degrees of freedom
P(24.43<40*lambda*Xbar<59.34) =0.95
Hence 95% confidence interval for lambda is
(24.43/(40*10), 59.34/(40*10))
= (0.061075,0.14835)

Confidence interval using chi square result / exact result is **narrower** (i.e. better)compared to result in part b.
Result in part b is impacted due to **smaller sample size**.
Larger sample could have resulted in **better / narrower** interval in part b (4)

[20 Marks]

Solution 7:

i) Option A (2)

ii) Chi square with 10 df can be written as Ga(5,0.5) distribution.
Hence, posterior distribution would be Ga(5+x, 1.5) using results from part 1 (1)

iii) Option B (1)

iv) Option D (3)
(working is not required)
Bayesian estimate under all-or-nothing loss is mode of the distribution.
Differentiating the log of the posterior distribution and equating with zero, we get,
4/p-14/(1-p) =0
Hence, 4(1-p)-14p=0
4-18p=0. Hence, p=4/18

v) **a)** Posterior distribution of theta can be written as, Normal((A+B)/(C+D),1/(C+D))
Where,
A = (n*x bar)/150^2
B =500/100^2
C =n/150^2
D=1/100^2 (2)

b) Mean of the posterior distribution is (A+B) /(C+D)
This can be written in the forms of
[(A/(x bar)) / (C+ D)]* (x bar) + ((B/500)/ (C+D)) * 500
i.e. (C/(C+D))* (x bar) + (D/(C+D) * 500

i.e., Z* x bar + (1-Z) * 500 which is a credibility estimate

where Z = (n/150^2)/((n/150^2) + (1/100^2))
MLE of 'theta' = x bar
Prior mean = 500 (2)

c) Impact on Z

If prior variance was 150^2 (instead of 100^2) –

- this will lead to reduction in denominator and Z will increase. (0.75)
- Increase in prior variance means that prior is less reliable and hence we need to rely more on data and hence Z will increase. (1.25)

if likelihood variance was 100^2 (instead of 150^2) –

- Z will increase with more increase in numerator compared to denominator. (0.75)
- Reduction in variance of the observed data means data is more reliable and hence more weight can be given to it and hence Z increases. (1.25)

[4]
[15 Marks]

Solution 8:

i) The sums of squares as given in the question

$S_{xx} = \sum(x_i-\bar{x})^2 = 2,800$ $S_{xy} = \sum((x_i-\bar{x})(y_i-\bar{y}) = 25,300$ (0.5)

$\bar{x} = \frac{\sum x_i}{n} = 35$, $\bar{y} = \frac{\sum y_i}{n} = 281$ (1)

$\hat{\beta} = \frac{S_{xy}}{S_x} = \frac{25,300}{2,800} = 9.04$ (1)

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 281 - 9.04 \times 35 = -34.82$ (0.5)

Hence the fitted regression line of y on x is y = -34.82 + 9.04x

[3]

ii) $S_{yy} = \sum(y_i-\bar{y})^2 = 2,70,832$

$\hat{\sigma}^2 = \frac{1}{n-2}(S_{yy} - \frac{S_{xy}^2}{S_{xx}}) = \frac{1}{5}(2,70,832 - \frac{25,300^2}{2,800}) = 8,445.69$ (1)

Now $\frac{5\hat{\sigma}^2}{\sigma^2} \sim$ chi square x_5^2 which gives a confidence interval for σ^2 of:

$(\frac{5\times 8,445.69}{11.07}, \frac{5\times 8,445.69}{1.145}) = (3,814.67, 36,880.74)$ (2)

[3]

iii) The proportion of the variability explained by the model is given by:

$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{25,300^2}{2800\times 2,70,832} = 84\%$ (2)

84% of the variance is explained by the model , which indicate that the fit is fairly good. It is still might be worthwhile to examine the residuals to double check that a linear model is appropriate. (1)

[3]

iv) Testing:

$H_0 : \beta = 0$ vs $H_1 : \beta > 0$

Now $\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_5$ [1]

The observed value of test statistics is

$\frac{9.04 - 0}{\sqrt{8445.63/2800}} = 5.20$ [1]

This exceeds the 0.5% critical value of the t_5 distribution of 4.032. So we have sufficient evidence at the 0.5% level to reject H_0 and the conclusion is that $\beta > 0$ hence the data are positively correlated (1)

[3]

v) The variance of the distribution of the mean number of COVID claims corresponding to an entry age of 60 is:

$[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}]\ \hat{\sigma}^2 = [\frac{1}{7} + \frac{(60 - 35)^2}{2800}] \times 8445.69 = 3,091.72$ (1)

The predicted value of number of COVID claims corresponding to age 60 is

$-34.82 + 9.04 \times 60 = 507.32$ (1)

We have t_5 distribution. Hence the 95% confidence interval is

$507.32 \pm 2.571 \times \sqrt{3091.72} = (364.37, 650.28)$ (2)

[4]

vi) **a)** The completed table of residuals are as follows

Age	5	15	25	35	45	55	65
Residual	91	19	-56	-95	-104	78	67

(2)

b) Clearly the trend of residual with progression of age is not pattern less. The residuals are not independent of the age. This means that the linear model is missing something and is not appropriate to these data (3)

[21 Marks]

Solution 9:

i) a) A distribution of the response variable Y

b) A "linear predictor" η

c) A "link function" g (2)

ii) The PDF of exponential distribution can be written as

f(y) = $\frac{1}{\mu}\, e^{\frac{-y}{\mu}}$ = exp$\left\{-\frac{y}{\mu} - log\mu\right\}$

Comparing the above with the standard PDF of exponential family of distribution

$\theta = -1/\mu$, b(θ) =logμ = -log(-θ) , $\emptyset = 1$, a($\emptyset$) = $\emptyset$ and c(y, $\emptyset$) = 0 (1)

iii)

a) The canonical link function from part (ii) that $\theta = -\frac{1}{\mu}$ (1)

b) The variance function is $b''(\theta)$. Differentiating b(θ) twice , $b''(\theta) = 1/\theta^2 = \mu^2$

So the variance function is μ^2 (1)

c) The dispersion parameter or scale parameter is $\emptyset = 1$ (1)

iv) The log of the likelihood function is

Log L(μ_i) = -$\sum \frac{y_i}{\mu_i}$ - $\sum \log \mu_i$

The canonical link function for the exponential distribution is g(μ_i) = 1/μ_i.

The canonical link function connects the mean response to the linear predictor , g(μ_i) = η_i

Hence we have

$\frac{1}{\mu_i}$ = α + βx_i

The log likelihood function in terms of α and β:

Log L(α, β) = $\sum y_i\,(\alpha + \beta x_i)$ + $\sum \log($ α + β$x_i)$

Differentiating the above equation with respect to α and β:

$\frac{\partial}{\partial\alpha}\log L(\alpha,\beta) = -\sum y_i + \sum \frac{1}{\alpha+\beta x_i}$

$\frac{\partial}{\partial\beta}\log L(\alpha,\beta) = -\sum x_i y_i + \sum \frac{x_i}{\alpha+\beta x_i}$

The equations satisfied by the MLEs of α and β are

$-\sum y_i + \sum \frac{1}{\hat\alpha+\hat\beta x_i} = 0$

$-\sum x_i y_i + \sum \frac{x_i}{\hat\alpha+\hat\beta x_i} = 0$

Substituting in the given data values gives the following equations

$$\frac{1}{\hat{\alpha}+30\hat{\beta}} + \frac{1}{\hat{\alpha}+35\hat{\beta}} + \frac{1}{\hat{\alpha}+40\hat{\beta}} + \frac{1}{\hat{\alpha}+45\hat{\beta}} + \frac{1}{\hat{\alpha}+50\hat{\beta}} - 890 = 0$$

$$\frac{30}{\hat{\alpha}+30\hat{\beta}} + \frac{35}{\hat{\alpha}+35\hat{\beta}} + \frac{40}{\hat{\alpha}+40\hat{\beta}} + \frac{45}{\hat{\alpha}+45\hat{\beta}} + \frac{50}{\hat{\alpha}+50\hat{\beta}} - 39550 = 0$$

(The above is for information purpose only. Students are not expected to provide the above derivation in the answer script) [4]

Correct answer is Option C

[10 Marks]

****