Institute of Actuaries of India

Subject CS1-Paper A – Actuarial Statistics

June 2019 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) Probability distribution of points (P) scored is:

$$= \frac{1}{\binom{52}{3}}\left(\binom{26}{0}\binom{26}{3} \text{ for P} = 0\right)$$

$$\frac{1}{\binom{52}{3}}\left(\binom{26}{1}\binom{26}{2} \text{ for P} = 1\right)$$

$$\frac{1}{\binom{52}{3}}\left(\binom{26}{2}\binom{26}{1} \text{for P} = 2\right)$$

$$\frac{1}{\binom{52}{3}}\left(\binom{26}{3}\binom{26}{0} \text{for P} = 3\right)$$

[1.5 Marks]

MGF of this function is

$$= \frac{1}{\binom{52}{3}}\left(\binom{26}{0}\binom{26}{3} \times \mathrm{e}^{0} + \binom{26}{1}\binom{26}{2} \times \mathrm{e}^{1\mathrm{t}} + \binom{26}{2}\binom{26}{1} \times \mathrm{e}^{2\mathrm{t}} + \binom{26}{3}\binom{26}{0} \times \mathrm{e}^{3\mathrm{t}}\right)$$

[1]

Hence MGF of X where $X=P_1+P_2$ (subscript 1 and 2 refer to the player 1 and player 2 respectively) can be stated as :

$E(e^{tX}) = E\left(e^{t\sum P_i}\right) = \prod E(e^{tP_i})$

[1]

As the draws are independent,

[0.5]

$$\prod E(e^{tP_i}) = \left(E(e^{tP})\right)^2$$

$$= \left(\frac{1}{\binom{52}{3}}\left(\binom{26}{3} + \binom{26}{1}\binom{26}{2} \times \mathrm{e}^{1\mathrm{t}} + \binom{26}{2}\binom{26}{1} \times \mathrm{e}^{2\mathrm{t}} + \binom{26}{3} \times \mathrm{e}^{3\mathrm{t}}\right)\right)^2$$

[1]

[5]

ii) Using MGF, $E(X) = \frac{d(E(e^{tX}))}{dt}\ for\ t = 0$ [1]

$$E(X) = \frac{2}{\binom{52}{3}}\left(\binom{26}{3} + \binom{26}{1}\binom{26}{2} \times \mathrm{e}^{1\mathrm{t}} + \binom{26}{2}\binom{26}{1} \times \mathrm{e}^{2\mathrm{t}} + \binom{26}{3} \times \mathrm{e}^{3\mathrm{t}}\right)^1 \times \left(\binom{26}{1}\binom{26}{2} \times \mathrm{e}^{1\mathrm{t}} + 2\binom{26}{2}\binom{26}{1} \times \mathrm{e}^{2\mathrm{t}} + 3.\binom{26}{3} \times \mathrm{e}^{3\mathrm{t}}\right) for\ t = 0$$

[1]

Hence $E(X) = \frac{2}{\binom{52}{3}}\left(2\binom{26}{3} + 2\binom{26}{1}\binom{26}{2}\right) \times \left(3\binom{26}{2}\binom{26}{1} + 3.\binom{26}{3}\right) = \frac{12}{\binom{52}{3}}\left(\binom{26}{3} + \binom{26}{1}\binom{26}{2}\right)^2$

[1]

[3]

[8 Marks]

Part (i) was not answered correctly by majority of the candidates and hence they could not attempt the Part (ii). Students who could identify the PDF of the event scored highly on the question.

Solution 2:

i)

$$\sum X_{Public} = 270 \quad [0.5]$$

$$(\bar{X}_{Public}) = 30 \quad [0.5]$$

$$\sum X_{Private} = 315.5 \quad [0.5]$$

$$(\bar{X}_{Private}) = 35.06 \quad [0.5]$$

$$\sum X_{Public}^2 = 8614.5 \quad [1]$$

(subscript 1 refers public hospitals)

$$S_1^2 = (8614.5 - 9 \times 30^2)/8 = 64.31 \quad [1]$$

$$\sum X_{Private}^2 = 11599.25 \quad [1]$$

$$S_2^2 = (11599.25 - 9 \times 35.06^2)/8 = 67.42 \quad [1]$$

[**6**]

ii) Two sided t-test can be applied in case the samples come from populations with equal variances.

[1]

We are testing $H_0 : \sigma_1^2 = \sigma_2^2 \; vs \; H_1 : \sigma_1^2 \neq \sigma_2^2$

[1]

Test statistic is $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$

[0.5]

Value of statistic is $\frac{64.31}{67.42} = 0.9542$

[0.5]

$F_{8,8}$ values at 5% levels are 0.2256 and 4.433 . Since the value of the test statistic is between the above values, we have insufficient evidence to reject the hypothesis and conclude that the population variances are equal.

[1]

[**4**]

iii) We are testing $H_0 : \mu_1 = \mu_2 \; vs \; H_1 : \mu_1 < \mu_2$

[0.5]

Test statistic is $\frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{S_P^2(1/n_1 + 1/n_2)}} \sim t_{n_1+n_2-2}$

[1]

Where $S_P^2 = \frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{(n_1+n_2-2)}$ [1]

Using the values in section (I) above,

$$S_P^2 = \frac{64.31\times8+67.42\times8}{16} = 65.86$$

[0.5]

value of test statistic is

$$\frac{(35.06-30)-0}{\sqrt{65.86(1/9+1/9)}} = 1.32$$

[0.5]

$$P(t_{16} > 1.32) = 20.5\%$$

[0.5]

This is higher than 95% hence we do not have sufficient evidence to reject the hypothesis and hence conclude that the cost of claims in private hospitals is similar to that in public hospital

[1]

[5]

[15 Marks]

Part (i) and (iii) of this question was generally well answered by most of the candidates. Many candidates could not attempt the part (ii) correctly.

Solution 3:

i)

$$\begin{aligned} f_X(x) &= \int_0^1 \left(\frac{9}{10}xy^2 + \frac{1}{5}\right) \mathrm{d}y \\ &= \left[\frac{3}{10}xy^3 + \frac{1}{5}y\right]_0^1 \\ &= \frac{3}{10}x + \frac{1}{5} \end{aligned}$$

[0.5 Marks for each step]

$$\begin{aligned} f_Y(y) &= \int_0^2 \left(\frac{9}{10}xy^2 + \frac{1}{5}\right) \mathrm{d}x \\ &= \left[\frac{9}{20}x^2y^2 + \frac{1}{5}x\right]_0^2 \\ &= \frac{9}{5}y^2 + \frac{2}{5} \end{aligned}$$

[0.5 Marks for each step]

[3]

ii)

$$\begin{aligned}
E(X) &= \int_x x f_X(x)\,\mathrm{d}x \\
&= \int_0^2 x\left(\frac{3}{10}x + \frac{1}{5}\right)\,\mathrm{d}x \\
&= \int_0^2 \left(\frac{3}{10}x^2 + \frac{1}{5}x\right)\,\mathrm{d}x \\
&= \left[\frac{1}{10}x^3 + \frac{1}{10}x^2\right]_0^2 \\
&= \frac{8}{10} + \frac{4}{10} - 0 - 0 \\
&= \frac{6}{5} = 1.2
\end{aligned}$$

[2]

$$\begin{aligned}
E(Y) &= \int_y y f_Y(y)\,\mathrm{d}y \\
&= \int_0^1 y\left(\frac{9}{5}y^2 + \frac{2}{5}\right)\,\mathrm{d}y \\
&= \int_0^1 \left(\frac{9}{5}y^3 + \frac{2}{5}y\right)\,\mathrm{d}y \\
&= \left[\frac{9}{20}y^4 + \frac{1}{5}y^2\right]_0^1 \\
&= \frac{9}{20} + \frac{1}{5} \\
&= \frac{13}{20} = 0.65
\end{aligned}$$

[2]

[4]

iii)

$$\begin{aligned}
Cov(X,Y) &= \int_x\int_y xy f_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x - E(X)E(Y) \\
&= \int_0^2\int_0^1 xy\left(\frac{9}{10}xy^2 + \frac{1}{5}\right)\,\mathrm{d}y\,\mathrm{d}x - \left(\frac{6}{5}\right)\left(\frac{13}{20}\right) \\
&= \int_0^2\int_0^1 \left(\frac{9}{10}x^2y^3 + \frac{1}{5}xy\right)\,\mathrm{d}y\,\mathrm{d}x - \frac{39}{50} \\
&= \int_0^2 \left[\frac{9}{40}x^2y^4 + \frac{1}{10}xy^2\right]_0^1\,\mathrm{d}x - \frac{39}{50} \\
&= \int_0^2 \left(\frac{9}{40}x^2 + \frac{1}{10}x\right)\,\mathrm{d}x - \frac{39}{50} \\
&= \left[\frac{3}{40}x^3 + \frac{1}{20}x^2\right]_0^2 - \frac{39}{50} \\
&= \frac{3}{5} + \frac{1}{5} - \frac{39}{50} \\
&= \frac{1}{50} = 0.02
\end{aligned}$$

[2]

[9 Marks]

This question was well answered by most of the candidates. Some candidates lost marks due to incorrect computation or evaluation of the integrals but the concepts were correctly applied by majority of the candidates.

Solution 4:

i) The slope and intercept parameters can be derived as the expected value of β and α in the following equation

$$y = \alpha + \beta x + \in$$ [0.5]

Where, $\in$ are the error terms that are assumed to be identical independently distributed normal random variables.

Under linear regression, α and β can be found by minimizing the squared errors - distance between the observed and predicted values of y.

The mathematical expressions of the expected values of α and β are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

[2]

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x}$$ [0.5]

Using the data given,

Slope = 6.825, intercept = 8.84 [3]

Alternative answer : intercept: 7.65, slope 6.78

****

Splitting the total sum of squares into regression and residual sum of squares:

$SS_{Total} = SS_{Regression} + SS_{Residual}$

Using standard notations; $SS_{Total} = S_{yy}$; $SS_{Regression} = S^2_{XY} / S_{XX}$

[6]

ii) R – squared = $SS_{Regression} / SS_{Total}$ = 130425.8 / 149597.4 = 0.8718

alternate answer : 86.86% **[2]**

iii) Standard error of $\beta = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{3834.34}{2800}} = 1.170216$ [1.5]

The two sided 95% confidence interval for β = $\hat{\beta} \pm t_{0.025,5} * se(\beta)$

i.e. $6.825 \pm 2.571 * 1.170216$ = (3.8164, 9.8336) [1.5]

alternate answer: 6.78 +/- 2.571*1.1782

[3]

iv) ANOVA Table

Source of variation	Degrees of freedom	SS	MSS
Regression	1	130425.8	130425.8
Residual	5	19171.68	3834.336
Total	6	149597.4	

F-test:

H_0: β = 0

F-statistic = 130425.8/ 3834.336 = 34.01521 on (1,5) degrees of freedom [1]

Critical value of F(1,5) = 10.01 [0.5]

Since, F-statistic is greater than the critical value,

so H_0 is rejected at the 2.5%% level. [0.5]

Hence, the slope parameter is statistically significantly different from zero. [1]

alternate answer:

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	129951.4	129951.4	33.07313	0.00223
Residual	5	19646.07	3929.213		
Total	6	149597.4			

Note for markers: At all levels of significance (1%, 2.5%, 5%), critical values for F distribution will be lower than the test statistic value. Marks should be provided for any level of significance used by the student. **[3]**

v) The residual plot is a U-shaped graph and the residuals are observed to follow a pattern [1]

The non-random pattern in the residuals indicates that the deterministic portion of the regression model is not capturing some explanatory information. The possibilities could include:

1. A missing variable
2. A missing of higher order term of a variable in the model to explain the polynomial trend in residuals [2]

From, the above graph it looks likely that including a higher order term of the independent variable should be able to resolve this problem. **[3]**

[17 Marks]

Question (except part (v)) was largely well answered. Computational errors were made by significant number of candidates. In part (v) most of the candidates identified that the distributaion residuals does not seem to be normal but could not provide any additional comments beyond that.

Solution 5:

(i) $\hat{\theta}$ said to be unbiased when $E(\hat{\theta}) = \theta$ [1]

(ii) measure of the 'bias' is given by $E(\hat{\theta}) - \theta$ [1]

(iii) Mean Square Error (MSE) of this estimator $\hat{\theta} = \left(E(\hat{\theta}) - \theta\right)^2$ [1]

(iv) $\tilde{\theta}$ is efficient as an estimator with lower MSE is said to be more efficient than one with higher MSE. [1]

(v) An estimator is termed as consistent if MSE converges to 0 as the sample size tends to ∞ [1]

(vi) θ can be estimated using: [mention any 2 methods, 1 mark each]

a. Method of moments: the population moments are equated to the sample moments to estimate the parameters.

b. Maximum likelihood method: A maximum likelihood function $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$ is generated. A maximum likelihood estimate of the parameter is given by solution to $\frac{dL(\theta)}{d\theta} = 0$

c. Bootstrap method: This is computer intensive method that allows us to avoid making assumption about the sampling distribution by forming an empirical sampling distribution which is possible due to re-sampling based on the available sample.

This was a bookwork question and was well answered. In part (vi) some students provided only the names of the methods without the accompanying narration.

[7 Marks]

Solution 6:

The residuals are based on differences between the observed responses, y, and the fitted responses, $\hat{\mu}$.

Pearson residuals

$$Pearson\ residual = \frac{y-\hat{\mu}}{\sqrt{Var(\hat{\mu})}}$$ [2]

Deviance residuals

$$Deviance\ residual = \ sign(y - \hat{\mu}) * d_i$$

$$where\ d_i\ is\ the\ contribution\ of\ y\ to\ the\ scaled\ deviance\ (\sum d_i^2)$$ [2]

[4 Marks]

This was a bookwork question and was generally well answered. Some students missed on providing the meaning of the terms used.

Solution 7:

If B1, B2, B3,…,Bk constitute a partition of a sample space S and $P(B_i) \neq 0$ for i=1,2,3,….k, then for any event A in S such that $P(A) \neq 0$:

$$P(B_r|A) = \frac{P(A|B_r)P(B_r)}{P(A)}\ where\ P(A) = \ \sum_{i=1}^{k} P(A|B_i) * P(B_i)\ \ for\ r = 1, 2, 3, \ldots . k$$

[1.5]

Derivation:

P (A∩B) = P (A) P (B|A)

On rearranging: $P(B|A) = \frac{P(A \cap B)}{P(A)}$

However, $P(A \cap B) = P(B \cap A) = P(B)P(A|B)$

Now, replacing B by B_r, we have:

$$P(B_r|A) = \frac{P(B_r \cap A)}{P(A)} = \frac{P(B_r)P(A|B_r)}{P(A)}$$

And from the law of total probability:

$P(A) = \sum_i P(A|B_i) * P(B_i)$ [2.5]

[4 Marks]

Many students failed to provide the proof. Majority of the attempts were limited to statement of the Bayes' Theorem

Solution 8:

i) Wickets taken per 500 balls follow Poi(5) distribution. As the number of trials (balls) is very high and poisson parameter >= 5, we can use normal approximation to Poison Distribution. [0.5]

Thus the wickets take approximately follow N(5,5). [0.5]

Hence we need 'X' such that:

$P\left(Z > \frac{X-5}{\sqrt{5}}\right) = 0.95$ [0.5]

Critical value at 95% confidence is 1.65 [0.5]

Thus

$P\left(1.65 > \frac{X-5}{\sqrt{5}}\right) = 0.95$

Hence $X < 8.68$ [0.5]

As number of wickets can only take whole values, we need to truncate the number to lower whole number. Hence the team takes upto 8 wickets at 95 % confidence level. [0.5]

[3]

ii) For team B the runs in 50 ball will follow Bin(50,0.4) [0.5]

The mean and variance for this Binomial distribution are $50 \times 0.4 = 20 \; and \; 50 \times 0.4 \times (1 - 0.4) = 12$ respectively [1]

For large number of trials and probability of success is close to 0.5 (or np>10), normal approximation can be applied & thus the runs per 50 balls follows approximately N(20,12) [1]

Probability of team B scoring 26 or more runs in 50 balls is thus $P\left(Z > \frac{26-20}{\sqrt{12}}\right) = 4.16\%$ [1]

The Poisson rate of taking wickets (by team A) is 1 per 100 balls i.e. 0.01 per ball. Hence, the wickets taken by team A in 50 balls has rate = 0.01 x 50 i.e it follows Poi(0.5) process. [1]

Probability of not taking any wicket is $\frac{0.5^0}{1} * e^{-0.5} = 60.65\%$ [1]

Hence probability of winning is 1-0.6065=39.34% [1]

Thus Team A has higher probability of winning. [0.5]

[7]

iii) Probability that team B bats for 30 balls = (Probability of waiting time (in terms of number of balls)> 30) x (Probability of A not scoring 26 runs in 30 balls) [0.5]

The waiting time has Exp(0.01) distribution, hence P(T>30)=exp(-0.01*30) = 74.08% [1]

Probability of A not scoring 30 runs = $P\left(Z < \frac{26-30\times0.4}{\sqrt{30\times0.4\times0.6}}\right) = 97.41\%$ [1]

Hence there is a 74.08% x 97.41% = 72.16% chance that team B will bat for 30 balls. [0.5]

[3]

[13 Marks]

Most of the students struggled with this question. Not applying CLT, not rounding –off the number of wickets were the common mistakes in part (i). In part (ii) most of the students computed probability of scoring 'exactly' 26 runs and used that in the answer. Only a handful of candidates attempted part (iii)

Solution 9:

i) Prior distribution of μ is Gamma (4,7)

$$f_{prior}(\mu) = \frac{7^4}{\Gamma(4)} * \mu^3 * e^{-7*\mu}$$

Thus $f_{prior}(\mu)\ is\ proportional\ to\ \mu^3 * e^{-7\mu}$ [1]

The likelihood is the product of the Poisson probabilities:

$$L(\mu) = \frac{\mu^{x_1}}{x_1!}e^{-\mu} * \frac{\mu^{x_2}}{x_2!}e^{-\mu} * \ldots\frac{\mu^{x_n}}{x_n!}e^{-\mu}$$

The likelihood function is proportional to

$L(\mu) is\ proportional\ to\ \mu^{\sum x_i} * e^{-n\mu}$ [1]

So, $f_{posterior}(\mu)\ is\ proportional\ to\ \mu^{3+\sum x_i} * e^{-7\mu-n\mu}$

The posterior distribution of μ thus takes the form of a Gamma ($4 + \sum x_i$,7+n). [1]

[3]

ii) (a) Squared error loss

When n=10 and $\sum x_i$ = 15, the posterior distribution of μ is Gamma (19, 17).

The Bayesian estimate of μ under squared error loss is the mean of the posterior distribution. [1]

Bayesian estimate = mean of posterior distribution = mean of Gamma (19, 17) = 19/17 = 1.1176 [1]

(b) All-or-nothing loss

The Bayesian estimate under all-or-nothing loss is the mode of the posterior distribution. [1]

To find the mode we need to differentiate the PDF and equate it to zero.

$f_{posterior}(\mu) = k * \mu^{18} * e^{-17\mu}$ *where k is a constant* [1]

Taking logs and differentiating:

$$\frac{d\ ln_{post}\mu}{d\mu} = \frac{18}{\mu} - 17$$

Equating the derivative to zero will give us the value of μ which maximizes the PDF and thus will give us the mode of the distribution.

μ = 18/17 [0.5]

Differentiating again gives us (-18/μ^2) which is less than zero. This is a check that the prior step gives us the maxima. [0.5]

So, the Bayesian estimate of all-or-nothing loss is **18/17**

(c) Absolute error loss

The Bayesian estimate under absolute error loss is the median of the posterior distribution. [1]

The posterior distribution follows Gamma (19, 17). Let X denote the posterior distribution. Hence X ~ Gamma (19, 17). Then 2*17X ~ Chi squared (2*19). [1]

The median of the posterior distribution is the value of M such that P (X<M) = 0.5
Equivalently, P ($\varkappa^2_{38}$ < 34M) = 0.5
From the tables we can see that the 50th percentile of $\varkappa^2_{38}$ is 37.34:
Hence, M = 37.34/34 = 1.098
So, the Bayesian estimate under absolute error loss is 1.098 [1]

[11 Marks]

Part (i) was answered nicely answered by the well prepared candidates. Students made mistakes in Part (ii) by equting the mean / median / mode to loss measures other than those given in the solution.

Solution 10:

Let Y_{ij} and P_{ij} be the claim amounts and number of employees covered for company i and year j respectively.

Denote $X_{ij} = \frac{Y_{ij}}{P_{ij}}$; N=2, n = 4

$\bar{P}_A = \sum_j \overline{P_{Aj}} = 121 + 119 + 120 + 110 = 470$

$\bar{P}_B = \sum_j \overline{P_{Bj}} = 150 + 135 + 122 + 145 = 552$

$\bar{P} = \overline{P_A} + \overline{P_B} = 1022$

$$P^* = \frac{1}{7} * \left[470 * \left(1 - \frac{470}{1022}\right) + 552 * \left(1 - \frac{552}{1022}\right)\right] = 72.53$$ [1]

Table for claims per unit employees, X_{ij}:

	Year1	Year2	Year3	Year4
Company A	33.75	36.87	37.92	31.42
Company B	30.89	23.73	16.99	30.93

Using the formulae from the tables,
$\overline{X_A} = 35.0745$
$\overline{X_B} = 26.0779$
$\bar{X} = 30.2074$

$$E[m(\theta)] = 30.2074$$ [2]

$$E[s^2(\theta)] = \frac{1}{N} * \sum_{i=1}^{2} \frac{1}{n-1} * \sum_{j=1}^{4} P_{ij}\left(\overline{X_{ij}} - \bar{X}_i\right)^2 = \frac{1}{2} * (1011.036 + 5904.065) = 3457.551$$

[2]

$$Var[m(\theta)] = \frac{1}{72.53} * \left(\frac{1}{7} * 41214.23 - 3457.551\right) = 33.5061$$ [2]

Putting the above derived values in the formulae:

$$z_A = \frac{470}{470 + \frac{3457.551}{33.5061}} = 0.81997$$ [1.5]

$$z_B = \frac{552}{552 + \frac{3457.551}{33.5061}} = 0.842501$$ [1.5]

Using credibility theory, credibility premium per unit risk volume is given by:

$$Company\ A: Z_A * \bar{X}_A + (1 - Z_A) * E[m(\theta)] = 34.1843$$
$$Company\ B: Z_B * \bar{X}_B + (1 - Z_B) * E[m(\theta)] = 26.72829$$

The EBCT claim amounts for the coming year for the two states are:
Company A: 4614.88; Company B: 4142.886 [2]

[12 Marks]

This question was attempted by majority of the students. Coputational errors were made by some while the rest scored highly in this question.

****