Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper A)

March 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)

Institute	Share	Prob(fails to join)		P(Institute i given fails to join)
	X	Y	Xi * Yi	(Xi*Yi)/(Summation of Xi *Yi over A to D)
A	0.2	**0.01**	0.002	**7.41%**
B	0.2	**0.02**	0.004	**14.81%**
C	**0.3**	**0.03**	0.009	**33.33%**
D	**0.3**	**0.04**	0.012	**44.44%**
	1.0		**0.027**	**100.00%**

Correct Share for C and D (0.5 Mark)
Correct Probability for all 4 institutes (0.5 Mark)
Correct Formula (1 Mark)
Correct Calculation / Final Answer for each Institute (1 Mark)

[3]

ii) Only statement 'a' is True. All other statements are incorrect

Corrected version

b – Under absolute error loss, median of posterior distribution minimises the expected loss function

c – Bayesian method assumes parameter to be a random variable

d – classical statistics assumes unknown parameter to be fixed and hence cannot assign probability statements to it.

[2]

iii) Correct Formula and answer –

Mean of the Gamma is alpha /lambda = 48/4 =12 [1]

iv)

Answer Option b

As Ga(a,b) is positively skewed, mean > median > mode. (0.5 mark)

Bayesian estimate under squared error loss is mean equal to 'S'. (0.5 mark)
(calculations not expected as simple logic needs to be applied to solve this 1-mark question)

[1]
[7 Marks]

Solution 2:

i) Factor analysis / Principal Component Analysis is -

A method for **reducing the dimensionality** of data [0.5]

It seeks to **identify key components necessary to model and understand data** [0.5]

Original variables may be

- **correlated** with each other [0.5]

While Newly identified **principal components** are chosen to be

- **uncorrelated** [0.5]
- **linear combinations of the original variables of the data** [0.5]
- **which maximise the variance** [0.5]

[3]

ii)

Principal Component	Diagonal entry (PCi)	PCi/ (Sum(PCi) over 1 to 5)	
PC1	0.456	65.0%	of total variance explained by PC1
PC2	0.137	19.5%	of total variance explained by PC2
PC3	0.08	11.4%	of total variance explained by PC3
PC4	0.0165	2.4%	of total variance explained by PC4
PC5	0.012	1.7%	of total variance explained by PC5
	0.7015	**100.0%**	

Correct formula (1 mark)
Sum of PCi (0.5 marks)
Correct calculation (2.5 Marks)

[4]

iii) As 1st 3 Principal components explain over 95% of total variance, dimensionality can be reduced to 3 for this dataset [1]

The 1st 3 Principal Components can then be used for building further classification or regression modelling purpose [1]

[2]

[9 Marks]

Solution 3:

i)

Let X denotes the sample with 5 values and Y denotes the sample with 17 values
Given that population variances are equal i.e. Sigma X = Sigma Y, [0.5]
Therefore, P (Sx^2 / Sy^2 > 3) = P($F_{4,16}$ > 3) [1]
As upper 5% point of $F_{4,16}$ distribution is 3.007. [0.5]
So, required probability is just over 5% [1]

(As, F4,16 at 10% is 2.333 and F4,16 at 2.5% is 3.729.

Hence required number 3 is between 5% and 10%)

Deduct half mark if **over** 5% is not mentioned / incorrectly mentions under 5% or 5%

[3]

ii) Answer Option d

Nc + Nd = 87 +123 = 210 [0.5]

N (N-1)/ 2 = 210 [0.5]

N^2-N-420 =0

N^2-21N+20N-420 = 0

N(N-21)+20(N-21) =0

(N+20) (N-21) =0

N=-20 / N=21 [0.5]
As N cannot be negative, N=21 [0.5]
[2]

iii) Size of the dataset [0.5]

Speed of arrival of the data [0.5]

Variety of different sources from which the data is drawn [0.5]

Reliability of the data elements might be difficult to ascertain [0.5]
[2]
[7 Marks]

Solution 4:

i) For Chi square –

if degrees of freedom = n then mean =n and variance = 2n [0.5]

Coefficient of variation = Sqrt (2n)/ n = Sqrt(2/n) [0.5]

Hence, as variance increases, coefficient of variation will reduce. [1]

For Poisson -

If Poisson parameter is L then mean = variance = L [0.5]

Coefficient of variation = Sqrt(L)/L = Sqrt(1/L) [0.5]

Hence, as variance L increases, coefficient of variation will reduce. [1]

For Exponential –

If exponential parameter is 1/L then mean =L and variance = L^2 [0.5]

Coefficient of variation = Sqrt (L^2)/L =1 = constant value [0.5]

Hence, Coefficient of variation will have no effect of increase (or any change) in variance [1]
[6]

ii) F(x) = u = 1-e^(-0.5x) [0.5]

1-u = e^(-0.5x)

LN (1-u) = -0.5x

X = -LN (1-u)/0.5 [0.5]

For u = 0.769, x = -2*LN(1-0.769) =2.931 and [0.5]

for u= 0.004, x = -2*LN(1-0.004) = 0.008 [0.5]
[2]
[8 Marks]

Solution 5:

i) P(X=2) = 0.3 + 0.2 + 0 = 0.5

Required expectation is summation of y* P(Y=y| X=2) [1]

= 1*0.3/0.5 + 2*0.2/0.5 +3*0/0.5

=3/5 + 4/5 = 7/5 =1.4 [1]

[2]

ii) For 5X - 4Y,
Mean is 5*E(X) – 4*E(Y) = 5*0 -4*0 =0 [0.5]
Variance is 5^2 * Var (X) + 4^2 * Var (Y) = 25*1 + 16*1 = 41 [1]
Hence required distribution is N(0,41) [0.5]

[2]

iii) (2*Lambda *X) ~ Chi square distribution with (2 *alpha) degrees of freedom [0.5]

P(X>50) = P(2*0.1*X > 2*0.1*50)

= P(Chi square >10) [1]

with 2*10 =20 degrees of freedom [0.5]

Hence,

Required Chi square expression is

P(Chi square > 10) where chi square distribution will have 20 degrees of freedom

Required chi square probability is equal to 1-0.0318 = 0.9682 [0.5]

Hence, probability of X greater than 50 is over 96.8% [0.5]

[3]

iv) E(X) = alpha /lambda = 10/0.1 = 100 [0.5]

Var(X) = alpha /lambda^2 = 10/0.1^2 = 1000 [0.5]

Hence using Central Limit theorem, using normal approximation [0.5]

(P X>50) = ~P(N(100,1000) >50)
~P(Z>(50-100)/sqrt(1000))
~P(Z(N(0,1)>-1.58114))
~Z(N(0,1)<1.58114))

For correct equation as above [1]

From tables

x	phi(x)
1.58	0.94295
1.59	0.94408

Answers using interpolation are accepted though not expected.

There is over 94.3% probability that X is greater than 50

using normal approximation to underlying gamma distribution

[0.5]

[3]

v) Probability calculated using normal distributional assumption is lower as compared to the answer obtained using chi square for the underlying Gamma distribution. [1]

As gamma distribution is positively skewed and has thick tail compared to normal and it will tend to be more like normal only when alpha tends towards infinity. As in this case, value of alpha is only 10, so normal approximation is not truly able to capture the correct thick tail found for Gamma.

[1]

[2]

[12 Marks]

Solution 6:

i) Prior mean (Mu0)=700 Var (sigma0)=70^2 = 4900

observed sample mean (x bar) over 6 (=n) years is 3600 /6 = 600 and Var of distribution (sigma)=100^2 = 10000

Posterior Variance = 1/ [(n/sigma^2) + (1/ (Sigma0^2)] [0.5]

= 1/ /[(6/10000)+(1/4900)] = 1/ 0.000804 =1243.655 [0.5]

posterior mean = [(n* x bar) / sigma^2) + (mu0 /Sigma0^2)] / [(n/sigma^2) + (1/ (Sigma0^2)]

[0.5]

= [(6*600/10000) + (700/4900)] /[(6/10000)+(1/4900)] = (0.36 + 0.142857)/0.000804

=625.3807 [0.5]

Hence posterior distribution of beta is N(625.3807, 1243.655) i.e (625.3807, 35.266^2)

[2]

ii) Required probability is

P(Z>600-625.3807/sqrt(1243.655))
~P(Z<0.7197039)

For correct equation as above [1]

From tables,

X	phi(x)
0.72	0.76424

Correct tabulated values [1]
There is over 76% probability that beta is greater than 600

[2]

iii)

	prior	**likelihood**	**posterior**
mean	700	600	625.38
sd	70	100	35.27

As posterior mean is closer towards likelihood mean, Credibility factor is expected to be greater than 0.5.

If we use Z = 0.75, using simple weighted average, we will get

posterior mean = 0.75 (600) + 0.25(700)

= 450 + 175 = 625

Hence Z is expected to be close to 0.75 [1]

If prior SD had been more than 70 then Credibility Factor Z would have increased (Higher the variance of prior, less reliable the prior belief would be) [1]

If likelihood SD had been lower than 100 then Credibility Factor Z would have increased. (Lower the variance of likelihood more reliable the data would be) [1]

[3]

[7 Marks]

Solution 7:

i) L(p) = constant * p^x*(1-p)^(n-x) [1]
logL(p) = constant + x logp+ (n-x)log(1-p) [1]
Taking derivative w.r.t. p
logL'(p) = x/p – (n-x)/(1-p) [1]
Equating to 0
x(1-p)-(n-x)p = 0
x – xp – np + xp = 0
p = x/n [1]

OR (if instead of "n", 5000 is substituted):

L(p) = constant * p^x*(1-p)^(5000-x) [1]
logL(p) = constant + x logp+ (5000-x)log(1-p) [1]
Taking derivative w.r.t. p
logL'(p) = x/p – (5000-x)/(1-p) [1]
Equating to 0
x(1-p)-(5000-x)p = 0
x – xp – 5000p + xp = 0
p = x/5000

[4]

ii) f(p) = 1/(1-0)
Let posterior distribution of p be denoted by P(p)

P(p) α L(p) * f(p) [1]
P(p) α p^x*(1-p)^(n-x) * 1 [1]
P(p) α p^(x+1-1)*(1-p)^(n-x+1-1)

Therefore, the posterior distribution is beta distribution with parameters x+1, n-x+1

[2]

[4]

OR if instead of "n", 5000 is substituted, posterior distribution would be beta distribution with parameters x+1, 5001-x+

iii) p = 200/500 = 0.4 [1]

iv) Under quadratic loss, the Bayesian estimator is the expectation of the posterior distribution. In this case, p = (200+1)/ (200+1+500-200+1) = 201/502 = 0.4004 [2]

v) The two estimates are almost equal, this is because the impact of prior distribution is very limited and the Bayesian estimator is mainly determined by the actual data [1]

vi) Posterior mean can be written in credibility form as:

p = (x+1)/(n+2) [1]
p = x/(n+2) + 1/(n+2)
= x/n * n/(n+2) + 2/(n+2) * (1/2) [1]
= E(X) * Z + E(p) * (1-Z)
Where E(X) = x/n and E(p) = ½ and Z = n/(n+2) = 500/502 [1]

[3]

[15 Marks]

Solution 8:

i) The scatter plot suggests an inverse relation between marks obtained and hours spent on social media per day [1]

ii) Sxx = 277.5 – 45^2/10 = 75
Syy = 43,956 – 644^2/10 = 2482.4
Sxy = 2,602 – 644*45/10 = -296
r = Sxy/√$(Sxx * Syy)$ = -0.686 [3]

-69% correlation co-efficient also implies a moderate negative linear relation between the two variables as visible from the scatterplot. [1]

[4]

iii) Null hypothesis H0: ρ = 0 against H1: ρ < 0 [1]

Need to assume that data come from a bivariate normal distribution. [1]

From page 25 of tables, r = 0.5 * ln(1-0.686/1.686) = -0.8404 [1]

And under H0, this should be a value from the N(0, 1/7) distribution. [1]

Fisher's standardized statistic = (-0.8404 – 0)/($\sqrt{1/7}$) = -2.22 [1]

This gives the p-value = P(z<-2.22) = 0.013 which is quite small and hence shows a strong evidence to reject the null hypothesis with 95% confidence. We can conclude that marks obtained and hours spent on social media are negatively correlated. [2]

[7]

iv) Beta = Sxy/Sxx = -296/75 = -3.9467
Alpha = mean of y – beta * mean of x = 644/10 + 3.9467 *45/10 = 82.16
Fitted line is y = 82.16 – 3.9476x

[3]

v) R^2 = -0.686^2 = 0.4706
This gives the proportion of total variation explained by the model. [2]

vi) For every additional hour spent on social media per day, the total marks reduce by 3.95 (~4 marks) basis the fitted equation. [1]

[18 Marks]

Solution 9:

i) Sum of Xi follows Gamma distribution with parameters 5n, λ [1.5]

If Y ~ gamma (α, λ) then 2 λ Y ~ chi squared distribution with degree of freedom 2α
Hence 2nλ$\bar{X}$ follows chi squared distribution with df 10n. [1.5]

[3]

ii) Option B is correct **[3]**

[6 Marks]

Solution 10:

i) Link function is g(μ) = logμ [1]

ii)

a) The linear predictor is $\alpha_i + \beta x$ where the intercept α_i for i = 1, 2 depends on gender [2]

b) The linear predictor is $\alpha_i + \beta_i x$ where both parameters depend on the gender [2]

[5 Marks]

Solution 11:

H0: There is no difference among industries [1]

H1: At least one industry differs significantly from the overall mean

SS_R = 19 (5^2 + 10^2 + 8^2) = 3591 [1.5]

Mean of resignation = (27+36+30)/3 = 31

SS_B = 20((27-31)^2 + (36-31)^2 + (30-31)^2) [1.5]

= 840

$F_{2, 57}$ = (840/2)/(3591/57) = 6.667 [1]

The 1% point from $F_{2,60}$ is 4.977 and since the test statistic is higher than this, the null hypothesis is rejected. We conclude that resignation rate is different across different industries. [1]

[6 Marks]

****