# Institute of Actuaries of India

# Subject CT6 – Statistical Methods

# November 2013 Examinations

# INDICATIVE SOLUTIONS

**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) Let X denotes total premium received from a single policy.

$P(X = 0) = 0.1$
$P(X = 1000n$ / No refund$) = 1/6$; $n = 1,2.3,\ldots\ldots,6$.
Where n is the number of year for which premiums are paid

$E(X/\text{no refund}) = 1000*3.5 = 3500$
$V(X/\text{no refund}) = 1000^2 * 2.916667 = 2916667$.

$E(X) = E(X/\text{no refund})\ E(\text{no refund}) = 3500 * 0.9 = 3150$.
$V(X) = V(E(X/\text{no refund})) + E(V(X/\text{no refund})) = 3500 + 2916667 = 2920167$. **(3 Marks)**

**ii)** The event space consists $6^{100}$ points. The number of event points contained in the required event is the no. of different set of integers $(x_1,x_2,..,x_{100})$ where $x_1 + x_2 +\ldots.+ x_{100} = j$, where $j = m / 1000$ and each of $x_1,x_2,x_3,\ldots,x_{100}$ can take the values 1,2,3,4,5,6.

Thus the required number of favorable events is the coefficient of $x^j$ in the expression of $(x + x^2 + \ldots.+x^6)^{100}$.

Now $x + x^2 + \ldots+ x^6 = x(1 – x^6) / (1 –x)$.
$(1 – x^6)^{100} = \sum(-1)^i\ {}^{100}C_i\ x^{6i}$ for $i = 0$ to $100$.
$(1 – x)^{-100} = \sum {}^{(100 + k – 1)}C_{(100 – 1)}\ x^k$, for $k = 0$ to $\infty$.

So, $(x + x^2+ \ldots + x^6)^{100} = \sum\sum (-1)^i\ {}^{100}C_i\ {}^{(100 + k -1)}C_{(100 – 1)}\ x^{(100+6i +k)}$, where i ranges from 0 to 100 and k ranges from 0 to $\infty$.

If $100 + 6i + k = j$ , Then $k = j – 100 – 6i$. As $k \geq 0$, $i \leq (j – 100) / 6$.

So, the coefficient of $x^j$ in the above expression is:
$\sum(-1)^i\ {}^{100}C_i\ {}^{(j – 6i – 1)}C_{(100 – 1)}$, where i ranges from 0 to n where n is the greatest integer not exceeding $(j – 100) / 6$.

Thus $P(X = j) = (6^{-100}) * \sum((-1)^i *{}^{100}C_i * {}^{(j – 6i – 1)}C_{99})$. where i ranges from 0 to n as defined above.

**(6 Marks)**

**iii)** Where m = 100000, j = 100, so n = 0 and by the above formula P(X = 100000) = $6^{-100}$.
And m = 100000 implies that each of the 100 policies has paid one premium only and so the required probability is $6^{-100}$. **(1 Mark)**

**[Total Marks -10]**

**Solution 2 :**

**i)** $X_t = (4/3) X_{t-1} - (7/12) X_{t-2} + (1/12) X_{t-3} + \varepsilon_t$,

Taking covariance with $X_{t-1}$, $\gamma_1 = Cov(X_t, X_{t-1}) = (4/3)\gamma_0 - (7/12)\gamma_1 + (1/12)\gamma_2$.
Dividing by $\gamma_0$, $\rho_1 = (4/3) – (7/12)\rho_1 + (1/12)\rho_2$, or $(19/12)\rho_1 = (1/12)\rho_2 + (4/3)$ …… (1)

Taking covariance with $X_{t-2}$, $\gamma_2 = Cov(X_t, X_{t-2}) = (4/3)\gamma_1 - (7/12)\gamma_0 + (1/12)\gamma_1$.
Dividing by $\gamma_0$, $\rho_2 = (4/3)\rho_1 – (7/12) + (1/12)\rho_1$, or $\rho_2 = (13/12)\rho_1 - (7/12)$ …… (2)

Solving equation (1) & (2) $(19/12)\rho_1 = (1/12)((13/12)\rho_1 - (7/12)) + (4/3)$,
Or, $((19/12) – (13/144))\rho_1 = (4/3) – (7/144)$, or $\rho_1 = (105/144) = (21/43)$.

So, $\rho_2 = (13/12)(21/43) – (7/12) = (273 – 301) / (12X43) = - 7/129$.

**(4 Marks)**

**ii)** The model can be written as :
$X_t - (4/3) X_{t-1} + (7/12) X_{t-2} - (1/12) X_{t-3} = \varepsilon_t$,

Using backward shift operator B, we get
$(1 – (4/3)B + (7/12)B^2 – (1/12)B^3) X_t = \varepsilon_t$.

The characteristic equation is :
$1 – (4/3)x + (7/12)x^2 – (1/12)x^3 = 0$
Or, $x^3 – 7x^2 + 16x – 12 = 0$,
Or $(x – 3)(x – 2)^2 = 0$,

So, the roots of the characteristic equation are 3 & 2, which are greater than 1.
Hence the time series is a stationary one.

**(3 Marks)**

**iii)** Partial auto correlation coefficients, $\Phi_1 = \rho_1 = 21/43$.
$\Phi_2 = (\rho_2 – \rho_1^2) / (1 - \rho_1^2) = -0.38447$. **(2 Marks)**

**[Total Marks-9]**

**Solution 3 :**

X = U $^{(1/4)}$), => U = $X^4$.
U = -U/3, or U = -3U =>U = -3 $X^4$ .
U = LN(1 – Z) => Z = 1 - exp(-3 $X^4$ ).

Now Z is the uniform (0,1) distribution, so, X can take any value between 0 to ∞ and Z should represent the corresponding Distribution function.
Thus F(x) = 1 + exp(3 $x^4$ ) is the corresponding distribution function.

Taking derivative to both sides with respect to x,
The density function, f(x) = $12x^3$ exp(3 $x^4$ ) = 3 . 4 . $x^{4-1}$ exp(3 $x^4$ ), where $0 \le x \le \infty$ .

This is clearly the density function of Weibull distribution with parameters 3 & 4.
Thus the student was generating the random variates for Weibull(3,4) distribution.

**(6 Marks)**

**Solution 4 :**

**i)** Let $X_i$ & $Y_i$ be the $i^{th}$ medical consultation & medicine expenses.
Let N be the total no. of claims over 1 year from the scheme and n be the number of employees of the employer.
So, N has a compound Poisson distribution with parameter 0.5n and S = Σ(Xi + Yi), where the sum is taken for i = 1 to N.
So, {$X_i + Y_i$}, for i = 1 to ∞,is a sequence of independent & identically distributed random variables, independent of N.

Thus, S has a compound Poisson distribution where the $i^{th}$ individual claim is $X_i + Y_i$.

E(S) = 0.5n (E(Xi + Yi)) and V(S) = 0.5n (E$(Xi + Yi)^2$) = 0.5n(E($Xi^2$) + 2E(Xi)E(Yi) + E($Yi^2$)),
Since $X_i$ & $Y_i$ are independent.

E(Xi)= α/λ, E(Yi) = (C + 100)/ 2.
E($Xi^2$) = α(α + 1)/$\lambda^2$ . E($Yi^2$) = ($C^2$ + 100C + 10000) / 3.
When, α = 5.5 , λ = 0.01 & C = 400,
E(S) = 0.5n (5.5/ 0.01 + 250) = 400n.
V(S) = 0.5n (550.650 + (160000+40000+10000) / 3 + 2.550.250)
= $592.66^2$ n.

Total yearly premium collected = Rs. 12 * 40n = Rs. 480n.

When S has an approximate normal distribution, then
P(S ≤480n) ≥ 0.99,
Or P(((S – 400n) / 592.66√n) < ((480n – 400n) / 592.66√n)) ≥ 0.99
Or, 80√n /592.66 ≥ 2.326
Or n > 296.93

So, the minimum number of employee the company should have is 297. **(7 Marks)**

**ii)** The worst possible combination for the employer is the set of values of α, λ & C which produces the highest possible values of E(S) & V(S).

Let m and d denotes the mean and std. dev. of total consultation and medicine expenses arising out of a single employee's family.

So, $E(S) = mn.$ & $V(S) = nd^2$ .

From the above the minimum value of n is derived as , $(480 – m) \sqrt{n} / d \geq 2.326$. Or, $n \geq (2.326d/ (480 –m))^2$ .

Highest value of n results from highest possible values of m & d , provided m < 480.
$m = 0.5 E(Xi + Yi) = 0.5((\alpha/\lambda) + (C+100)/2))$

$d^2 = 0.5 (E(Xi^2) +2 E(Xi)E(Yi) + E(Yi^2))$
$= 0.5((\alpha(\alpha + 1)/\lambda^{2)} + (C^2 + 100C + 10000) / 3 + \alpha(C+100)/\lambda)$
So, m and d are maximized when α & C are maximum and λ is minimized.

So, the required combination is α = 6, C = 500 and λ = 0.0095.

This combination gives m = 465.79 & d = 688.35.

Which gives $n \geq (2.326.688.35/ (480 –465.79))^2$
Or, n ≥ 12695.5 Or n ≥ 12696 (Rounded to next higher figure).

(Some printing mistakes were there. So, credit was given for any sensible approach)

**(6 Marks)**
**[Total Marks-13]**

**Solution 5 :**

**i)** The equation for adjustment coefficient is $M_x(r) = 1 + (1 + \theta)m_1r$.

We have X is an exp(0.0001) variable. So, $M_x(r) = 1 / (1 – 10000r)$. θ = 0.25 and $m_1 = E(X) = 10000$.
Thus the equation becomes :
1 / (1 – 10000r) = 1 + 1.25 * 10000r.

Or, 1 – 10000r + 12500 r – 12500 * 10000 = 1

Or, r = 2500 / (12500 * 10000) = 1 / 50000. (Neglecting r = 0 as r > 0)

This is the adjustment coefficient for the insurer.

From Lundburg's inequality, the upper bound of the probability of ruin is given by:
ψ(U) ≤ exp(-1000000 / 50000) or ψ(U) ≤ exp(-20).

**(6 Marks)**

**ii)** Clearly, the above expression is independent of the Poisson parameter $\lambda$. The higher value of Poisson parameter speeds up the whole process of claim and so, the claim arises quickly.
So, in this case the ruin will happen early rather than late. However, it does not affect the probability of ruin.

The value of the Poisson parameter determines the time when the ruin will occur if ruin occurs.

So, the probability of ruin does not depend on $\lambda$ and $\lambda$ determines the timing (or speed) of ruin if there is a ruin at all.

**(4 Marks)**
**[Total Marks-10]**

**Solution 6 :**

**i)** The Poisson distribution is given by $f(y) = \exp(-\mu)\mu^y / y!$ for y = 0, 1, 2 ……

The function can be written as $f(y) = \exp((y\log\mu - \mu)/1 - \log y!)$.

Standard exponential family of distribution is denoted by it's form:
$g(y) = \exp((y\theta - b(\theta))/ a(\varphi) + c(y,\varphi))$.

So, if we substitute the values by $\theta = \log\mu$, $b(\theta) = \mu = \exp(\theta)$, $a(\varphi) = \varphi = 1$ and $c(y,\varphi) = -\log y!$, it can be a member of the exponential family of distribution.

**(2 Marks)**

**ii)** The log-likelihood function can be written as

$\log L(\mu_{I}, \mu_{II}, \mu_{III}) = \sum(y_i \log\mu_i) - \sum\mu_i - \sum\log y_i!$.

So, the log likelihood function for model I becomes :

$\log L = a\sum y_i + b\sum y_i + c\sum y_i - 10\exp(a) - 5\exp(b) - 20\exp(c) - \sum\log y_i!$.
(the 1st sum is taken for i = 1 to 10, 2nd sum for i = 11 to 15 , 3rd sum for i = 16 to 35 and the 4th sum is taken for i = 1 to 35)

$= 11a + 3b + 4c - 10\exp(a) - 5\exp(b) - 20\exp(c) - \sum\log y_i!$

By taking partial derivative of logL with respect to a, b & c and equating to 0, we get the maximum likelihood estimators of a, b & c as:
a= log 1.1 = 0.9531, b = log(.6) = -.51083 , c = log(.2) = -1.60944. **(4 Marks)**

**iii)** For model II, the log likelihood function becomes:
$\log L = a\sum y_i - 35\exp(a) - \sum\log y_i! = 18a - 35\exp(a) - \sum\log y_i!$.

Differentiating with respect to a and equating to 0, it becomes:
$18 - 35\exp(a) = 0$, or, $a = \log(18/35) = -.66498$. **(2 Marks)**

**iv)** The scaled deviance for model I is $2(\log L_s - \log L_I)$, where $L_s$ is the value of the log-likelihood function of the saturated model and $L_I$ is the value of the log likelihood function for model I.
For the saturated model we can replace $\mu_i$ with $y_i$ in the equation (1) above as it fits the observed data perfectly.
So, the expected results are the observed result. Thus, $\log L_s = \sum(y_i \log \mu_i) - \sum \mu_i - \sum \log y_i!$.

$= 4.2\log 2 - 18 - 4\log 2$ (ylogy = 0 for y = 0 & 1, = 2log2 for y = 2)

$= 2\log 2 - 18 = -15.2274$.

$\log L_I = 11a + 3b + 4c - 10\exp(a) - 5\exp(b) - 20\exp(c) - \sum \log y_i!$
$= -27.6944$.

Thus the scaled deviance for model I = 2(-15.2274 – ( - 27.6944)) = 24.93.

Similarly, $\log L_{II} = 18a - 35\exp(a) - \sum \log y_i!$
$= 18\log(18/35) - 18.\ 4\log 2 = -32.7422$.

Thus the scaled deviance for model II = 2(-15.2274 – ( - 32.7422)) = 35.03.

**(6 Marks)**

**v)** We can use the chi-square distribution to compare model I and model II. The difference in the scaled deviance = $2(\log L_{II} - \log L_I)$ = 35.03 – 24.93 = 10.10.

The test statistic $2(\log L_{II} - \log L_I)$ should have a chi-square distribution with 3 -1 = 2 degrees of freedom which has a critical value of 5.991 at the upper 5% level. Since 10.10 > 5.991, our value is significant here.

So, model I is a significant improvement over model II. We prefer model I here.

**(3 Marks)**
**[Total Marks-17]**

**Solution 7 :**

**i)** Let b = a+ 1. So, a > -1 gives b > 0. Thus the prior distribution becomes
$f(p) \propto \{p(1-p)\}^{b-1}$ where b > 0.

So, the total monthly number of claims becomes a binomial distribution with parameters m & p.

The likelihood function bases on the number of monthly claims in the last n months becomes $L(p) = {}^mC_{x1}\, p^{x1} (1-p)^{m-x1} * {}^mC_{x2}\, p^{x2} (1-p)^{m-x2} * \ldots\ldots * {}^mC_{xn}\, p^{xn} (1-p)^{m-xn}$. So, $L(p) \propto p^{\sum xi} * (1-p)^{mn - \sum xi}$.

Since the posterior distribution of p is proportional to the product of likelihood function and prior distribution,

Posterior distribution of p $\propto \{p(1-p)\}^{b-1} * p^{\sum xi} * (1-p)^{mn - \sum xi}$
$= p^{\sum xi + b - 1} * (1-p)^{mn + b - \sum xi - 1}$
Which is a form of another beta distribution with parameters $\sum x_i + b$ and $mn + b - \sum x_i$.

**(4 Marks)**

**ii)** The likelihood function based on the observed data

$L(p) = C * p^{\sum xi} * (1-p)^{mn - \sum xi}$, where C is a constant.
$Log(L) = LogC + \sum x_i \log p + (mn - \sum x_i) \log(1-p)$.

Differentiating w.r.t. p and equating to 0, it gives:
$\sum x_i / p - (mn - \sum x_i) / (1-p) = 0$, Or, $\underline{p} = \sum x_i / mn$.

Taking second derivative, the expression becomes:
- $\sum x_i / p^2 - (mn - \sum x_i) / (1-p)^2$.

Since $mn \geq \sum x_i$, The expression < 0.
So, the maximum likelihood estimate of p is $\sum x_i / mn = p$ **(3 Marks)**

**iii)** The Bayesian estimate under quadratic loss is the mean of the posterior distribution.
So, the Bayesian estimate of p is given as :
$(b + \sum x_i) / (b + \sum x_i + b + mn - \sum x_i) = (b + \sum x_i) / (2b + mn)$.

We have to rewrite it as $Z\underline{p} + (1-Z)k$, where k is the mean of the prior distribution, so k = ½.

The above expression can be rewritten as :
$(mn/(2b + mn)) * (\sum x_i / mn) + (2b/(2b + mn)) * ½$
$= Z(\sum x_i / mn) + (1-Z) * ½$, where $Z = mn / (2b + mn)$.

**(3 Marks)**

**iv)** When n increases, Z increases and for very large values of n, for a given b, Z tends to 1. It means for a given b, as the size of past observations increases, more and more weight is assigned to M.L.E of p and lesser weight is assigned to prior estimates of p.

**(1 Mark)**

**v)** When a = 0, b = 1 and So, p* = 16 / 1202 = 8/601. Z = 1200 / 1202.
When a = 3, b = 4 and So, p* = 19 / 1208 = 1 / 80. Z = 1200 / 1208. **(2 Marks)**

**vi)** When a = 0, b = 1 & so, prior variance = 1 / 2.2.3 = 1/ 12.
When a = 3, b = 4 & so, prior variance = 4.4/8.8.9 = 1/36.
So, as a increases, prior variance of p decreases. Though the prior mean remains same as ½, but with higher value of a, we are more confident about p around ½. **(2 Marks)**

**[Total Marks-15]**

**Solution 8 :**
**i)**
**(a)**
$F(x) = 1 - exp\,(-cx^{1/4})$
Differentiating with respect to x

$$f(x) = \frac{1}{4} c\, x^{-3/4} e^{-cx^{1/4}}$$

Thus, the m$^{th}$ non – central moment is:
$$E(X^m) = \int_0^\infty \frac{1}{4} c\, x^{-\frac{3}{4}} e^{-cx^{\frac{1}{4}}} dx$$

Substituting $x^{\frac{1}{4}} = y \Rightarrow x = y^4 \Rightarrow dx = 4\, y^3 dy$

$$E(X^m) = \int_0^\infty y^{4m} \frac{1}{4}\, cy^{-3} e^{-cy}\, 4\, y^3 dy$$

$= c \int_0^\infty y^{4m} e^{-cy} dy$ **(4 Marks)**

**(b)**
Comparing the above integrand with gamma function gives $\alpha = 4m + 1, \lambda = c.$

Total probability of distribution = 1

$$\int_0^\infty \frac{c^{4m+1}}{\Gamma(4m+1)}\, y^{4m} e^{-cy} dy = 1$$

$$\Rightarrow \int_0^\infty y^{4m} e^{-cy}\, dy = \frac{\Gamma(4m+1)}{c^{4m+1}}$$

Using expression for $E(X^m)$ from (a):

$E(X^m) = c * \Gamma(4m + 1) / c^{4m+1} = c^{-4m} (4m)!$ , where ! denotes factorial function **(3 Marks)**

**ii)**

The likelihood function with 100 observations:
$L = \prod_{i=1}^{100} \frac{1}{4}\, c\, x_i^{-3/4} e^{-c x_i^{1/4}} = c^n\, e^{-c \sum x_i^{1/4}} * constant$

$$\Rightarrow \log L = n \log c - c \sum x_i^{1/4} + k$$

Differentiating with respect to c

$\frac{\partial \log L}{\partial c} = \frac{n}{c} - \sum x_i^{1/4}$

Equating to 0;

$\frac{n}{c} = \sum x_i^{1/4} => c = \frac{n}{\sum x_i^{1/4}}$

$=> c = \frac{100}{1430} = 0.07$

Thus the fitted distribution is:

$F(x) = 1 - e^{-0.07\, x^{1/4}}; x > 0$

**(4 Marks)**

**iii)**

**(a)**

The values of the distribution function at the critical values are:

$F(0) = 0$

$F(100) = 1 - e^{-0.07 * 100^{1/4}} = 0.1985$

$F(1{,}000) = 1 - e^{-0.07 * 1{,}000^{1/4}} = 0.3254$

$F(10{,}000) = 1 - e^{-0.07 * 10{,}000^{1/4}} = 0.5034$

$F(100{,}000) = 1 - e^{-0.07 * 100{,}000^{1/4}} = 0.7120$

$F(\infty) = 1$

The expected numbers can then be calculated by multiplying the probabilities for each range by n.

| Band | Actual Number | Probability | Expected Number |
|---|---|---|---|
| 0 < x < 100 | 12.00 | F(100) - F(0) = 0.1985 | 19.85 |
| 100 ≤ x < 1,000 | 15.00 | F(1000) - F(100) = 0.1269 | 12.69 |
| 1,000 ≤ x < 10,000 | 18.00 | F(10,000) - F(1,000) = 0.1780 | 17.80 |
| 10,000 ≤ x < 100,000 | 18.00 | F(100,000) - F(10,000) = 0.2086 | 20.86 |
| x ≥ 100,000 | 37.00 | F(∞) - F(100,000) = 0.2880 | 28.80 |
| Total | 100.00 | | 100.00 |

**(4 Marks)**

**(b)** The $\chi^2$ goodness of fit statistic is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(12-19.85)^2}{19.85} + \frac{(15-12.69)^2}{12.69} + \frac{(18-17.80)^2}{17.80} + \frac{(18-20.86)^2}{20.86} + \frac{(37-28.80)^2}{28.80}$$

$= 3.104 + 0.42 + 0.002 + 0.392 + 2.335 = 6.253$

**(2 Marks)**

**(c)**

We are using 5 groups. The expected numbers have been calculated based on the total for the actual numbers. We have estimated one parameter. So the total numbers of degrees of freedom is 5-1-1 = 3.

The observed values of 6.25 at 3 degrees of freedom is less than 7.815, the upper 5% point of $\chi_3^2$ distribution.

So we cannot reject the Weibull distribution at 5% level.\The components of the chi – square statistic are largest at the extremes of the distribution, i.e. the lower and upper tail of the distribution.

The weibull model appears to overestimate the numbers in the lower tail and underestimate the numbers in the upper tail.

**(3 Marks)**

**[Total Marks-20]**

******************************