Institute of Actuaries of India

Subject CT3 – Probability and Mathematical Statistics

November 2010 Examinations

INDICATIVE SOLUTIONS

Sol. 1) **(a)** sample mean = 3125
sample standard deviation = 1316.64

(b) for k = 2
upper bound = 3125+2*1316.64 = 5758.29
lower bound = 3125-2*1316.25 = 491.71

(c) total no. of data under the above limits = 18 = 90%
according to mathematician, minimum proportion of data in the interval = $1-1/2^3$ = 87.5%
Thus, mathematician's theorem is valid for k = 2 for the given data **[6]**

Sol. 2) X is U[0,k]
r in the interval 0 to 1
X_1 lies in the interval 0 to k
X_2 lies in the interval k to 0 i.e. 0 to k
X_3 lies in the interval 0 to 1/k which do not cover the full interval
Hence X_1 and X_2 are valid simulated observations of X **[2]**

Sol. 3) **(a)**Suppose that person1 is born on a particular day i, where i can be any one of the 365 days. So probability = 1/365
If person2 does not share the same birthday, then he should have born on any one of the rest 364 days. So probability = 364/365
Finally, there can be such 365 combinations as i can be any of the 365 days
Probability of 2 persons not having the same birthday is
$1/365 * 364/365 * 365 = 365*364/365^2$
Thus, probability of 2 persons having the same birthday is $1-365*364/365^2 = 0.002740$

(b) Let all 3 persons do not share the same birthday whose probability based on above argument is $1/365 *364/365 * 363/365*365 = 365*364*363/365^3$
Thus, probability of at least 2 persons having the same birthday is
$1-365*364*363/365^3 = 0.008204$
For a group of 4 persons, the probability at least 2 persons having same birthday would be
$1- 365*364*363*362/365^4 = 0.016356$

(c) Let the group size be 15
Thus, for a group of 15 persons, the probability that at least 2 persons have the same birthday would be
$1- 365*364*\ldots(365-15+1)/365^{15}$
$= 1- 0.7471 = 0.2529$

[9]

Sol.4) bias = $E(\hat{\theta})-\theta=\frac{k}{k+1}E(X)-\theta=\frac{k}{k+1}\theta-\theta=\frac{-\theta}{k+1}$

$$Var(\hat{\theta})=Var(\frac{kX}{k+1})=\frac{k^2\theta^2}{25(k+1)^2}$$

MSE = variance + bias^2

Thus,

$$MSE_{\hat{\theta}}(\theta)=\frac{k^2\theta^2}{25(k+1)^2}+\left(\frac{-\theta}{k+1}\right)^2$$

But, it is given $MSE_{\hat{\theta}}(\theta)=2[bias_{\hat{\theta}}(\theta)]^2=2\left(\frac{-\theta}{k+1}\right)^2$

Thus, equating the above equations

$$\frac{\theta^2}{(k+1)^2}=\frac{k^2\theta^2}{25(k+1)^2}\Rightarrow k^2=25\Rightarrow k=5 \text{ as given } k>0$$

[4]

Sol.5) Expected claims for the insurance company = 1*0.2 + 2*0.1 = 0.4

Insurance Premium = 110% * 0.4 = 0.44

Revenue = 1.70

Other Expenses = 20% * 1.70 = 0.34

Profit before repair cost = 1.70 – 0.34 – 0.44 = 0.92

Retained repair cost = 0 with probability 0.4

= 1 with probability 0.6

Profits after retained repair cost = 0.92 with probability 0.4

= -0.08 with probability 0.6

Thus, expected dividends = 0.92*0.4 + 0*0.6 = 0.368 **[4]**

Sol.6) $F_Y(y)=P(Y\le y)=1-P(Y>y)$

= 1 – $P(X_1 > y)\ldots P(X_n > y)$ as X_i's are mutually stochastic independent variables

= 1 – $P(X_i > y)^n$

$$P(X_i\le y)=\int_0^y 3(1-u^2)du=\left|3(u-\frac{u^3}{3})\right|_0^y=3y-y^3$$

Hence, $P(X_i > y) = 1 - 3y + y^3$; Therefore, $F_Y(y) = 1 - (1 - 3y + y^3)^n$

$$f_Y(y)=\frac{d}{dy}F_Y(y)=-n(1-3y+y^3)^{n-1} \text{ X } (-3+3y^2)=3n(1-y^2)(1-3y+y^3)^{n-1} \quad 0 < y < 1$$

[4]

Sol. 7) Let S = score

$$E(S) = E[E(S/\theta)] = E(\theta) = 75$$

$$Var(S) = E[Var(S/\theta)] + Var[E(S/\theta)]$$

$$= E(8^2) + Var(\theta) = 64 + 36 = 100 = 10^2$$

$$P[S < 90/S > 65] = P(65 < S < 90)/P(S > 65)$$

$$= \frac{F(90) - F(65)}{1 - F(65)} = \frac{\Phi\left(\frac{90-75}{10}\right) - \Phi\left(\frac{65-75}{10}\right)}{1 - \Phi\left(\frac{65-75}{10}\right)}$$

$$= \frac{\Phi(1.5) - \Phi(-1)}{1 - \Phi(-1)} = \frac{0.9332 - (1 - 0.8413)}{1 - (1 - 0.8413)}$$

$$= 0.9206$$

[5]

Sol. 8)

(a) $M_X(t) = E(e^{tX}) = \int_0^{0.5} 4xe^{tx}dx + \int_{0.5}^{1} 4(1-x)e^{tx}dx$

$$= 4\int_0^{0.5} xe^{tx}dx + 4\int_{0.5}^{1} e^{tx}dx - 4\int_{0.5}^{1} xe^{tx}dx$$

Now,

$$\int_{0.5}^{1} xe^{tx}dx = \left|\frac{xe^{tx}}{t}\right|_{0.5}^{1} - \frac{1}{t^2}\left|e^{tx}\right|_{0.5}^{1} = \frac{e^t}{t} - \frac{0.5e^{0.5t}}{t} - \frac{e^t}{t^2} + \frac{e^{0.5t}}{t^2}$$

Similarly

$$\int_0^{0.5} xe^{tx}dx = \left|\frac{xe^{tx}}{t}\right|_0^{0.5} - \frac{1}{t^2}\left|e^{tx}\right|_0^{0.5} = \frac{0.5e^{0.5t}}{t} - \frac{e^{0.5t}}{t^2} + \frac{1}{t^2}$$

and

$$\int_{0.5}^{1} e^{tx}dx = \left|\frac{e^{tx}}{t}\right|_{0.5}^{1} = \frac{e^t}{t} - \frac{e^{0.5t}}{t}$$

Thus,

$$M_X(t) = \left(\frac{2e^{0.5t}}{t} - \frac{4e^{0.5t}}{t^2} + \frac{4}{t^2}\right) + \left(\frac{4e^t}{t} - \frac{4e^{0.5t}}{t}\right) - \left(\frac{4e^t}{t} - \frac{2e^{0.5t}}{t} - \frac{4e^t}{t^2} + \frac{4e^{0.5t}}{t^2}\right)$$

$$M_X(t) = \frac{4}{t^2}\left(e^{\frac{t}{2}} - 1\right)^2$$

(b) (i) $M_Y(t) = E(e^{t.\frac{1}{2}(X_1+X_2)}) = E(e^{.\frac{t}{2}X_1}).E(e^{.\frac{t}{2}X_2})$

$= [M_X\left(\frac{t}{2}\right)]^2 = \frac{4}{t^2}(e^{\frac{t}{2}} - 1)^2$

(b) (ii) $\overline{X} = \frac{1}{2}(X_1 + X_2)$ is the mean of two random samples from U[0,1]

Based on the results of (a) and (b) (i) above, we can say that $\overline{X}$ has the triangular distribution as specified in (a)

[8]

Sol.9)

(a)

$$E(XY) = \int_0^1 \int_0^1 xy\, f_{X,Y}(x,y)dxdy$$

$$= \int_0^1 y \int_0^1 x\, f_{X,Y}(x,y)dxdy = \int_0^1 y \int_0^1 x\,[1 - \alpha(1-2x)(1-2y)]dxdy$$

$$= \int_0^1 y \int_0^1 [x(1 - \alpha + 2\alpha y) + 2x^2\alpha(1-2y)]dxdy$$

$$= \int_0^1 y[\frac{x^2(1-\alpha+2\alpha y)}{2} + \frac{2}{3}(\alpha - 2\alpha y)x^3]\,|_0^1 dy$$

$$= \int_0^1 y(\frac{1}{2} + \frac{1}{6}\alpha - \frac{1}{3}\alpha y)dy = \left[\frac{y^2}{4} + \frac{\alpha y^2}{12} - \frac{\alpha y^3}{9}\right]|_0^1 = \frac{1}{4}(1 - \frac{\alpha}{9})$$

(b)

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)dy = \int_0^1 [1 - \alpha(1-2x) + 2\alpha y(1-2x)]dy$$

$= \left|y - a(1-2x)y + a(1-2x)y^2\right|_0^1 = 1$

Similarly, $f_Y(y) = \int_0^1 f_{X,Y}(x,y)dx = 1$

$$E(X) = \int_0^1 x f_X(x)dx = \int_0^1 xdx = \frac{x^2}{2}\,|_0^1 = \frac{1}{2}$$

Similarly,

$$E(Y) = \int_0^1 y f_Y(y) dy = \frac{1}{2}$$

(c) The covariance between X and Y is:

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{4}\left(1 - \frac{\alpha}{9}\right) - \frac{1}{2} * \frac{1}{2} = -\frac{1}{36}\alpha$$

To prove: Independence implies null covariance

If both random variables are independent, this means that: $f_{X,Y}(x,y) = f_X(x) * f_Y(y) = 1$

The only way for the joint density to be equal to 1 is to have α = 0. But if α = 0, then Cov(X,Y) = 0. Therefore, the independence implies that the covariance is equal to 0.

To prove: Null covariance implies independence

If the covariance is equal to 0, we must have α = 0. But then if α = 0, the joint density is equal to 1, which means it is equal to the product of the marginal densities. Thus, the two random variables are independent.

Hence, the statement is true. **[10]**

Sol.10) **(a)** Since S_n has exponential distribution with mean 5, hence, the parameter λ is 1/5 = 0.2

$\Pr[S_n > \Pi] = e^{-0.2\Pi} = 0.01$

$\Pi = 23.03$

(b) S_n follows lognormal (μ, σ^2)

Therefore $e^{\mu+0.5\sigma^2} = 5$ and $e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right) = 25$

dividing, we have $\left(e^{\sigma^2} - 1\right) = 1$

$\sigma^2 = 0.69315$ or $\sigma = 0.83256$

Hence, $\mu = 1.26287$

$\Pr[S_n > \prod] = \Pr[\log S_n > \log \prod]$

$$= P\left[\frac{\log S_n - 1.26286}{0.83255} > \frac{\log \prod - 1.26286}{0.83255}\right] = 0.01$$

$\downarrow$ $\downarrow$

Z ~ N(0,1) 2.326

Giving $\prod = 24.52$

(c) $\Pr[(S_1 \le 23.03) \cap (S_1 + S_2 \le 46.06)] =$

$\int_0^{23.03} \frac{1}{5} e^{-x/5} P[S_1 + S_2 \le 46.06 | S_1 = x] dx$ conditioning on the value of S_1

$= \int_0^{23.03} \frac{1}{5} e^{-x/5} P[S_2 \le 46.06 - x] dx$

$= \int_0^{23.03} \frac{1}{5} e^{-\frac{x}{5}} (1 - e^{-(46.06-x)/5}) dx$

$= 1 - e^{-23.03/5} - (23.03/5)e^{-46.06/5}$

$= 0.9895$ **[10]**

Sol.11) n = 30

x	F(x)	f(x)	Expected Claims in each interval
310	0.16	0.16	4.80
500	0.27	0.11	3.30
2,498	0.44	0.17	5.10
4,876	0.65	0.21	6.30
7,498	0.83	0.18	5.40
12,930	0.95	0.12	3.60
infinity	1.00	0.05	1.50

Hence, the first two and the last two intervals need to be combined so that the expected number of claims in each interval is at least 5.

x	Expected (E)	Observed (O)	O – E	$(O-E)^2 / E$
500	8.10	3	-5.10	3.21
2,498	5.10	8	2.90	1.65
4,876	6.30	9	2.70	1.16
7,498	5.40	2	-3.40	2.14
infinity	5.10	8	2.90	1.65
Total	**30**	**30**	**0**	$\mathbf{X^2 = 9.81}$

The $\mathbf{X^2}$ distribution has 4 degrees of freedom because there are five categories (5 – 1 = 4).

The critical value for a chi-square test with 4 degrees of freedom and 5% level of significance is 9.49. Since, chi-square calculated is greater than tabulated, the hypothesis is rejected. **[5]**

Sol.12) **(a)**

$f(x) = \frac{1}{\theta}\, x^{\frac{1-\theta}{\theta}}$

Let L $(\theta) = \prod f(x_i) = \frac{1}{\theta^n} \prod x_i^{\frac{1-\theta}{\theta}}$

$$=> \log_e L(\theta) = -n \log_e \theta + \frac{1-\theta}{\theta} \sum_{i=1}^{n} \log_e x_i$$

$$\frac{d \log_e L(\theta)}{d\theta} = -\frac{n}{\theta} + \sum_{i=1}^{n} \log_e x_i \left(-\frac{1}{\theta^2}\right) = 0 => \hat{\theta} = -\frac{1}{n} \sum_{i=1}^{n} \log_e x_i$$

$$\frac{d^2 \log_e L(\theta)}{d\theta^2} = \frac{n}{\theta^2} + \sum_{i=1}^{n} \log_e x_i \left(\frac{2}{\theta^3}\right) = \frac{n}{\hat{\theta}^2} + (-n\hat{\theta}) \times \left(\frac{2}{\hat{\theta}^3}\right) = -\frac{n}{\hat{\theta}^2} < 0$$

Hence, $\hat{\theta} = -\frac{1}{n} \sum_{i=1}^{n} \log_e x_i$ is the maximum likelihood estimator of θ

(b) $E(\log_e x) = \int_0^1 \log_e x\ \frac{1}{\theta}\, x^{\left(\frac{1-\theta}{\theta}\right)} dx = \frac{1}{\theta} \int_0^1 \frac{\log_e x}{x}\, x^{\frac{1}{\theta}}\, dx$

Put $\log_e x = y, => x = e^y => dx = e^y\, dy$ and,

X = 1 => y = 0 and x = 0 => y = - ∞

$$\frac{1}{\theta} \int_{-\infty}^{0} y\, e^{y\left(\frac{1-\theta}{\theta}\right)}\, e^y dy = \int_{-\infty}^{0} \frac{y}{\theta}\, e^{\frac{y}{\theta}}\, dy =$$

$= \left[\frac{y}{\theta} \int e^{\frac{y}{\theta}} dy\right]_{-\infty}^{0} - \int_0^1 [\frac{d(\frac{y}{\theta})}{dy} \int e^{\frac{y}{\theta}} dy] dy$

$= \left[\frac{y}{\theta}\ \theta\, e^{\frac{y}{\theta}}\right]_{-\infty}^{0} - \int_{-\infty}^{0} \frac{1}{\theta} \theta\, e^{\frac{y}{\theta}} dy$

$= 0 - \int_{-\infty}^{0} e^{\frac{y}{\theta}} dy = \left[-\theta\, e^{\frac{y}{\theta}}\right]_{-\infty}^{0} = -\theta + 0 = -\theta$

$$E(\hat{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\log_e x_i) = -\frac{1}{n} \times -n\theta = \theta$$

Hence, $\hat{\theta}$ is an unbised estimator of θ. **[8]**

Sol.13) Restating the table to be the City Automobile Club's cost after 10% payment by the auto owner:

Towing Cost (x)	p(x)
72	50%
90	40%
144	10%

Then E(X) = 0.5*72 + 0.4*90 + 0.1*144 = 86.4

$E(X^2) = 0.5*72^2 + 0.4*90^2 + 0.1*144^2 = 7905.6$

$Var(X) = 7905.6 – 86.4^2 = 440.64$

Let N be the number of towings each year.

As N is Poisson distributed, E(N) = Var(N) = 1000

Let S be the aggregate towing cost, then

E(S) = E(X)E(N) = 86.4*1000 = 86,400

$Var(S) = E(N)Var(X) + E(X)^2Var(N) = 1000*440.64 + 86.4^2*1000 = 7,905,600$

Pr(S > 90000) =

$$\Pr\left(\frac{S - E(S)}{\sqrt{Var(S)}} > \frac{90000 - 86400}{\sqrt{7905600}}\right) = \Pr(Z > 1.28) = 1 - \boldsymbol{\phi}(1.28) = 0.1$$ **[5]**

Sol.14) **(a)** Hypotheses are

$H_0 : \mu_{STD} = \mu_{LAC} = \mu_{VEG}$ – the mean protein intake is same on all three diets

H_1 : at least two μ's are different

We have, G = 3 groups and n = 26 data points. So the dof are G-1=2, n-G=23 and n-1=25.

$$SSB = \sum_{j=1}^{G} n_j \bar{Y}_j^{\,2} - \frac{\sum Y_{ij}{}^2}{n}$$

$SSB = 10(75)^2 + 10(57)^2 + 6(47)^2 – (10*75 + 10*57 + 6*47)^2/26 = 3286.15$

$$SSW = \sum_{j=1}^{G} (n_j - 1)\, s_j^{\,2}$$

$SSW = 9(9)^2 + 9(13)^2 + 5(17)^2 = 3695$

MSB = SSB / (G – 1) = 1643.08

MSW = SSW / (n – G) = 160.65

F = MSB / MSW = 10.23

From tables, $F_{2,\,23,\,95\%} = 3.42$

Since F observed is greater than tabulated, H_0 is rejected and the diets do not all have the same Protein intake (reject even at 1%)

(b) 95% confidence interval is given by:

$$\bar{y} \pm t_{2.5\%,23}\,\hat{\sigma}\sqrt{\frac{1}{n}}$$

$t_{2.5\%,\,23} = 2.069$

STD:
$75 \pm 2.069\sqrt{160.65/10} = [66.7, 83.3]$

LAC:
$57 \pm 2.069\sqrt{160.65/10} = [48.7, 65.3]$

VEG:
$47 \pm 2.069\sqrt{160.65/6} = [36.3, 57.7]$

The mean protein intake on the **standard diet** (STD) appears to be higher than that on other two diets as the confidence intervals do not overlap. However, the CIs for lacto-vegetarian and strict vegetarian diet do overlap so we cannot be sure their protein intake differs.

(c) 95% confidence interval is given by:

$$(\overline{y_1} - \overline{y_2}) + t_{2.5\%,23}\,\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

STD vs LAC:

$$(75 - 57) \pm 2.069 \times 12.675\sqrt{\frac{1}{10} + \frac{1}{10}} = [6.27, 29.73]$$

STD vs VEG:

$$(75 - 47) \pm 2.069 \times 12.675\sqrt{\frac{1}{10} + \frac{1}{6}} = [14.44, 41.55]$$

VEG vs LAC:

$$(57 - 47) \pm 2.069 \times 12.675\sqrt{\frac{1}{10} + \frac{1}{6}} = [-3.55, 23.55]$$

The two intervals involving the Standard diet do not contain 0, meaning we can be sure that the protein intake is higher on the standard diet than on either of the two vegetarian diets. However, the CI for the two vegetarian diets contains 0, so we do not have sufficient evidence to say that the protein intake on these two diets is different.

To do hypothesis tests, we consider two diets at one time for any difference in the mean protein intake.

STD vs LAC:

H_0 : $\mu_{STD} = \mu_{LAC}$ – the protein intake on the standard and lacto diets is same

H_1 : the two μ's are different – the protein intake on the standard and lacto diets is not the same

$$t_{obs} = \frac{(\bar{y}_{STD} - \bar{y}_{LAC}) - 0}{\sqrt{MSW(\frac{1}{n_{STD}} + \frac{1}{n_{LAC}})}} = \frac{18}{5.67} = 3.17$$

From tables, $t_{calc} = t_{2.5\%,23} = 2.069$

Therefore, null hypothesis is rejected.

STD vs VEG:

H_0 : $\mu_{STD} = \mu_{VEG}$ - the protein intake on the standard and vegetarian diets is same

H_1 : the two μ's are different – the protein intake on the standard and vegetarian diets is not the same

$$t_{obs} = \frac{(\bar{y}_{STD} - \bar{y}_{VEG}) - 0}{\sqrt{MSW(\frac{1}{n_{STD}} + \frac{1}{n_{VEG}})}} = \frac{28}{6.55} = 4.27$$

From tables, $t_{calc} = t_{2.5\%,23} = 2.069$

Therefore, null hypothesis is rejected.

LAC vs VEG:

H_0 : $\mu_{LAC} = \mu_{VEG}$ - the protein intake on the lacto and vegetarian diets is same

H_1 : the two μ's are different – the protein intake on the lacto and vegetarian diets is not the same

$$t_{obs} = \frac{(\bar{y}_{LAC} - \bar{y}_{VEG}) - 0}{\sqrt{MSW(\frac{1}{n_{LAC}} + \frac{1}{n_{VEG}})}} = \frac{10}{6.55} = 1.53$$

From tables, $t_{calc} = t_{2.5\%,23} = 2.069$

[15]

Therefore, we do not sufficient evidence to reject H_0.

Sol.15)

	1	2	3	4	5	6	7	8	9	10	Total
x	40	10	100	110	120	150	20	90	80	130	**850**
y	56	62	195	240	170	270	48	196	214	286	**1,737**
xy	2,240	620	19,500	26,400	20,400	40,500	960	17,640	17,120	37,180	**182,560**
x^2	1,600	100	10,000	12,100	14,400	22,500	400	8,100	6,400	16,900	**92,500**

$S_{xy} = \sum x_i y_i - \sum x_i \sum y_i / n = 182560 - 850 * 1737/10 = 34915$

$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 92500 - 850^2/10 = 20250$

$\hat{b} = S_{xy} / S_{xx} = 1.72$

$\hat{a} = \bar{y} - b\bar{x} = (1737/10) - 1.72 * (850/10) = 27.14$

y = 27.14 + 1.72x

(b) Gradient represents the amount of hours per rupee spent

[5]

****