Institute of Actuaries of India

Subject CT3 – Probability & Mathematical Statistics

November 2013 Examinations
Indicative Solutions

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 approaches leading to a valid answer and examiners have given credit for any alternative approach or interpretation which they consider to be reasonable

Solution 1 :

The percentages of customers who bought newspaper ***A*** from a magazine stall in city ***K*** for Monday to Friday in a randomly selected week are:

62% 55% 63% 58% 62%

i. Mean = (62% + 55% + 63% + 58% + 62%)/5 = 60%

Median = (5+1)/2 i.e. the 3rd observation when the data is sorted from smallest to largest = 62%

(2 Marks)

ii. **a%** and **b%** are the respective percentages of customers who bought newspaper ***A*** from the stall for Saturday and Sunday in that week.

a) Using the whole week's data, the median will be the (7+1)/2 i.e. the 4th observation when the data is sorted from smallest to largest.

The value of the median will be least when both a% and b% are no more than the smallest value of 5-days' data i.e. less than 55%. In which case, the value of median will be 58% as it will be the 4th observation when the data is sorted from smallest to largest.

(1 Mark)

b) We are told that the mean and the median after adding the Saturday and Sunday's data will remain the same as without these two days data.

Without loss of generality, assume a ≤ b.

Setting the means equal:

$$60 = \frac{a + b + 5 * 60}{7} \; i.e. \; a + b = 120$$

Setting the medians equal … In order that the median is 62% we have 2 possibilities:

- Possibility 1: a ≤ 62 and b ≥ 62
- Possibility 2: a, b ≥ 62

Given that we must have a + b = 120, we can only have possibility 1.

So a pair of possible values of (a, b) can be (55, 65).

(3 Marks)

iii. The data is relevant for only 1 week. So, it cannot be really inferred that the mean and median will exceed 50% in every other week. Hence the stall keeper's claims cannot be substantiated for want of sufficient credible data.

(1 Mark)

[Total Marks-7]

Solution 2 :

We can assume that the events concerning the first urn are independent of events concerning the second urn.

Let R_i denote drawing a red ball from urn *i* and let B_i denote drawing a blue ball from urn *i*.
Let *n* be the number of blue balls in the second urn. Then:

$$0.44 = \mathbb{P}\big((R_1 \cap R_2) \cup (B_1 \cap B_2)\big)$$
$$= \mathbb{P}(R_1 \cap R_2) + \mathbb{P}(B_1 \cap B_2) \ \ldots mutually\ exclusive$$
$$= \mathbb{P}(R_1) * \mathbb{P}(R_2) + \mathbb{P}(B_1) * \mathbb{P}(B_2) \ \ldots independence$$
$$= \frac{4}{10} * \frac{16}{16+n} + \frac{6}{10} * \frac{n}{16+n}$$
$$= \frac{64+6n}{160+10n} \quad \Rightarrow n = 4$$

[Total Marks- 3]

Solution 3 :

The lifetime, **T**, of an electronic device is a random variable having a probability density function:

$$f_T(t) = 0.5\, e^{-0.5\,t}\,; \quad t > 0$$

The device is given an efficiency value V = 5 if it fails before time t = 3. Otherwise, it is given a value V = 2 T. Thus V is defined over the range v ≥ 5.

Note: If t < 3, v = 5; If t ≥ 3, v = 2t ≥ 6.

i. Let f_V denote the probability density function of V.

For $v < 5$,

$f_V(v) = 0$ as the smallest value of V is 5.

For $v = 5$,

$$f_V(v) = \mathbb{P}(V = 5)$$
$$= \mathbb{P}(T < 3)$$
$$= \int_0^3 0.5\, e^{-0.5\,t}\, dt = 1 - e^{-1.5}$$

For $5 < v < 6$,

$f_V(v) = 0$ as V does not take any value between 5 and 6.

For $v \geq 6$,

$$\begin{aligned}\mathbb{P}(V \le v) &= \mathbb{P}(V < 5) + \mathbb{P}(V = 5) + \mathbb{P}(5 < V < 6) + \mathbb{P}(6 \le V \le v)\\ &= 0 + (1 - e^{-1.5}) + 0 + \mathbb{P}(6 \le 2T \le v)\\ &= (1 - e^{-1.5}) + \mathbb{P}(3 \le T \le 0.5\,v)\\ &= (1 - e^{-1.5}) + \int_3^{0.5\,v} 0.5\,e^{-0.5\,t}\,dt\\ &= (1 - e^{-1.5}) + (e^{-1.5} - e^{-0.25\,v})\\ &= 1 - e^{-0.25\,v}\end{aligned}$$

Thus,

$$f_V(v) = \frac{d}{dv}[\mathbb{P}(V \le v)] = \frac{d}{dv}[1 - e^{-0.25\,v}] = 0.25\,e^{-0.25\,v}$$

Thus, the PDF (f_v) of V is given as below:

$$f_V(v) = \begin{cases} 0 & v < 5,\ \ 5 < v < 6 \\ 1 - e^{-1.5} & v = 5 \\ 0.25\,e^{-0.25\,v} & v \ge 6 \end{cases}$$

(4 Marks)

ii. The expected efficiency value of the electronic device is given as:

$$\mathbb{E}[V] = 5\,f_V(5) + \int_6^{\infty} v\,f_V(v)\,dv$$

$$= 5\,(1 - e^{-1.5}) + \int_6^{\infty} 0.25\,v\,e^{-0.25\,v}\,dv$$

$$= 5\,(1 - e^{-1.5}) + \int_0^{\infty} 0.25\,(x + 6)\,e^{-0.25(x+6)}\,dx \quad \text{... setting } x = v - 6$$

$$= 5\,(1 - e^{-1.5}) + e^{-1.5}\int_0^{\infty} 0.25\,x\,e^{-0.25\,x}\,dx + 6\,e^{-1.5}\int_0^{\infty} 0.25\,e^{-0.25\,x}\,dx$$

$$= 5\,(1 - e^{-1.5}) + e^{-1.5} * 4 + 6\,e^{-1.5} * 1$$

$$\left[\begin{array}{c}\int_0^{\infty} 0.25\,\mathrm{x}\,\mathrm{e}^{-0.25\,\mathrm{x}}\,\mathrm{dx} = \text{mean of an Exp}(0.25)\text{ random variable}\\ \int_0^{\infty} 0.25\,\mathrm{e}^{-0.25\,\mathrm{x}}\,\mathrm{dx} = \text{total probability of an Exp}(0.25)\text{ random variable}\end{array}\right]$$

$$= 5\,(1 + e^{-1.5}) \ \text{ i.e. } 6.116$$ **(3 Marks)**

[Total Marks – 7]

Solution 4 :

Consider a random sample, $X_1, X_2 \ldots X_n$ from a normal $N(\mu, \sigma^2)$ distribution, with sample mean $\bar{X}$ and sample variance S^2.

i. To show:

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2$$

By definition:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$RHS = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \bar{X} + \bar{X} - X_j)^2$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(X_i - \bar{X})^2 - 2(X_i - \bar{X})(X_j - \bar{X}) + (X_j - \bar{X})^2 \right]$$

$$= \frac{1}{2n(n-1)} \left[\sum_{i=1}^{n} n(X_i - \bar{X})^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \bar{X})(X_j - \bar{X}) + \sum_{j=1}^{n} n(X_j - \bar{X})^2 \right]$$

$$= \frac{1}{2n(n-1)} \left[n(n-1)S^2 - 2 \sum_{i=1}^{n} (X_i - \bar{X}) \sum_{j=1}^{n} (X_j - \bar{X}) + n(n-1)S^2 \right]$$

$$= \frac{1}{2n(n-1)} [n(n-1)S^2 - 0 + n(n-1)S^2]$$

$$\left[\because \sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = 0 \right]$$

$$= S^2.$$

(4 Marks)

ii. Consider any pair of (I, j) such that i ≠ j,

- $E(X_i - X_j) = E(X_i) - E(X_j) = \mu - \mu = 0$
- $Var(X_i - X_j) = Var(X_i) + Var(X_j) - 2\,Cov(X_i, X_j) = \sigma^2 + \sigma^2 - 2 * 0 = 2\,\sigma^2$

[$Cov(X_i, X_j) = 0$ as X_i & X_j are independent]

Thus: $E[(X_i - X_j)^2] = Var(X_i - X_j) + [E(X_i - X_j)]^2 = 2\,\sigma^2$

$$E(S^2) = E\left[\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2\right]$$

$$= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(X_i - X_j)^2\right]$$

$$= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i=1}^{n}E\left[(X_i - X_j)^2\right] \quad \text{as for i = j the terms equate to 0}$$

$$= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i=1}^{n}2\,\sigma^2$$

$$= \frac{1}{2n(n-1)} * 2\,\sigma^2 * n(n-1)$$

$$= \sigma^2$$

Thus, S^2 is an unbiased estimator of σ^2. **(3 Marks)**

[Total Marks- 7]

Solution 5 :

There are two independent random variables **X** and **Y** with probability density functions *g* and *h* respectively, where for any x > 0, we have:

$$g(x) = 3^{18}\,x^{17}\,\frac{e^{-3x}}{17!}\,; \qquad h(x) = 3^6\,x^5\,\frac{e^{-3x}}{5!}$$

We have: S = X + Y

The probability density function of S is given as (for any s > 0):

$$f_S(s) = \int_{-\infty}^{\infty} g(x)\;h(s-x)\,dx$$

[Using the convolution formula given]

$$= \int_0^s 3^{18}\, x^{17}\, \frac{e^{-3x}}{17!} * 3^6\, (s-x)^5\, \frac{e^{-3(s-x)}}{5!}\, dx \quad [\because x > 0\ \&\ s - x > 0]$$

$$= \frac{3^{24}}{17!\ 5!} e^{-3s} \int_0^s x^{17}\ (s-x)^5\, dx$$

$$= \frac{3^{24}}{17!\ 5!} e^{-3s}\, s^{22} \int_0^s \left(\frac{x}{s}\right)^{17} \left(1 - \frac{x}{s}\right)^5\ dx$$

$$= \frac{3^{24}}{17!\ 5!} s^{23}\, e^{-3s} \int_0^s \left(\frac{x}{s}\right)^{17} \left(1 - \frac{x}{s}\right)^5\ d\left(\frac{x}{s}\right)$$

$$= \frac{3^{24}}{17!\ 5!} s^{23}\, e^{-3s}\ \frac{17!\ 5!}{23!} * \int_0^1 \frac{23!}{17!\ \ 5!}\, u^{17}(1-u)^5\, du$$

$$\left[\text{Setting } u = \frac{x}{s}\right]$$

$$= \frac{3^{24}}{23!} s^{23}\, e^{-3s} * 1$$

as the integrand equals the total probability of a $Beta(18, 6)$ *distribution*

$$= \frac{3^{24}}{23!}\ s^{23}\ \ e^{-3s}$$

[Total Marks- 5]

Solution 6 :

Let **N** be the number of claims on a motor insurance policy in one year. Suppose the claim amounts X_1, X_2 … are independent and identically distributed random variables, independent of N. Let **S** be the total amount claimed in one year for that insurance policy.

i. The mean and variance of S are given as:

- $E(S) = E(N)*E(X_1)$
- $Var(S) = E(N)*Var(X_1) + Var(N)*[E(X_1)]^2$.

(1 Mark)

ii. For the i^{th} policy, the first two moments for the number of claims under Option 1 are as below:

- Mean = β_i
- Variance = $\beta_i\ (1 - \beta_i)$

Thus, if S_i denotes the total claim amount under i^{th} policy, we have:

- $E(S_i) = \beta_i \mu$
- $Var(S_i) = \beta_i \sigma^2 + \beta_i (1 - \beta_i) \mu^2$

Therefore, the mean of T is

$$E(T) = E\left(\sum_{i=1}^{100} S_i\right) = \sum_{i=1}^{100} E(S_i) = \mu \sum_{i=1}^{100} \beta_i$$

The variance of T is

$$Var(T) = Var\left(\sum_{i=1}^{100} S_i\right)$$
$$= \sum_{i=1}^{100} Var(S_i) \quad \text{... The policies are independent and thus are } S_i$$
$$= \sum_{i=1}^{100} [\beta_i \sigma^2 + \beta_i (1 - \beta_i) \mu^2]$$

(3 Marks)

iii. For the i^{th} policy, the first two moments for the number of claims under Option 2 are as below:

- Mean = β_i
- Variance = β_i

Thus, if S_i denotes the total claim amount under i^{th} policy, we have:

- $E(S_i) = \beta_i \mu$
- $Var(S_i) = \beta_i \sigma^2 + \beta_i \mu^2$

Therefore, the mean of T' is

$$E(T') = E\left(\sum_{i=1}^{100} S_i\right) = \sum_{i=1}^{100} E(S_i) = \mu \sum_{i=1}^{100} \beta_i$$

The variance of T' is

$$Var(T') = Var\left(\sum_{i=1}^{100} S_i\right)$$
$$= \sum_{i=1}^{100} Var(S_i) \quad \text{... The policies are independent and thus are } S_i$$
$$= \sum_{i=1}^{100} [\beta_i \sigma^2 + \beta_i \mu^2]$$

Comparing with the mean and variance of T:

$$E(T') = \mu \sum_{i=1}^{100} \beta_i = E(T)$$

$$Var(T') = \sum_{i=1}^{100} [\beta_i \sigma^2 + \beta_i \mu^{\,2}]$$

$$= \sum_{i=1}^{100} [\beta_i \sigma^2 + \beta_i (1 - \beta_i) \mu^{\,2}] + \mu^2 \sum_{i=1}^{100} {\beta_i}^2$$

$$= \mathrm{Var}(T) + \underbrace{\mu^2 \sum_{i=1}^{100} {\beta_i}^2}_{>0}$$

$$> Var(T)$$

(4 Marks)

[Total Marks- 8]

Solution 7 :

Consider a random variable *U* that has a uniform distribution on [0, 1] and let *F* be the cumulative distribution function of the standard normal distribution.

Define a random variable X as below:

$$X = \begin{cases} -F^{-1}\left(U + \frac{1}{2}\right) & \text{if } \ 0 \le \mathrm{U} < \frac{1}{2} \\ -F^{-1}\left(U - \frac{1}{2}\right) & \text{if } \ \frac{1}{2} < U \le 1 \end{cases}.$$

i. The range of X for each sub-part is:

- $\begin{cases} \mathrm{U} = 0 \ \Rightarrow \ X = -\mathrm{F}^{-1}(0.5) = 0 \\ \mathrm{U} \to {}^{1}/_{2}- \ \Rightarrow \ \mathrm{X} = -\mathrm{F}^{-1}(1) \to -\infty \end{cases}$
- $\begin{cases} \mathrm{U} \to {}^{1}/_{2}+ \ \Rightarrow \ \mathrm{X} = -\mathrm{F}^{-1}(0) \to +\infty \\ \mathrm{U} = 1 \ \Rightarrow \ X = -\mathrm{F}^{-1}(0.5) = 0 \end{cases}$

For -∞ < x ≤ 0,

$$P[-\infty < X \le x] = P\left[-\infty < -F^{-1}\left(U + \frac{1}{2}\right) \le x\right]$$

$$= P\left[F(-x) \le U + \frac{1}{2} < 1\right]$$

$$= P\left[1 - F(x) - \frac{1}{2} \le U < \frac{1}{2}\right]$$

$$= P\left[\frac{1}{2} - F(x) \le U < \frac{1}{2}\right]$$

$$= \frac{1}{2} - \left[\frac{1}{2} - F(x)\right] \quad \text{... given U follows Uniform}(0,1) \text{ distribution}$$

$$= F(x)$$

For $0 \le x < +\infty$,

$$P[x \le X < +\infty] = P\left[x \le -F^{-1}\left(U - \frac{1}{2}\right) < +\infty\right]$$

$$= P\left[0 < U - \frac{1}{2} \le F(-x)\right]$$

$$= P\left[\frac{1}{2} < U \le \frac{1}{2} + 1 - F(x)\right]$$

$$= P\left[\frac{1}{2} < U \le \frac{3}{2} - F(x)\right]$$

$$= \left[\frac{3}{2} - F(x)\right] - \frac{1}{2} \quad \text{... given U follows Uniform}(0,1) \text{ distribution}$$

$$= 1 - F(x)$$

Thus, X has a standard normal distribution.

(7 Marks)

ii. Given u = 0.619,

$$x = -F^{-1}\left(0.619 - \frac{1}{2}\right)$$

$$= -F^{-1}(0.119)$$

$$= -F^{-1}(1 - 0.881)$$

$$= -F^{-1}\big(1 - F(1.18)\big)$$

$$= -F^{-1}\big(F(-1.18)\big)$$

$$= -(-1.18)$$

$$= 1.18$$

Given u = 0.483,

$$x = -F^{-1}\left(0.483 + \frac{1}{2}\right)$$

$$= -F^{-1}(0.983)$$

$$= -2.12$$

(3 Marks)

[Total Marks- 10]

Solution 8 :

Let θ ε {0, 1 … 10} be the number of weapon-producing nuclear plants in country A.

Let X be the number of nuclear plants found to be producing nuclear weapons by the secret agent. Thus X is a random variable with possible values 0, 1 or 2.

i. The decision making process of country B can be formulated as below:

"**Test H_0: θ = 0 against H_1: θ > 0**"

This is equivalent to testing none of the plants are weapon-producing under null hypothesis against at least 1 plant is weapon-producing under alternate hypothesis.

(1 Mark)

ii. The type of error made by B if she does not invade A when some of the nuclear plants in A are indeed producing weapons is **Type II Error**. **(1 Mark)**

iii. The **critical region** adopted by country B is given as: [x ε {0, 1, 2} : x > 0] as H_0 will be rejected if at least 1 plant is found to be weapon-producing. **(1 Mark)**

iv. The **probability of a Type I error** is given by P[Reject H_0 | H_0 is true] i.e. P[X > 0 |θ = 0].

Given θ = 0, this means there are no weapon producing nuclear plants. This implies X = 0 with probability 1. Therefore, the probability of a Type I error = $P[X > 0 \mid \theta = 0] = 0$

(1 Mark)

[Total Marks- 4]

Solution 9 :

For the study involving '***n***' independent three-toss experiments, the frequencies *a, b, c, d, e, f, g* and *h* of the eight possible outcome sequences and the associate probabilities are as follows:

HHH	**HHT**	**HTH**	**THH**
a	*b*	*c*	*d*
$\frac{1}{2}\,\theta^2$	$\frac{1}{2}\,\theta\,(1-\theta)$	$\frac{1}{2}\,(1-\theta)^2$	$\frac{1}{2}\,\theta\,(1-\theta)$

TTH	**THT**	**HTT**	**TTT**
e	*f*	*g*	*h*
$\frac{1}{2}\,\theta\,(1-\theta)$	$\frac{1}{2}\,(1-\theta)^2$	$\frac{1}{2}\,\theta\,(1-\theta)$	$\frac{1}{2}\,\theta^2$

i. Let S be a random variable denoting the total number of outcomes of type HH or TT observed in the above '*n*' three-toss experiments.

Thus, the observed value of S in terms of frequencies as:

s = # of outcomes of the type HH or TT

= a * #{HHH} + b* #{HHT} + c * #{HTH} + d * #{THH} + e * #{TTH} + f * #{THT} + g * #{HTT} + h * #{TTT}

= a * 2 + b * 1 + c * 0 + d * 1 + e * 1 + f * 0 + g * 1 + h * 2

= 2a + b + d + e + g + 2h **(2 Marks)**

ii. The likelihood equation for the given data will be:

$$L(\theta;\ \underline{X})$$

$$= \left[\frac{1}{2}\,\theta^2\right]^a * \left[\frac{1}{2}\,\theta\,(1-\theta)\right]^b * \left[\frac{1}{2}\,(1-\theta)^2\right]^c * \left[\frac{1}{2}\,\theta\,(1-\theta)\right]^d * \left[\frac{1}{2}\,\theta\,(1-\theta)\right]^e * \left[\frac{1}{2}\,(1-\theta)^2\right]^f * \left[\frac{1}{2}\,\theta\,(1-\theta)\right]^g * \left[\frac{1}{2}\,\theta^2\right]^h$$

$$= \left(\frac{1}{2}\right)^{a+b+c+d+e+f+g+h} * \theta^{2a+b+d+e+g+2h} * (1-\theta)^{b+2c+d+e+2f+g}$$

$$\propto\ \theta^s(1-\theta)^{2n-s}$$

NB:

- **s** = 2a + b + d + e + g + 2h
- **n** = a + b + c + d + e + f + g + h
- **2n – s** = 2*(a + b + c + d + e + f + g + h) – (2a + b + d + e + g + 2h)
  = b + 2c +d + e + 2f + g

Taking logarithm of the likelihood function,

$$l(\theta;\ \underline{X}) = \text{constant} + s * log_e(\theta) + (2n-s) * log_e(1-\theta)$$

Differentiating w. r. t. $\boldsymbol{\theta}$ we get:

$$\frac{\partial l}{\partial \theta} = \frac{s}{\theta} - \frac{2n-s}{1-\theta}$$

Solving: $\frac{\partial l}{\partial \theta} = 0$, we get

$$\hat{\theta}_{MLE} = \frac{s}{2n}$$

$$\left[Check:\quad \frac{\partial^2 l}{\partial \theta^2} = -\frac{s}{\theta^2} - \frac{2n-s}{(1-\theta)^2}\bigg|_{\hat{\theta}_{MLE}} < 0 \quad \Rightarrow maximum\right]$$

(4 Marks)

iii. For large n, $\hat{\theta}_{MLE}$ is approximately normal, and is unbiased with variance given by the Cramer-Rao lower bound, that is:

$$\hat{\theta}_{MLE} \approx N(\theta, CRLB)$$

$$\text{where,} \quad \text{CRLB} = \frac{1}{-E\left[\frac{\partial^2 l}{\partial \theta^2}\right]}$$

Now,

$$-E\left[\frac{\partial^2 l}{\partial \theta^2}\right] = E\left[\frac{S}{\theta^2} + \frac{2n - S}{(1-\theta)^2}\right]$$

$$= \frac{1}{\theta^2} * E(S) + \frac{1}{(1-\theta)^2} * [2n - E(S)]$$

$$= \frac{1}{\theta^2} * 2n\theta + \frac{1}{(1-\theta)^2} * [2n - 2n\theta] \ \ldots since\ E(S) = 2n\theta$$

$$= 2n\left[\frac{1}{\theta} + \frac{1}{1-\theta}\right]$$

$$= \frac{2n}{\theta\,(1-\theta)}$$

The asymptotic variance of the MLE of θ is **θ (1 – θ) / 2n.**
Thus, the approximate asymptotic distribution of the MLE of θ is:

$$\hat{\theta}_{MLE} \approx N\left[\theta, \frac{\theta\,(1-\theta)}{2n}\right]$$

(3 Marks)

iv. In a particular study involving 1000 three-toss experiments, the observed frequencies were:

a	***b***	***c***	***d***	***e***	***f***	***g***	***h***
135	131	125	125	123	115	125	121

Here:

- n = 1000
- s = 2a + b + d + e + g + 2h = 1016

$$\hat{\theta}_{MLE} = \frac{s}{2n} = \frac{1016}{2000} = 0.508$$

A large-sample 95% confidence interval for θ is given as:

$$\hat{\theta}_{MLE} \pm z_{0.025} * \sqrt{\frac{\hat{\theta}_{MLE}\left(1 - \hat{\theta}_{MLE}\right)}{2n}}$$

$$= \ 0.508 \ \pm 1.96 * \sqrt{\frac{0.508 * 0.492}{2000}}$$

$$= \ 0.508 \ \pm 0.022$$

$$= \ (0.486, 0.530)$$

(4 Marks)

v. It has been suggested that the above model is incorrect in its assumption that the probability of a head on the first toss is 0.5.

In order to test this, we can set up the following hypothesis testing problem:

H_0: p = 0.5 *against* H_1: p ≠ 0.5

Here: p = probability of getting a head on first toss.

Let X denote the number of heads observed in the first toss in the above n (=1000) 3-toss experiments. As a random variable, X follows Binomial (n, p) distribution.

Given ***n*** is large and so using Normal approximation, a 95% asymptotic confidence interval for 'p' is stated as below:

$$\hat{p} \ \pm 1.96 * \frac{\sqrt{n\hat{p}(1-\hat{p})}}{n} \qquad \text{where } \hat{p} = \frac{X}{n}$$

The given data reveals that: X = a + b + c + g = 516. This means $\hat{p} = 0.516$.

The 95% confidence interval: 0.516 ± 0.031 = (0.485, 0.547) contains the value 0.5. Hence, we can conclude that the criticism does **not** hold water at the 5% level of significance.

(5 Marks)

[Total Marks-18]

Solution 10 :

i. Comments on the plot

- The centers of the distributions differ for all the four cities. Thus there is a prima facie case for suggesting that the underlying means are different.
- The difference between the mean time taken to commute to office in peak hours and non-peak hours are in the order City A (highest), City D (lowest).
- The variation in the data for City C is *lowest* compared to City D which appears to be *highest*. However, with only 7 observations for each city, we cannot be sure that there is a real underlying difference in variance.

(2 Marks)

ii. Following are the assumptions underlying analysis of variance:

- The populations must be **normal**.
- The populations have a **common variance**.
- The observations are **independent**.

(2 Marks)

iii. We are carrying out the following test:

H_0: The mean of differences is same for each city

against

H_1: The mean of differences are not the same for all of the cities

To carry out the ANOVA, we must first compute the Sum of Squares

$$\mathrm{SS_T} = 1{,}495 - \frac{103^2}{28} = 1{,}116.11$$

$$\mathrm{SS_B} = \frac{1}{7}\left(64^2 + 44^2 + 8^2 + (-13)^2\right) - \frac{103^2}{28} = 516.11$$

$$SS_R = SS_T - SS_B = 600.00$$

The ANOVA table is:

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	24	600.00	25.00	
Total	27	1,116.11		

Under $H_0, F = \frac{172.04}{25.00} = 6.88$, using the $F_{3,24}$ distribution.

The 5% critical point is 3.009, so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that there are underlying differences between the cities.

(4 Marks)

iv. Analysis of the mean differences

Since, $\bar{y}_{1*} = 9.14;\ \bar{y}_{2*} = 6.29;\ \bar{y}_{3*} = 1.14;\ \bar{y}_{4*} = -1.86$

we can write:

$$\bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$$

$$\hat{\sigma}^2 = \frac{SS_R}{n-k} = 25$$

The least significant difference between any pair of means is:

$$t_{24,0.025} * \hat{\sigma}\sqrt{\left(\frac{1}{7}+\frac{1}{7}\right)} = 2.064 * \sqrt{25} * \sqrt{\left(\frac{1}{7}+\frac{1}{7}\right)} = 5.52$$

Now we can examine the difference between each of the pairs of means. If the difference is less than the least significant difference then there is no significant difference between the means.

We have:

$\bar{y}_{1*} - \bar{y}_{2*} = 2.85;\ \bar{y}_{2*} - \bar{y}_{3*} = 5.15;\ \ \bar{y}_{3*} - \bar{y}_{4*} = 3.00$

Observing that all these 3 differences are less than 5.52, we underline these pairs to show that they have no significant difference:

$$\underline{\bar{y}_{1*} > \bar{y}_{2*}} > \underline{\bar{y}_{3*}} > \underline{\bar{y}_{4*}}$$

Examining to see if the first two groups can be combined:

$\bar{y}_{1*} - \bar{y}_{3*} = 8.00$

There is a significant between means 1 and 3, so we cannot combine the first two groups.

Examining to see if the last two groups can be combined:

$\bar{y}_{2*} - \bar{y}_{4*} = 8.15$

There is a significant between means 2 and 4, so we cannot combine the last two groups.

Therefore the diagram remains as before.

(6 Marks)

[Total Marks-14]

Solution 11 :

i. Fitted Linear Regression Equation

The relevant summary statistics to fit the equation are:

$\sum x$ = 385.2; $\sum x^2$ = 12,666.58;
$\sum y$ = 1,162.5; $\sum y^2$ = 119,026.9;
$\sum xy$ = 38,191.41; n = 12.

$$S_{xx} = \sum x^2 - n\overline{x}^2 = 12666.58 - 12 * \left(\frac{385.2}{12}\right)^2 = 301.66$$
$$S_{xy} = \sum xy - n\overline{xy} = 38191.41 - 12 * \left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right) = 875.16$$
$$S_{yy} = \sum y^2 - n\overline{y}^2 = 119026.90 - 12 * \left(\frac{1162.5}{12}\right)^2 = 6409.71$$

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{875.16}{301.66} = 2.90$$
$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{1162.5}{12}\right) - 2.90 * \left(\frac{385.2}{12}\right) = 3.78$$

Therefore, the fitted regression line is: $\mathbf{y = \hat{\alpha} + \hat{\beta}x = 3.78 + 2.90\,x}$

(4 Marks)

ii. Confidence interval for β

Assuming normal errors with a constant variance:

95% confidence interval for β: $\hat{\beta} \pm t_{n-2}(2.50\%) * s.e.(\hat{\beta})$

Here: $s.e.(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[S_{yy} - \frac{{S_{xy}}^2}{S_{xx}}\right] = 387.07$$

$$s.e.(\hat{\beta}) = \sqrt{\frac{387.07}{301.66}} = 1.13$$

95% confidence interval for β: $2.90 \pm 2.228 * 1.13 = (0.38, 5.42)$

(5 Marks)

iii. <u>95% confidence intervals for the mean IBM share price</u>

$$\hat{y}_{x_0} \pm t_{n-2}(2.50\%)\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

The Dell Share price is US $ 40 (x_0).

$\hat{y}_{x_0} = 3.78 + 2.90 * 40 = 119.78$

Thus, 95% Confidence interval:

$= 119.78 \pm 2.228 * \sqrt{387.07 * \left[\frac{1}{12} + \frac{(40 - 32.1)^2}{301.66}\right]}$

$= 119.78 \pm 2.228 * 10.5989$

$= (96.17, 143.39)$

(4 Marks)

[Total Marks-13]

Solution 12 :

We are carrying out the following test:

$H_0: \mu = 300 \quad v/s \quad H_1: \mu < 300$

Under H_0,

$$\frac{\overline{X} - 300}{30/\sqrt{n}} \sim N(0,1)$$

The test statistic is

$$\frac{290 - 300}{30/\sqrt{60}} = -2.5819$$

From the Tables, -2.3263 is the critical value for a left-sided one-tailed 1% test. The test statistic of -2.5819 is ***less*** than this critical value, so we do not have sufficient evidence to accept H_0. Therefore, it can be concluded that the quality control suspicions are true at the 1% level of significance.

[Total Marks-4]

XXXXXXXXXXXXXXXXXXXX