Institute of Actuaries of India

Subject CT3 – Probability & Mathematical Statistics

May 2011 Examinations

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 examiners have given credit for any alternative approach or interpretation which they consider to be reasonable.

Q. 1) From the definition of the normal density, we have

$$P[Z > a] = \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Let x = z – a.

$$P[Z > a] = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-(x+a)^2/2} dx$$

$$= \frac{e^{-a^2/2}}{\sqrt{2\pi}} \int_0^\infty e^{-x^2/2}\, e^{-xa} dx$$

$$\leq \frac{e^{-a^2/2}}{\sqrt{2\pi}} \int_0^\infty e^{-x^2/2}\, dx \qquad \text{since } e^{-xa} \leq 1 \text{ for all x } \varepsilon \text{ [0, } \infty) \text{ and a > 0}$$

Now using the fact $\int_0^\infty e^{-x^2/2} dx = \sqrt{\frac{\pi}{2}}$

we see that

$$P[Z > a] \leq \frac{1}{2} e^{-a^2/2}$$

[Total: 5]

Q2 Let X denotes the variable stating the gross income in units of 1,000.

We are given:

- $\sum x = 321.6$
- $\sum x^2 = 10628.31$

This implies:

- $\bar{x} = \frac{1}{10} \sum x = 32.16$
- $s_x = \sqrt{\frac{\sum x^2 - 10\bar{x}^2}{10-1}} = 5.6338$

Let Y denote the variable stating the income net of income tax in units of 1,000.

We have: Y = X – 30% * (X – 10)
= 0.7 * X + 3

Note this formula holds only if no gross income is lower than 10,000. We assume this holds for rest of the computations.

This means:

- $\bar{y} = 0.7 * \bar{x} + 3$
  $= 0.7(32.16) + 3$
  $= 25.512$
- $s_y = 0.7 * s_x$
  $= 0.7(5.6338)$
  $= 3.944$

Thus, the sample mean of the 10 net incomes is 25,500 while the standard deviation of the same is 3,900. (The numbers are rounded to nearest hundred as specified in the problem)

[Total: 3]

Q3 As R has values in (0, 1), the density function $f_R(r)$ is non-zero only for r ε (0, 1).

$$F_R(r) = P(R \le r) = P\left(X \le \frac{1}{2}, \frac{X}{1-X} \le r\right) + P\left(X > \frac{1}{2}, \frac{1-X}{X} \le r\right)$$

$$= P\left(X \le \frac{1}{2}, X \le \frac{r}{r+1}\right) + P\left(X > \frac{1}{2}, X \ge \frac{1}{r+1}\right)$$

$$= P\left(X \le \frac{r}{r+1}\right) + P\left(X \ge \frac{1}{r+1}\right) \qquad \left[\because \frac{r}{r+1} \le \frac{1}{2} \; and \; \frac{1}{r+1} \ge \frac{1}{2}\right]$$

$$= \frac{r}{r+1} + \left[1 - \frac{1}{r+1}\right] \qquad [\because X \sim U(0, 1)]$$

$$= \frac{2r}{r+1}$$

For r ε (0, 1), the density for R equals to

$$f_R(r) = \frac{d}{dr} F_R(r)$$

$$= \frac{2}{r+1} - \frac{2r}{(r+1)^2}$$

$$= \frac{2}{(r+1)^2}$$

[Total: 5]

Q4 The probability function of X is

$$P(X = k) = \frac{e^{-2}2^k}{k!} \; for \; k = 0,1,2 \ldots . \quad \text{given that } \mu_x = 2$$

The general probability function of a geometric distribution on 0, 1, 2… is of the form

$$P(Y = k) = p(1-p)^k \; for \; k = 0, 1, 2 \ldots$$

This has a mean $\mu_Y = \frac{1-p}{p}$

Given $\mu_Y = 2$ we have

$$\frac{1-p}{p} = 2 \quad \Rightarrow \quad p = \frac{1}{3}$$

Thus,

$$P(Y = k) = \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^k \; for \; k = 0, 1, 2 \ldots$$

Now,

$$P(X = Y) = \sum_{k=0}^{\infty} P(X = Y = k)$$

$$= \sum_{k=0}^{\infty} P(X = k)\, P(Y = k) \quad as \; X \; and \; Y \; are \; independent$$

$$= \sum_{k=0}^{\infty} \frac{e^{-2}2^k}{k!} \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^k$$

$$= \frac{e^{-2}}{3} \sum_{k=0}^{\infty} \frac{\left(4/3\right)^k}{k!}$$

$$= \frac{e^{-2}}{3} \, e^{4/3} \qquad as \; e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \; with \; x = \frac{4}{3}$$

$$= \frac{1}{3} e^{-2/3}$$

[Total: 5]

Q5 Let X is the outcome of the treasure hunt. Note that X ≥ 0.

We have F(0) = 0.8.

For $1000 \le x \le 5000$, $F(x) = 0.8 + \frac{0.2}{5000-1000}(x - 1000)$
$= 0.75 + 0.00005x.$

Thus, $F(t) = \begin{cases} 0 & for\ x < 0 \\ 0.8 & for\ 0 \le x < 1000 \\ 0.75 + 0.00005x & for\ 1000 \le x \le 5000 \\ 1 & for\ x > 5000 \end{cases}$

By inverse transformation method we will have for u generated randomly from U (0, 1):

$$x = F^{-1}(u) = \begin{cases} 0 & if\ 0 \le u < 0.8 \\ 20000\ (u - 0.75) & if\ 0.8 \le u \le 1 \end{cases}$$

The random samples for the first five trials are 0.95, 0.65, 0.75, 0.55 and 0.85.

$u_1 = 0.95 \Rightarrow x_1 = 20000\ (0.95 - 0.75) = 4000$

$u_2 = 0.65 \Rightarrow x_2 = 0$

$u_3 = 0.75 \Rightarrow x_3 = 0$

$u_4 = 0.55 \Rightarrow x_4 = 0$

$u_5 = 0.85 \Rightarrow x_5 = 20000\ (0.85 - 0.75) = 2000$

Thus the average of the outcomes $= \frac{4000+0+0+0+2000}{5} = 1200.$

[Total: 3]

Q6

(a) S is approximately Normal for large n by Central Limit Theorem.

Using results from the table for $X_i \sim U(-0.5, 0.5)$

- $E[S] = n\,E[X_i] = n * 0 = 0$
- $Var[S] = n\,Var[X_i] = n/12$

So, $S \sim N(0, n/12)$

(b) S approx $\sim N(0, 1)$ if $n = 12$.

(c) The distribution of each X_i is symmetric and so the distribution of the sum $S = \sum_{i=0}^{n} X_i$ is also symmetric.

So skewness is zero, as for any normal distribution.

[Total: 4]

Q7 Note here the random variable X takes values at discrete points and continuous regions with probability greater than zero.

Let $P[X < 1] = c$

For $x < 1$,

$$x^2 = F[x \mid X < 1] = \frac{F(x)}{P(X<1)} = \frac{F(x)}{c}$$

$$\therefore F(x) = cx^2 \; for \; 0 \leq x < 1$$

Thus $F(1) = P(X < 1) + P(X = 1) = c + 0.25$

$\therefore P(X > 1) = 1 - F(1) = 0.75 - c$

For $1 < x \leq 2$,

$$x - 1 = F(x \mid X > 1) = \frac{F(x) - F(1)}{P(X>1)}$$

$\therefore F(x) = 1 - (0.75 - c)(2 - x)$ for $1 < x \leq 2$

Then,

$$S(x) = 1 - F(x) = \begin{cases} 1 - cx^2 & for \; 0 \leq x < 1 \\ (0.75 - c)(2 - x) & for \; 1 \leq x \leq 2 \end{cases}$$

$E(X) = \int_0^2 S(x)dx$

$$= \int_0^1 (1 - cx^2)\, dx + \int_1^2 (0.75 - c)(2 - x)\, dx$$

$$= \frac{33-20c}{24}$$

Given E(X) = 1, we have $\frac{33-20c}{24} = 1$ which means $c = 0.45$

Therefore, F(1) = c + 0.25 = 0.70

[Total: 5]

Q8 The observed value of X is 200.

This problem is equivalent to finding the θ_1 and θ_2 such that

- P(X ≥ 200) = 0.025 under Poisson(θ_1)
- P(X ≤ 200) = 0.025 under Poisson(θ_2)

Under Poisson(θ_1), we have:

$$P(X \geq 200) = 0.025$$
$$\Leftrightarrow P(X > 199.5) = 0.025 \quad \textit{using continuity correction}$$
$$\Leftrightarrow P\left(Z > \frac{199.5 - \theta_1}{\sqrt{\theta_1}}\right) = 0.025 \quad \textit{using Normal approximation with } Z \sim N(0,1)$$
$$\Leftrightarrow \frac{199.5 - \theta_1}{\sqrt{\theta_1}} = 1.96 \quad \because \Phi(1.96) = 0.975 = 1 - 0.025$$

Solving for θ_1:

$$(199.5 - \theta_1)^2 = 1.96^2\, \theta_1$$
$$\Leftrightarrow \theta_1^2 - 402.8416\, \theta_1 + 39800.25 = 0$$
$$\Leftrightarrow \theta_1 = \frac{402.8416 \pm 55.50094}{2} = 173.67 \textit{ or } 229.17$$

Given this will be the lower bound, θ_1 = 173.67.

Under Poisson(θ_2), we have:

$$P(X \leq 200) = 0.025$$
$$\Leftrightarrow P(X < 200.5) = 0.025 \quad \textit{using continuity correction}$$
$$\Leftrightarrow P\left(Z < \frac{200.5 - \theta_2}{\sqrt{\theta_2}}\right) = 0.025 \quad \textit{using Normal approximation with } Z \sim N(0,1)$$
$$\Leftrightarrow \frac{200.5 - \theta_2}{\sqrt{\theta_2}} = -1.96 \quad \because \Phi(-1.96) = 0.025$$

Solving for θ_2:

$$(200.5 - \theta_2)^2 = (-1.96)^2\, \theta_2$$

$$\Leftrightarrow \theta_2^2 - 404.8416\,\theta_2 + 40200.25 = 0$$
$$\Leftrightarrow \theta_2 = \frac{404.8416 \pm 55.6392}{2} = 174.60 \; or \; 230.24$$

Given this will be the upper bound, $\theta_2 = 230.24$.

Thus, the approximate 95% symmetrical confidence interval for θ will be (173.67, 230.24).

[Total: 7]

Q9 We have $f(x) = \frac{\theta}{(\theta+x)^2}; \quad 0 < x < \infty, \; \theta > 0$

$$\ln f(x) = \ln\theta - 2\ln(\theta + x)$$

$$\frac{\partial lnf(x)}{\partial\theta} = \frac{1}{\theta} - \frac{2}{\theta+x}$$

$$\frac{\partial^2 lnf(x)}{\partial\theta^2} = -\frac{1}{\theta^2} + \frac{2}{(\theta+x)^2}$$

Thus, $E\left[\frac{\partial^2 lnf(x)}{\partial\theta^2}\right] = -\frac{1}{\theta^2} + \int_0^\infty \frac{2\theta}{(\theta+x)^4} dx$

$$= -\frac{1}{\theta^2} + \left[\frac{-2\theta}{3(\theta+x)^3}\right]_0^\infty$$

$$= -\frac{1}{\theta^2} + \frac{2}{3\theta^2} = -\frac{1}{3\theta^2}.$$

We have got a sample of n observations from this distribution.

By properties of asymptotic distributions of MLE $\hat{\theta}$ of θ, we know:

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

Here, $CRLB = \frac{1}{n\,E\left[-\frac{\partial^2 lnf(x)}{\partial\theta^2}\right]}$

$$= \frac{3\theta^2}{n}.$$

Thus the asymptotic variance of the maximum likelihood estimator of θ is $\frac{3\theta^2}{n}$.

[Total: 5]

Q10 We have the random variable X with a density function:

$$f(x\,|\theta) = \frac{\Gamma(3\theta)}{\Gamma(\theta)\Gamma(2\theta)} x^{\theta-1}(1-x)^{2\theta-1}$$

(a) The mode of this distribution will be that value of x where the first derivative of f(x) vanishes. Computing this derivative the expression $\frac{df}{dx} = 0$ implies:
(Ignoring constant factors)

$$\frac{df}{dx}(x) = (\theta-1)x^{\theta-2}(1-x)^{2\theta-1} + (2\theta-1)x^{\theta-1}(1-x)^{2\theta-2}(-1) = 0$$
$$\Longrightarrow x^{\theta-2}(1-x)^{2\theta-2}[(2-3\theta)x + (\theta-1)] = 0$$

This can be solved for x_m that makes this an equality and gives

$$x_m = \frac{\theta-1}{3\theta-2} \; assuming \; \theta \neq \frac{2}{3}$$

One can check that at this value of x, the second derivative is indeed negative.

We need to ensure that $0 \leq \frac{\theta-1}{3\theta-2} \leq 1$. This means $\theta \geq 1$.

Note: In case $\theta = 1$, the mode is at $x = 0$.

(b) The most powerful test will be given by the Neyman-Pearson Lemma using the Likelihood Ratio technique. It states that if C is a critical region of size α and there exists a constant k such that $L_0/L_1 \leq k$ inside C and $L_0/L_1 \geq k$ outside C, then C is a most powerful critical region of size α for testing the simple hypothesis $H_0: \theta = 1$ against $H_1: \theta = 2$.

Thus most powerful critical region will be of the form:

$$\frac{Likelihood\ under\ H_0}{Likelihood\ under\ H_1} < crtical\ value$$

We have: $L_0 = 2(1-x)$ & $L_1 = 20x(1-x)^3$

$$\frac{L_0}{L_1} = \frac{2(1-x)}{20x(1-x)^3} = \frac{1}{10x(1-x)^2}$$

Thus the critical region: $\left\{x \in (0,1): \frac{1}{10x(1-x)^2} < k_1\right\}$ for some constant k_1
Or, $\{x \in (0,1): x(1-x)^2 > k\}$ where $k = \frac{1}{10k_1}$.

(c) X has been observed as $1/7$.

Here given the sample observed we are trying to compute the exact shape of such a critical region so that it can be used to compute the p-value as desired in part (d).

Looking at the critical region we obtained in (b), the test should take the form:

Reject H_0 if

$$X(1-X)^2 - \frac{1}{7}\left(1-\frac{1}{7}\right)^2 > 0$$
$$\Leftrightarrow X(1-X)^2 - \frac{36}{343} > 0$$
$$\Leftrightarrow \left(X-\frac{1}{7}\right)\left(X-\frac{4}{7}\right)\left(X-\frac{9}{7}\right) > 0$$

This holds only if $\frac{1}{7} < X < \frac{4}{7}$

Thus we get: $Reject\ H_0\ if\ X \in \left(\frac{1}{7}, \frac{4}{7}\right)$

(d) The p-value for the most powerful test will be

$$P_{H_0}(Reject\ H_0) = P_{H_0}\left(\frac{1}{7} < X < \frac{4}{7}\right)$$
$$= \int_{\frac{1}{7}}^{\frac{4}{7}} 2(1-x)\,dx$$
$$= \frac{27}{49}$$

[Total: 10]

Q11 The moment generating function of X

$$M(t) = E[e^{tX}]$$

$$= \int_{-\infty}^{\infty} e^{tx} f(x)\ dx$$

$$= \frac{1}{2}\left[\int_{-\infty}^{0} e^{(1+t)x}\ dx + \int_{0}^{\infty} e^{-(1-t)x}\ dx\right]$$

$$= \frac{1}{2}\left[\frac{1}{1+t} + \frac{1}{1-t}\right] \quad \text{for } |t| > 1$$

$$= \frac{1}{1-t^2}$$

Now,

$$M'(t) = (-1)(1-t^2)^{-2}(-2t) = 2t(1-t^2)^{-2}$$

$$M''(t) = 2(1-t^2)^{-2} + 2t(1-t^2)^{-3}(-2t) = 2(1-t^2)^{-2} - 4t^2(1-t^2)^{-3}$$

Thus,

$$E(X) = M'(0) = 0$$

$$\text{Var(X)} = \text{E}(\text{X}^2) - [\text{E(X)}]^2$$

$$= M''(0) - 0$$

$$= 2$$

[Total: 5]

Q12

(a) The likelihood function is

$$L(\theta; x) = \begin{cases} \frac{2^n}{\theta^{2n}} \prod_{i=1}^{n} x_i & 0 \le \min_i x_i \le \max_i x_i \le \theta \\ 0 & otherwise \end{cases}$$

Here $L(\theta; x) = 0$ when $\theta < \max_i X_i$
For $\theta \ge \max_i X_i$, it is clear that $L(\theta; x)$ decreases as θ increases.
Hence, $L(\theta; x)$ obtains its maximum at $\max_i X_i$, i.e., MLE $\hat{\theta} = \max_{1 \le i \le n} X_i$.

(b) Let Y = $\max_{1\le i\le n} X_i$

For $0 \le y \le \theta$

$$F_Y(y) = P[\, Y \le y \,]$$
$$= P\left[\max_{1\le i\le n} X_i \le y\right]$$
$$= P[(X_1 \le y) \cap (X_2 \le y) \cap \dots \cap (X_n \le y)]$$
$$= \prod_{i=1}^{n} P[X_i \le y]$$
$$= \prod_{i=1}^{n} \int_0^y 2\theta^{-2} x\, dx$$
$$= \left(\frac{y}{\theta}\right)^{2n}$$

Thus,

$$f_y(y) = \frac{d}{dy}\left[\left(\frac{y}{\theta}\right)^{2n}\right] = \frac{2ny^{2n-1}}{\theta^{2n}} \; for\; 0 \le y \le \theta$$

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var(\hat{\theta}) + \left[E(\hat{\theta}) - \theta\right]^2$$

$$E(\hat{\theta}) = \int_0^\theta y\, \frac{2n}{y} \left(\frac{y}{\theta}\right)^{2n} dy = \int_0^\theta \frac{2n}{\theta^{2n}}\, y^{2n}\, dy = \frac{2n\theta}{2n+1}$$

$$E(\hat{\theta}) - \theta = -\frac{\theta}{2n+1}$$

$$E(\hat{\theta}^2) = \int_0^\theta y^2\, \frac{2n}{y} \left(\frac{y}{\theta}\right)^{2n} dy = \int_0^\theta \frac{2n}{\theta^{2n}}\, y^{2n+1}\, dy = \frac{n\theta^2}{n+1}$$

Thus,

$$Var(\hat{\theta}) = E(\hat{\theta}^2) - [E(\hat{\theta})]^2 = \frac{n\theta^2}{n+1} - \frac{4n^2\theta^2}{(2n+1)^2} = \frac{n\theta^2}{(n+1)(2n+1)^2}$$

Therefore,

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[E(\hat{\theta}) - \theta\right]^2 = \frac{n\theta^2}{(n+1)(2n+1)^2} + \frac{\theta^2}{(2n+1)^2} = \frac{\theta^2}{(n+1)(2n+1)}$$

[Total: 8]

Q13 For simplicity, it is assumed that the decisions $X_1, X_2, \ldots X_n$ of the different voters are iid $\mathfrak{B}(p)$.

Then we know that
$\left[\mu_n - \frac{2\sigma_n}{\sqrt{n}},\ \mu_n + \frac{2\sigma_n}{\sqrt{n}}\right]$ is the 95% confidence interval for μ.

We have
$\sigma_n = Var(X_n) = p\,(1-p) \leq \frac{1}{4}$
Thus,
$\left[\mu_n - \frac{1}{\sqrt{n}},\ \mu_n + \frac{1}{\sqrt{n}}\right]$ contains the 95% confidence interval for μ.
If we want the error to be ±3%,
then $\left|\frac{1}{\sqrt{n}}\right| \leq 3\%$
$n \geq \left(\frac{1}{0.03}\right)^2 = 1112$

[Total: 3]

Q14

(a) The Neyman-Pearson theory states that the most powerful test for H_0 against H_1would be that one which for given α = 0.3 and critical region C which maximises $P_{H_1}(X \in C)$

Choices of C for which $P_{H_0}(X \epsilon C) = \alpha = 0.3$ $are\ C = \{1,3\},\{1,7\},\{2\},\{5\}$

It is clear that max $P_{H_1}(X \epsilon C) = 0.6$ when C = {1,3}
This means the most powerful critical region C = {1,3}

Therefore, the most powerful test is the one which rejects H_0 if X=1 or 3
The power of this test = $P_{H_1} = (X = 1\ or\ X = 3)$
$= 0.3 + 0.3 = 0.6$

(b) In contrast, the power of the test which rejects H_0 if X = 5 $P_{H_1}(X = 5) = 0.2$
This test is clearly weaker than the most powerful one which gives power of 0.6

[Total: 5]

Q15

(a) We have $\ln(X_1) \sim N(\mu, \sigma^2)$ under the proposed model "M".

$$f_1(\mu,\sigma) = P[\ln(X_1) \le 1]$$
$$= P\left[Z \le \frac{1-\mu}{\sigma}\right] \; with\; Z = \frac{\ln(X_1)-\mu}{\sigma} \sim N(0,1)$$
$$= \Phi(\tfrac{1-\mu}{\sigma})$$

For j = 2, 3 … 9:

$$f_j(\mu,\sigma) = P\left[\frac{j}{2} < \ln(X_1) \le \frac{j+1}{2}\right]$$
$$= P\left[\ln(X_1) \le \frac{j+1}{2}\right] - P\left[\ln(X_1) \le \frac{j}{2}\right]$$
$$= P\left[Z \le \frac{\frac{j+1}{2}-\mu}{\sigma}\right] - P\left[Z \le \frac{\frac{j}{2}-\mu}{\sigma}\right]$$
$$= \Phi\left(\frac{\frac{j+1}{2}-\mu}{\sigma}\right) - \Phi(\tfrac{\frac{j}{2}-\mu}{\sigma})$$

$$f_{10}(\mu,\sigma) = P[\ln(X_1) > 5]$$
$$= P\left[Z > \frac{5-\mu}{\sigma}\right]$$
$$= 1 - \Phi(\tfrac{5-\mu}{\sigma})$$

(b) We are told μ = 3 and σ = 0.6

$$E_5 = 900 * f_5(3, 0.6)$$
$$= 900 * [\,\Phi\left(\tfrac{3-3}{0.6}\right) - \Phi\left(\tfrac{2.5-3}{0.6}\right)\,]$$
$$= 900 * [\Phi(0) - \Phi(-0.8333)]$$
$$= 900 * [0.5 - (1 - 0.79673)]$$
$$= 900 * 0.29673$$
$$= 267.057$$

Given that μ = 3, we observe that the intervals are symmetric around 3. So, we can say:

$E_6 = E_5 = 267.057$
$E_7 = E_4 = 140.229$
$E_8 = E_3 = 37.125$
$E_9 = E_2 = 5.202$
$E_{10} = E_1 = 0.387$

Observe $\sum_{j=1}^{10} E_j = 900$ which is as expected.

(c) The chi-square goodness-of-fit test is to test:

H_0: Data is a sample from the model "M"

v/s H_1: Data is not a sample from the model "M"

Observed = O_i

Expected = E_i

Chi-Square = $\frac{(O_i - E_i)^2}{E_i}$

The expected numbers in the first and the last group are too small; hence we combine them individually with the immediate next group so that the expected frequencies for all cells is at least 5.

Interval	Observed	Expected	Chi-Square
(-∞, 1.5]	7.000	5.589	0.356
(1.5, 2]	40.000	37.125	0.223
(2, 2.5]	121.000	140.229	2.637
(2.5, 3]	290.000	267.057	1.971
(3, 3.5]	250.000	267.057	1.089
(3.5, 4]	156.000	140.229	1.774
(4, 4.5]	34.000	37.125	0.263
(4.5, ∞)	2.000	5.589	2.305
	900.000	**900.000**	**10.618**

The chi-square statistic for this goodness-of-fit is $\chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = 10.618$

The degrees of freedom (d.f.) = (10 - 2) – 1 – 0 = 7

[Note: Combined two pairs of groups & no estimation of any estimator is required]

We will compute $p\ value = P[\mathrm{X}_7^2 > 10.618]$

From the tables: $F_{\chi_7^2}[10.5] = 0.8380$ & $F_{\chi_7^2}[11] = 0.8614$

By linear interpolation:

$$F_{\chi_7^2}[10.618] = 0.8380 + \left(\frac{10.618 - 10.5}{11 - 10.5}\right) * (0.8614 - 0.8380) = 0.84352$$

Thus, $p\ value = P[\mathrm{X}_7^2 > 10.618] \approx 0.15648$

[Using CHIDIST function of MS Excel, p value = 0.15619]

Therefore, there is not enough evidence to reject H_0.

(d) Here, we would have estimated the two parameters μ and σ. This means we will lose 2 more degrees of freedom and d.f. = (10 – 2) – 1 – 2 = 5.

We are given revised $\chi^2 = 11.355$
We will need to compute $p\ value = P[X_5^2 > 11.355]$

From the tables: $F_{\chi_5^2}[11] = 0.9486$ & $F_{\chi_5^2}[11.5] = 0.9577$

By linear interpolation:

$$F_{\chi_5^2}[11.355] = 0.9486 + \left(\frac{11.355-11}{11.5-11}\right) * (0.9577 - 0.9486) = 0.95506$$

Thus, $p\ value = P[X_5^2 > 11.355] \approx 0.04494$

[Using CHIDIST function of MS Excel, p value = 0.04478]

Here the p-value has significantly decreased compared to earlier circumstance and at this value there might be some evidence that the null hypothesis might not hold.

[Total: 10]

Q16

(a) The variable x takes values 1, 2 …. 30. Sample size n = 30.

$$\begin{aligned} S_{xx} &= \sum(x_i - \bar{x})^2 \\ &= \sum x_i^2 - n\bar{x}^2 \\ &= \frac{n(n+1)(2n+1)}{6} - n\left[\frac{n(n+1)}{2n}\right]^2 \\ &= \frac{n(n^2-1)}{12} \\ &= \frac{(30)(30^2-1)}{12} \\ &= 2247.5 \end{aligned}$$

[Alternatively, this can be derived from first principles]

Hence $\hat{\beta} = S_{xy}/S_{xx} = \frac{599.62}{2247.5} = 0.2668$

(b) $\hat{\sigma}^2 = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$

This can be written as

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - S_{xy}^2/S_{xx}\right)$$

$$= \frac{1}{28}\left(1020.6 - \frac{599.62^2}{2247.5}\right) = 30.73$$

(c) Using $E[SS_{RES}] = (n-2)\sigma^2$, SS_{RES} value is as derived above 860.48

(d) We have $SS_{TOTAl} = S_{yy}$
Hence $SS_{REG} = S_{yy} - SS_{RES} = 1020.6 - 860.48 = 160.12$

The proportion of the total variability of the responses "explained" by a model is called the coefficient of determination, denoted R^2, and is estimated by

$$R^2 = SS_{REG}/SS_{TOT} = 160.12/1020.6 = 15.7\%$$

(e) We are testing $H_0: \beta = 0$ the "no linear relationship" hypothesis $vs\ H_1: \beta \neq 0$

We know the test statistic

$$t = (\hat{\beta} - \beta)/se(\hat{\beta})$$

has t-distribution with n – 2 degrees of freedom, where

$se(\hat{\beta})$ denotes the estimated standard error of $\hat{\beta} = \sqrt{\hat{\sigma}^2/S_{xx}}$

Under null hypothesis $H_0: \beta = 0$,

$$t = 0.2668/\sqrt{\hat{\sigma}^2/S_{xx}} = 0.2668/\sqrt{30.73/2247.5} = 2.24$$

We have n – 2 = 28 degrees of freedom. From tables: $t_{0.025} = 2.048$ and $t_{0.005} = 2.763$.

We may not have sufficient evidence to reject H_o under 5% significance but there may be some evidence to do so under 1% significance. Note, we got R^2 = 15.7% which indicates only 15.7% of the variation is explained by the model meaning the linear relation is sufficiently weak.

[Total: 10]

Q17 (a) In this problem:

n = number of observations = 27
k = number of categories = 4

i. Estimate for overall serotonin level =
$\hat{\mu} = \bar{y}.. = (66 + 46 + 20 + 38)/(7 + 6 + 6 + 8) = 170/27 = 6.2962$

ii. Estimate for common underlying variance in serotonin level =
$\hat{\sigma}^2 = \frac{1}{n-k} SS_R$

Here: $SS_R = SS_T - SS_B$

$$SS_T = \sum\sum y_{ij}^2 - \frac{1}{n}y_{..}^2 = (676 + 382 + 78 + 196) - 170^2/27$$
$$= 261.630$$

$$SS_B = \sum \frac{1}{n_i}y_{i.}^2 - \frac{1}{n}y_{..}^2 = \left(\frac{66^2}{7} + \frac{46^2}{6} + \frac{20^2}{6} + \frac{38^2}{8}\right) - \frac{170^2}{27}$$
$$= 1222.12 - 1070.37$$
$$= 151.749$$
$$SS_R = 261.63 - 151.749 = 109.88$$

Hence $\hat{\sigma}^2 = \frac{109.88}{(27-4)} = 4.777$

(b) **ANOVA**

We are testing the hypotheses:
H_0: Mean Serotonin level is same across all groups v/s
$H_{1:}$ There are differences between the mean levels of serotonin

The ANOVA table is:

Source of Variation	Sum of Squares	Degree of freedom	Mean Square	F
Between the groups	151.749	3	50.583	10.589
Residual	109.881	23	4.777	
Total	261.63			

The variance ratio is 10.589, which has $F_{3.23}$ distribution.

5% critical value for $F_{3.23}$ is 3.028 << 10.589, hence we have sufficient evidence to reject H_0 at the 5% level. We conclude that there are differences between the mean levels of serotonin.

[Total: 7]

[Total Marks 100]

****