

Institute of Actuaries of India

Subject CT3 – Probability & Mathematical Statistics

May 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 :

As a die can show either an even or odd face when rolled, we have $P(A) = P(B) = 0.5$. Similarly, the sum of faces of two dice can be either of even or odd. Thus, $P(C) = 0.5$.

Pair-wise independence

$$P(A \cap B) = P(\text{Blue die shows even face} \cap \text{Red die shows even face})$$

Now there are only 4 combinations of (Blue die face, Red die face) possible such as: (Even, Even), (Even, Odd), (Odd, Even) & (Odd, Odd). Thus, (Even, Even) is just 1 of 4 possible outcomes.

$$\begin{aligned} \text{Hence: } P(A \cap B) &= 0.25 \\ &= (0.5) * (0.5) \\ &= P(A) \cdot P(B) \end{aligned}$$

Hence, A & B are independent.

[NB: If this argument is made using combinatorics like as below

$$P(A \cap B) = \frac{\binom{3}{1} \binom{3}{1}}{\binom{6}{1} \binom{6}{1}} = 0.25$$

This will be incorrect as it assumes independence of events A & B]

$$\begin{aligned} P(B \cap C) &= P(\text{Red die shows even face} \cap \text{Sum of two faces is even}) \\ &= P(\text{Red die shows even face} \cap \text{Blue die shows even face}) \text{ [as sum of two evens is even]} \\ &= 0.25 \\ &= P(B) \cdot P(C) \quad \text{[following the earlier argument]} \end{aligned}$$

Hence, B & C are independent.

Following similar argument, we can establish C & A are also independent.

Thus all three pairs are pair-wise independent.

Mutual independence

$$\begin{aligned} P(A \cap B \cap C) &= P(\text{Blue die shows even face} \cap \text{Red die shows even face} \cap \text{Sum of two faces is even}) \\ &= P(\text{Blue die shows even face} \cap \text{Red die shows even face}) \\ &\quad \text{[as if Blue and Red die are even then sum will be even]} \\ &= 0.25 \\ &\neq 0.125 = P(A) \cdot P(B) \cdot P(C) \end{aligned}$$

Thus, the events are **not** mutually independent.

[Total 4]

Solution 2 :

Set $x = 0$:

$$\mathbb{P}\left(X > \frac{1}{2}\right) = \mathbb{P}\left(X \leq \frac{1}{2}\right) \Rightarrow 1 - \mathbb{P}\left(X \leq \frac{1}{2}\right) = \mathbb{P}\left(X \leq \frac{1}{2}\right)$$

Thus:

$$\mathbb{P}\left(X \leq \frac{1}{2}\right) = \frac{1}{2} \Rightarrow \text{Median}(X) = \frac{1}{2}.$$

[Total 2]

Solution 3 :

$\{Z_0, Z_1, Z_2 \dots\}$ are independent random variables, each having a standard Normal distribution.

We know:

$$\sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

i. Take $n = 5$.

$$\sum_{i=1}^5 Z_i^2 \sim \chi_5^2$$

Thus by definition of the t-distribution and given that Z_0 and $\sum_{i=1}^5 Z_i^2$ are independent:

$$\frac{Z_0}{\sqrt{\sum_{i=1}^5 Z_i^2 / 5}} \sim t_5 \text{ distribution}$$

[An alternate formulation for this is as below:

$$\text{Set } \bar{Z} = \frac{1}{6} \sum_{i=0}^5 Z_i \quad \& \quad S^2 = \frac{1}{5} \sum_{i=0}^5 (Z_i - \bar{Z})^2$$

Using the facts:

★ $\bar{Z}$ and S^2 are independent

★ The random variable $\bar{Z} \sim N\left(0, \frac{1}{6}\right)$ distribution

★ The random variable $5S^2 \sim \chi_5^2$ distribution

we can have $\frac{\bar{Z}}{S/\sqrt{6}} \sim t_5$ distribution

]

- ii. Take blocks of 9 and 16 standard normal variates:

$$\sum_{i=1}^9 Z_i^2 \sim \chi_9^2 \quad \& \quad \sum_{i=10}^{25} Z_i^2 \sim \chi_{16}^2$$

As none of the subscripts of Z overlap, $\sum_{i=1}^9 Z_i^2$ & $\sum_{i=10}^{25} Z_i^2$ are independent.

Thus by definition of the F-distribution and using the independence property:

$$\frac{(\sum_{i=1}^9 Z_i^2 / 9)}{(\sum_{i=10}^{25} Z_i^2 / 16)} \sim F_{9,16} \text{ distribution}$$

[Total 4]

Solution 4 :

X has a gamma distribution with mean $\alpha\lambda$ and variance $\alpha\lambda^2$. This means that the parameters of X are α and $1/\lambda$.

The MGF of X is given as:

$$M_X(t) = (1 - t\lambda)^{-\alpha} \quad \text{for } t < \frac{1}{\lambda}$$

The cumulant generating function (CGF) is defined as: $C_X(t) := \log_e[M_X(t)]$

As $(-1 < -t\lambda \leq 0)$, we can use the series expansion formula for $\log_e$ function (from the tables):

$$\begin{aligned} C_X(t) &= \log_e[M_X(t)] \\ &= \log_e[(1 - t\lambda)^{-\alpha}] \\ &= -\alpha * \log_e[1 + (-t\lambda)] \\ &= -\alpha * \left[(-t\lambda) - \frac{(-t\lambda)^2}{2} + \frac{(-t\lambda)^3}{3} - \frac{(-t\lambda)^4}{4} + \dots \right] \\ &= (\alpha\lambda)t + \left(\frac{\alpha\lambda^2}{2}\right)t^2 + \left(\frac{\alpha\lambda^3}{3}\right)t^3 + \left(\frac{\alpha\lambda^4}{4}\right)t^4 + \dots \end{aligned}$$

Let the i^{th} cumulant of the distribution of X be denoted κ_i .

We know that κ_i is the coefficient of $\frac{t^i}{i!}$ for $i = 1, 2, 3, 4, \dots$

Hence:

$$\kappa_i = \left(\frac{\alpha \lambda^i}{i} \right) * i! = (i-1)! \alpha \lambda^i \text{ for } i = 1, 2, 3, 4, \dots$$

So:

$$\begin{aligned}\kappa_2 &= \alpha \lambda^2 \\ \kappa_3 &= 2\alpha \lambda^3 \\ \kappa_4 &= 6\alpha \lambda^4\end{aligned}$$

[Alternate Approach:

$$\kappa_i = C_X^{(i)}(0) = (i-1)! \alpha \lambda^i \text{ for } i = 1, 2, 3, 4, \dots$$

Here: $C_X^{(i)}(0)$ is the value of i^{th} derivative of $C_X(t)$ w.r.t. t calculated at $t = 0$.]

[Total 4]

Solution 5 :

Observe that for a randomly selected person:

- $E[X + Y] = E[X] + E[Y] = 50 + 20 = 70$
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \text{Cov}[X, Y] = 50 + 30 + 2(10) = 100$

Using **Central Limit Theorem**, T is *approximately* Normal with

- Mean: $E[T] = 100(70) = 7000$
- Variance: $\text{Var}[T] = 100(100) = 100^2$

Therefore:

$$\mathbb{P}(T < 7050) = \mathbb{P}\left[\frac{T - 7000}{100} < \frac{7050 - 7000}{100}\right] \approx \mathbb{P}[Z < 0.5] = 0.69146$$

Here: Z is a standard normal variable

[Total 3]

Solution 6 :

i. We know:

Area under each bar of a class-interval is **proportional** to the Frequency

Note:

- For 1st class-interval: Area = $(60.5 - 59.5) * 6 = 6$ & Frequency = 12
- For 3rd class-interval: Area = $(65.5 - 61.5) * 2 = 8$ & Frequency = 16

Similarly if one checks the remaining intervals for which we know the frequency, we see that that the frequency is 2 times the area under the bar.

This means that the frequency of the interval (75.5 – 78.5) will be 2 times the area of the bar. In other words: Frequency = $2 * [(78.5 - 75.5) * 1.5] = 9$.

For the last interval, Frequency = $2 * [(90.5 - 78.5) * 0.5] = 12$ using similar arguments.

[Alternately, the Frequency of last interval can be obtained as $140 - \Sigma(freq) = 140 - 128 = 12$ where Σ is over the first 7 class intervals.]

- ii. Expand the frequency distribution table and compute cumulative frequency (CumFreq):

t	Frequency	CumFreq	Proportion	
59.5 - 60.5	12	12	0.09	Q ₁ : 25 th Percentile
60.5 - 61.5	14	26	0.19	
61.5 - 65.5	16	42	0.30	Q ₂ : 50 th Percentile
65.5 - 67.5	24	66	0.47	
67.5 - 70.5	33	99	0.71	Q ₃ : 75 th Percentile
70.5 - 75.5	20	119	0.85	
75.5 - 78.5	9	128	0.91	
78.5 - 90.5	12	140	1.00	
	140			

Using the **Proportion** column we can conclude that Q₁ lies in interval (61.5 – 65.5), Q₂ lies in the interval (67.5 – 70.5) and Q₃ lies in interval (70.5 – 75.5) using the definitions of the quartiles.

As this is a grouped data, the i^{th} quartile (Q_i) will correspond to the $N * (i/4)^{\text{th}}$ observation for $i = 1, 2 \text{ \& } 3$ and $N = 140$.

Assuming the values are distributed uniformly and applying linear interpolation:

$$Q_1: 61.5 + (65.5 - 61.5) * \left[\frac{140 * \left(\frac{1}{4}\right) - 26}{42 - 26} \right] = 63.75$$

$$Q_2: 67.5 + (70.5 - 67.5) * \left[\frac{140 * \left(\frac{1}{2}\right) - 66}{99 - 66} \right] = 67.86$$

$$Q_3: 70.5 + (75.5 - 70.5) * \left[\frac{140 * \left(\frac{3}{4}\right) - 99}{119 - 99} \right] = 72.00$$

- iii. Using the given definition of skewness:

$$\begin{aligned} skew &= \frac{(Q_1 - 2Q_2 + Q_3)}{IQR} \\ &= \frac{(Q_1 - 2Q_2 + Q_3)}{Q_3 - Q_1} \\ &= \frac{63.75 - 2 * 67.86 + 72.00}{72.00 - 63.75} = 0.003 \end{aligned}$$

This indicates that there is (slight) positive skewness in the data.

[Total 10]

Solution 7 :

A random variable X has the probability density function

$$f(x) = \begin{cases} c \exp\left(-\frac{x^2}{2}\right), & x < 0 \\ \frac{c}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right), & x \geq 0 \end{cases}$$

Here: $\sigma > 0$ is an unknown parameter and c is a given constant.

- i. For f(x) to be valid density, we must have:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= 1 \\ \text{or, } \int_{-\infty}^0 f(x) dx + \int_0^{\infty} f(x) dx &= 1 \\ \text{or, } c * \left[\int_{-\infty}^0 \exp\left(-\frac{x^2}{2}\right) dx + \int_0^{\infty} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \right] &= 1 \\ \text{or, } c &= \left[\int_{-\infty}^0 \exp\left(-\frac{x^2}{2}\right) dx + \int_0^{\infty} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \right]^{-1} \end{aligned}$$

We know: Median $[N(0, 1)] = 0$. This means:

$$\int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{2} \Rightarrow \int_{-\infty}^0 \exp\left(-\frac{x^2}{2}\right) dx = \frac{\sqrt{2\pi}}{2}$$

Similarly: Median $[N(0, \sigma^2)] = 0$. This means:

$$\int_0^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \frac{1}{2} \Rightarrow \int_0^{\infty} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \frac{\sqrt{2\pi}}{2}$$

Thus:

$$\begin{aligned} c &= \left[\int_{-\infty}^0 \exp\left(-\frac{x^2}{2}\right) dx + \int_0^{\infty} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \right]^{-1} \\ &= \left[\frac{\sqrt{2\pi}}{2} + \frac{\sqrt{2\pi}}{2} \right]^{-1} = \frac{1}{\sqrt{2\pi}} \end{aligned}$$

[Alternately:

As the value of 'c' does not depend on choice of σ , we can derive 'c' by setting the value of $\sigma = 1$.
In that case:

$$\begin{aligned} c &= \left[\int_{-\infty}^0 \exp\left(-\frac{x^2}{2}\right) dx + \int_0^{\infty} \exp\left(-\frac{x^2}{2}\right) dx \right]^{-1} \\ &= \left[\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx \right]^{-1} \\ &= \frac{1}{\sqrt{2\pi}} * \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \right]^{-1} \\ &= \frac{1}{\sqrt{2\pi}} \text{ given that the integral equals to 1 as it is total probability of a } N(0,1) \text{ distribution} \end{aligned}$$

]

- ii. If $\sigma = 1$, then the given distribution is the standard Normal (i.e. Normal distribution with mean 0 and variance 1).
- iii. The likelihood equation for the given data will be:

$$L(\sigma; \underline{X}) = \left(\frac{1}{\sqrt{2\pi}} \right)^5 \exp \left[-\frac{(-0.42)^2 + (-1.47)^2 + (-0.95)^2 + (-0.39)^2 + (-0.51)^2}{2} \right]$$

$$* \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^5 \exp \left[-\frac{(2.80)^2 + (0.31)^2 + (0.20)^2 + (1.15)^2 + (2.18)^2}{2\sigma^2} \right]$$

Taking logarithm,

$$\begin{aligned} l(\sigma; \underline{X}) &= \text{constant} - 5 * \log_e(\sigma) - \left[\frac{(2.80)^2 + (0.31)^2 + (0.20)^2 + (1.15)^2 + (2.18)^2}{2\sigma^2} \right] \\ &= \text{constant} - 5 * \log_e(\sigma) - \frac{14.051}{2\sigma^2} \end{aligned}$$

Differentiating w.r.t. σ we get:

$$\frac{\partial l}{\partial \sigma} = -\frac{5}{\sigma} + \frac{14.051}{\sigma^3}$$

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

$$\hat{\sigma}_{MLE}^2 = \frac{1}{5} * 14.051 = 2.8102$$

$$\Rightarrow \hat{\sigma}_{MLE} = \sqrt{2.8102} = 1.6764$$

$$\left[\text{Check: } \frac{\partial^2 l}{\partial \sigma^2} = \frac{5}{\sigma^2} - \frac{3 * 14.051}{\sigma^4} \right]_{\hat{\sigma}_{MLE} = 1.6764} < 0 \Rightarrow \text{maximum}$$

[Total 10]

Solution 8 :

i. The pivotal quantity of the form $g(\underline{X}, \theta)$ should possess the following properties:

- it is a function of the sample values and the unknown parameter θ
- its distribution is completely known
- it is monotonic in θ .

ii.

(a) Using the given values of the sample (in units of ₹ '000):

- Sample mean: $\bar{x} = 645.20$

- Sample standard deviation: $s = \sqrt{\frac{1}{4} \sum (x_i - \bar{x})^2} = \sqrt{25.70} = 5.0695$

As the sample come from a Normal distribution $N(\mu, \sigma^2)$,

$$\frac{\bar{X} - \mu}{S/\sqrt{5}} \sim t_4 \text{ distribution}$$

From the statistical tables, we have:

$$0.95 = \mathbb{P}[-2.776 < t_4 < 2.776]$$

$$\text{or, } 0.95 = \mathbb{P}\left[-2.776 < \frac{\bar{X} - \mu}{S/\sqrt{5}} < 2.776\right]$$

$$\text{or, } 0.95 = \mathbb{P}[\bar{X} - 2.776 * S/\sqrt{5} < \mu < \bar{X} + 2.776 * S/\sqrt{5}]$$

$$\text{or, } 0.95 = \mathbb{P}[645.20 - 2.776 * 5.0695/\sqrt{5} < \mu < 645.20 + 2.776 * 5.0695/\sqrt{5}]$$

$$\text{or, } 0.95 = \mathbb{P}[638.9 < \mu < 651.5]$$

So, a 95% confidence interval for the average salary μ is (638.9, 651.5).

(b) As the sample come from a Normal distribution $N(\mu, \sigma^2)$,

$$\frac{4 S^2}{\sigma^2} \sim \chi_4^2 \text{ distribution}$$

From the statistical tables, we have:

$$0.95 = \mathbb{P}[0.7107 < \chi_4^2]$$

$$\text{or, } 0.95 = \mathbb{P}\left[0.7107 < \frac{4 S^2}{\sigma^2}\right]$$

$$\text{or, } 0.95 = \mathbb{P}\left[\sigma^2 < \frac{4 S^2}{0.7107}\right]$$

$$\text{or, } 0.95 = \mathbb{P}\left[\sigma < \frac{2 (5.0695)}{\sqrt{0.7107}}\right]$$

$$\text{or, } 0.95 = \mathbb{P}[\sigma < 12.0]$$

So, a 95% confidence interval of the form " $\sigma < L$ " is (0, 12.0).

[Total 11]

Solution 9 :**i. Male Bowlers: Paired-Data**

$$n_1 = 10, \bar{x}_1 = 24.9, s_1 = (213.656)^{0.5} = 14.617$$

A 95% confidence interval for male group (assuming normality) can be calculated:

$$\begin{aligned} & \bar{x}_1 \pm t_9 (0.025) \frac{s_1}{\sqrt{n_1}} \\ &= 24.9 \pm 2.262 * \frac{14.617}{\sqrt{10}} \\ &= (14.444, 35.356) \end{aligned}$$

ii. Female Bowlers: Paired-Data

$$n_2 = 10, \bar{x}_2 = 20.7, s_2 = (326.456)^{0.5} = 18.068$$

A 95% confidence interval for female group (assuming normality) can be calculated:

$$\begin{aligned} & \bar{x}_2 \pm t_9 (0.025) \frac{s_2}{\sqrt{n_2}} \\ &= 20.7 \pm 2.262 * \frac{18.068}{\sqrt{10}} \\ &= (7.775, 33.625) \end{aligned}$$

iii. None of the intervals include zero, and therefore there is sufficient evidence (at 5% significance) that the special diet has an effect on the bowling speed, i.e. it increases the bowling speed for both males and female bowlers.

iv. Testing Common Variance

Assuming that the impacts of diet data come from normal distribution: $N(\mu_1, \sigma_1^2)$ & $N(\mu_2, \sigma_2^2)$:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{9,9} \text{ distribution}$$

To test: $H_0: \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 \neq \sigma_2^2$

The test statistic value under H_0 is $\frac{S_1^2}{S_2^2} = 0.654$

The $F_{9,9}$ distribution has lower and upper 2.5% critical points at 0.248 and 4.026. Our observed value (0.654) is well within the range between the critical points. Therefore there is no evidence (at 5% level) to suggest that the variances differ of the impact in the male and female samples.

v. Two-Sample t-test

Using the inference from part (iv), we have

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ distribution}$$

$$\text{Here: the pooled variance: } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 270.056 = (16.433)^2$$

To test: $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$

The test statistic value under H_0 is

$$\frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{24.9 - 20.7}{16.433 \sqrt{\frac{1}{10} + \frac{1}{10}}} = 0.571$$

$$t_{18}(0.025) = -2.101 \text{ \& } t_{18}(0.975) = 2.101$$

There is no evidence to reject the null hypothesis at the 5% level that the mean impact due to the specialised diet is same for male and female bowlers.

[Total 16]

Solution 10 :

i. The relevant summary statistics to compute correlation coefficient are:

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 207 - 10 * \left(\frac{39}{10}\right)^2 = 54.90$$

$$S_{xy} = \sum xy - n\bar{x}\bar{y} = 2853 - 10 * \left(\frac{39}{10}\right)\left(\frac{562}{10}\right) = 661.20$$

$$S_{yy} = \sum y^2 - n\bar{y}^2 = 40508 - 10 * \left(\frac{562}{10}\right)^2 = 8923.60$$

$$\text{Correlation Coefficient } r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{661.20}{\sqrt{54.90}\sqrt{8923.60}} = 0.945$$

ii. Fitted Linear Regression Equation

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.20}{54.90} = 12.04$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{562}{10}\right) - 12.04 * \left(\frac{39}{10}\right) = 9.23$$

Therefore, the fitted regression line is: $y = \hat{\alpha} + \hat{\beta}x = 9.23 + 12.04x$

iii. Relation: $SS_{TOT} = SS_{REG} + SS_{RES}$

$$SS_{TOT} = S_{yy} = 8923.60$$

$$SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 8923.60 - \frac{(661.20)^2}{54.90} = 960.30$$

$$SS_{REG} = S_{TOT} - S_{RES} = 8923.60 - 960.30 = 7963.30$$

iv. Coefficient of Determination:

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{SS_{REG}}{SS_{TOT}} = \frac{7963.30}{8923.60} = 0.8924$$

For the simple linear regression model, the value of the coefficient of determination is the square of the correlation coefficient for the data, since,

$$0.945 = r = \frac{S_{xy}}{(S_{xx} * S_{yy})^{0.5}} = \sqrt{R^2} = \sqrt{0.8924}$$

[Total 10]

Solution 11 :

i. We are carrying out the following test:

H_0 : No significant difference between mean fees being charged by each parlour
v/s

H_1 : Significant difference between mean fees being charged by at least two parlours

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

$$SS_T = 2,647,876 - \frac{(8,854)^2}{30} = 34,765.47$$

$$SS_B = \frac{1}{5} (1,494^2 + 1,318^2 + 1,654^2 + 1,480^2 + 1,418^2 + 1,490^2) - \frac{8,854^2}{30} = 12,109.47$$

$$SS_R = SS_T - SS_B = 22,656.00$$

Source	df	SS	MS	F
Treatments	5	12,109.47	2,421.89	2.57
Residual	24	22,656.00	944.00	
Total	29	34,765.47		

From tables $F_{5,24}(5\%) = 2.621$

And observed $F < 2.621$

Therefore there are no significant differences, at the 5% level, between mean fees being charged by each parlour.

ii. We have

$$\frac{SS_R}{\sigma^2} \sim \chi_{30-6}^2 = \chi_{24}^2 \text{ distribution}$$

From the statistical tables, we have:

$$0.95 = \mathbb{P}[12.40 < \chi_{24}^2 < 39.36]$$

$$\text{or, } 0.95 = \mathbb{P}\left[12.40 < \frac{SS_R}{\sigma^2} < 39.36\right]$$

$$\text{or, } 0.95 = \mathbb{P}\left[\frac{SS_R}{39.36} < \sigma^2 < \frac{SS_R}{12.40}\right]$$

$$\text{or, } 0.95 = \mathbb{P}\left[\sqrt{\frac{22656}{39.36}} < \sigma < \sqrt{\frac{22656}{12.40}}\right]$$

$$\text{or, } 0.95 = \mathbb{P}[23.99 < \sigma < 42.74]$$

So, a 95% confidence interval for σ is (23.99, 42.74).

[Total 8]

Solution 12 :

P is a random variable having a beta distribution with parameters $\alpha (> 0)$ and $\beta (> 0)$ defined over the region $(0, 1)$.

- i. The probability density function of P is defined as (*same available in the Tables*):

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

The support for this density function is $0 < p < 1$.

- ii. It is expected that we derive the k^{th} raw moment from first principles.

For $k = 0$, the given equation is an identity.

For any other $k > 0$, we have:

$$\begin{aligned} \mathbb{E}[P^k] &= \int_0^1 p^k f(p) dp \\ &= \int_0^1 p^k \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} dp \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + k)}{\Gamma(\alpha + \beta + k)} \int_0^1 \frac{\Gamma(\alpha + k + \beta)}{\Gamma(\alpha + k) \Gamma(\beta)} p^{\alpha+k-1} (1-p)^{\beta-1} dp \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + k)}{\Gamma(\alpha + \beta + k)} \end{aligned}$$

[The integral equals to 1 as it is the total probability of a Beta($\alpha+k, \beta$) random variable]

Put $k = 1$:

$$\mathbb{E}[P] = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + \beta + 1)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\alpha \Gamma(\alpha)}{(\alpha + \beta) \Gamma(\alpha + \beta)} = \frac{\alpha}{\alpha + \beta}$$

NB: It is incorrect to set $\Gamma(\alpha+1) = \alpha!$. This is because ' α ' is not necessarily an integer.

Put $k = 2$:

$$\begin{aligned} \mathbb{E}[P^2] &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha + \beta + 2)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{(\alpha + 1) \alpha \Gamma(\alpha)}{(\alpha + \beta + 1) (\alpha + \beta) \Gamma(\alpha + \beta)} \\ &= \frac{\alpha}{\alpha + \beta} \frac{\alpha + 1}{\alpha + \beta + 1} \end{aligned}$$

Thus:

$$\begin{aligned}\text{Var}[P] &= \mathbb{E}[P^2] - (\mathbb{E}[P])^2 = \frac{\alpha}{\alpha + \beta} \frac{\alpha + 1}{\alpha + \beta + 1} - \left(\frac{\alpha}{\alpha + \beta}\right)^2 \\ &= \frac{\alpha [(\alpha + 1)(\alpha + \beta) - \alpha(\alpha + \beta + 1)]}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}\end{aligned}$$

iii. $\mathbb{E}[P(1 - P)]$

$$\begin{aligned}&= \int_0^1 p(1 - p) \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1 - p)^{\beta-1} dp \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + 1)\Gamma(\beta + 1)}{\Gamma(\alpha + \beta + 2)} \int_0^1 \frac{\Gamma(\alpha + \beta + 2)}{\Gamma(\alpha + 1)\Gamma(\beta + 1)} p^{\alpha+1-1} (1 - p)^{\beta+1-1} dp\end{aligned}$$

[The integral equals to 1 as it is the total probability of a Beta($\alpha+1, \beta+1$) random variable]

$$\begin{aligned}&= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + 1)\Gamma(\beta + 1)}{\Gamma(\alpha + \beta + 2)} \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\alpha\Gamma(\alpha)\beta\Gamma(\beta)}{(\alpha + \beta + 1)(\alpha + \beta)\Gamma(\alpha + \beta)} \\ &= \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)}\end{aligned}$$

[Alternately: $\mathbb{E}[P(1-P)] = \mathbb{E}[P] - \mathbb{E}[P^2]$ and plug in values obtained in part (ii)]

iv. For $i = 1, 2 \dots n$, we have:

- X_i is the random variable which takes the value of 1 if the trial is successful for the i^{th} patient and 0 otherwise;
- P_i denotes the probability that the drug trial will be successful for the i^{th} patient. P_i follows a Beta distribution with parameters $\alpha (> 0)$ and $\beta (> 0)$.

Thus: $X_i \mid P_i \sim \text{Bernoulli}(P_i)$

$S = \sum_{i=1}^n X_i$ is the total number of successful trials among the n patients.

$$(a) \quad \mathbb{E}[S] = \sum_{i=1}^n \mathbb{E}[X_i]$$

$$\begin{aligned}
&= \sum_{i=1}^n \mathbb{E}[\mathbb{E}\{X_i|P_i\}] \quad \text{using conditional expectation theory} \\
&= \sum_{i=1}^n \mathbb{E}[P_i] \quad \text{as } X_i|P_i \sim \text{Bernoulli}(P_i) \Rightarrow \mathbb{E}[X_i|P_i] = P_i \\
&= \frac{n\alpha}{\alpha + \beta} \quad \text{as } P_i \sim \text{Beta}(\alpha, \beta) \Rightarrow \mathbb{E}[P_i] = \frac{\alpha}{\alpha + \beta} \text{ from part (ii)}
\end{aligned}$$

$$\begin{aligned}
(b) \quad \text{Var}[S] &= \text{Var}\left[\sum_{i=1}^n X_i\right] \\
&= \sum_{i=1}^n \text{Var}[X_i] \quad \text{as } X_i \text{ s are independent} \\
&= \sum_{i=1}^n [\text{Var}[\mathbb{E}\{X_i|P_i\}] + \mathbb{E}[\text{Var}\{X_i|P_i\}]] \quad \text{using conditional expectation theory} \\
&= \sum_{i=1}^n [\text{Var}[P_i] + \mathbb{E}[P_i(1 - P_i)]] \quad \because X_i|P_i \sim \text{Bernoulli}(P_i) \Rightarrow \text{Var}[X_i|P_i] = P_i(1 - P_i) \\
&= \sum_{i=1}^n \left[\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \right] \quad \text{using results in (ii) \& (iii)} \\
&= \frac{n\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} [1 + (\alpha + \beta)] \\
&= \frac{n\alpha\beta}{(\alpha + \beta)^2}
\end{aligned}$$

[Total 18]
