Institute of Actuaries of India

Subject CT3 – Probability & Mathematical Statistics

September 2017 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 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:

i) Median = $\left(\frac{1}{2}n + \frac{1}{2}\right)^{th}$ $value$

$Q_1 = \left(\frac{1}{4}n + \frac{1}{2}\right)^{th}$ value and $Q_3 = \left(\frac{3}{4}n + \frac{1}{2}\right)^{th}$ value .

For city *A*:

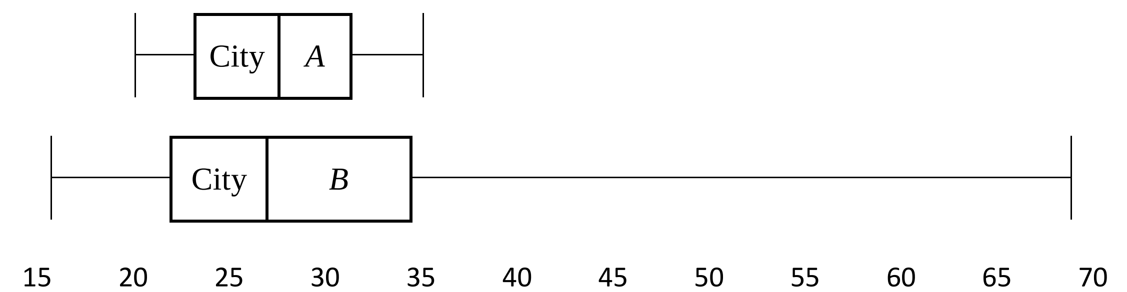
Median = 27.50

$Q_1 = 23.50$; $Q_3 = 31.50$

For city *B*:

Median = 27.00

$Q_1 = 22.50$; $Q_3 = 34.50$



For quartiles, alternative method are
(Alternate method 1)
$Q_1 = \left(\frac{1}{4}n + \frac{1}{4}\right)^{th}$ *value and* $Q_3 = \left(\frac{3}{4}n + \frac{3}{4}\right)^{th}$ *value*

For city A: $Q_1 = 22.25$ $Q_3 = 31.75$
For city B: $Q_1 = 21.25$ $Q_3 = 35.75$

(Alternate method 2)

$Q_1 = \left(\frac{1}{4}n + \frac{1}{4}\right)^{th}$ *value and* $Q_3 = \left(\frac{3}{4}n + \frac{3}{4}\right)^{th}$ *value*

Q1= 3.25th item an Q3=9.75th item
For city A:
Q1 = 4th item – 0.25 (4th item- 3rd item) = 26-0.25*(26-21)=24.75*
Q3 = 10th item – 0.75 (10th item- 9th item) = 32-0.75*(32-31)=31.25*

For city B:
Q1 = 4th item – 0.25 (4th item- 3rd item) = 25-0.25*(25-20)=23.75*
Q3 = 10th item – 0.75 (10th item- 9th item) = 37-0.75*(37-32) =33.25*

(The above answers are in line with answers using quartile functions used in excel spreadsheet)

Either of the methods is acceptable as long as students are consistent.

[4]

ii) Looking at the box plots, we see that the median of both distributions are close to 27° Celsius. This suggests that the monthly maximum temperatures for City *A* and *B* may have the averages close to each other.

However, the overall spread of the figures for city *B* appears to be greater than the corresponding spread for city *A* which can be confirmed by measuring by IQR. This suggests that the variability in the monthly maximum temperatures for City *B* is greater than the corresponding variability for city *A* (although conclusions drawn from such small sample sizes should be treated with caution).

The value of 68 for city *B* could be an outlier.

City *A* distribution seems symmetric and City *B* distribution is clearly positively skewed. For comparison:

- for city *A*: the modes (26.00) ≈ median (27.50) = mean (27.50)
- for city *B*: the mode (25.00) < median (27.00) < mean (30.75) [3]

[7 Marks]

Solution 2:

i) In order that $f(x)$to be a pdf :

$$f(x) = \frac{k}{(4+x)^2} \geq 0 \;; \quad x > 0 \quad \text{and} \quad \int_0^\infty f(x)\, dx = 1$$

We can see that $f(x) = \frac{k}{(4+x)^2} \geq 0$, provided k≥ 0

$$\int_0^\infty \frac{k}{(4+x)^2}\, \mathrm{dx} = \left[-\frac{k}{(4+x)}\right]_0^\infty = \frac{k}{4} - \frac{k}{(4+\infty)} = \frac{k}{4}$$

Therefore, $k = 4$ (hence it is satisfying the condition k≥ 0) [1]

ii) To simulate a random variable we need the distribution function $F(x)$:

$$F(x) = \int_0^{x} \frac{4}{(4+t)^2}\, dt = \left[-\frac{4}{(4+t)}\right]_0^{x} = \frac{4}{4} - \frac{4}{(4+x)} = 1 - \frac{4}{(4+x)}$$

We can now use the inverse transform method:

$u = 1 - \frac{4}{(4+x)} => x = \frac{4}{(1-u)} - 4 = \frac{4u}{(1-u)}$

Substituting in our values of u, we obtain:

$x_1 = \frac{4\,(0.914)}{(1-0.914)} = 42.512$

$x_2 = \frac{4\,(0.683)}{(1-0.683)} = 8.618$

$x_3 = \frac{4\,(0.257)}{(1-0.257)} = 1.384$ [4]

[5 Marks]

Solution 3:

i) The random variable X is Poisson with mean 1.

Hence, $P(X$ takes ositive even values $)$ = P(X=2) + P(X=4) +…

= $e^{-1} \times(\frac{1}{2!} + \frac{1}{4!} + \ldots)$

We know that $e^1 = 1+1+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+\ldots$

and

$e^{-1} = 1-1+\frac{1}{2!}-\frac{1}{3!}+\frac{1}{4!}+\ldots$

Hence, $e^1 + e^{-1} = 2+2\left(\frac{1}{2!}+\frac{1}{4!}+\frac{1}{6!}+\ldots\right)$

$= 2 * \left(1+\frac{1}{2!}+\frac{1}{4!}+\frac{1}{6!}+\ldots\right)$

$(e^1 + e^{-1})/2 - 1 = \left(\frac{1}{2!}+\frac{1}{4!}+\frac{1}{6!}+\ldots\right)$

$P(X$ takes positive even values $)$ = $e^{-1}\ [(e^1 + e^{-1})/2 - 1\ ]$

$=(1+ e^{-2})/2 - e^{-1}$ [4]

ii) We know that

$P(X$ takes positive even values $) + P(X$ takes postive odd values $)$

$+P(X$ takes zero $) = 1$

Hence, $P(X$ takes positive odd values $)$ = $1 -[(1+ e^{-2})/2 - e^{-1}]- e^{-1}$

$=(1- e^{-2})/2$ [2]

[6 Marks]

Solution 4:

The χ_9^2 distribution has mean 9 and variance 18.

Setting the expressions for the mean and variance of the lognormal distribution to the above values and solving the two equations:

$$e^{\mu + \frac{1}{2}\sigma^2} = 9 \; ; \qquad e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right) = 18$$

Squaring the first equation and substituting into the second, we get:

$$81\left(e^{\sigma^2} - 1\right) = 18$$

Solving this, we get $\sigma^2 = \log\frac{99}{81} = 0.2007$

Now substituting σ^2in any one equation, we get μ = 2.0969.

If X ~ log Normal then log X ~ Normal, so: $P\ (X > 9) = P\ (\log X > \log 9)$

$$= \mathrm{P}\ \left(Z > \frac{\log 9 - 2.0969}{\sqrt{0.2007}}\right) = \mathrm{P}\ (Z > \ 0.224)$$

Using interpolation from tabulated values for 0.22 and 0.23 in page 160

$$= 1 - P\ (Z < 0.224) = 1 - 0.5886 = 0.4114$$ **[5 Marks]**

Solution 5:

Let X be the time for the first exotic cake and Y be the time for the second. Then:

X ~ N (120, 225) and Y ~ N (120, 225) [in terms of minutes]

We require: $P\ (|X - Y| < 25)$

The distribution of X - Y, is:

$$(X - Y) \sim N\ (120 - 120,\ 225 + 225) \sim N\ (0,\ 450)$$

$$P\ (|X - Y| < 25) = \ P\ (-25 < X - Y < 25)$$

$$= \ P\ \left(\frac{-25}{\sqrt{450}} < Z < \frac{25}{\sqrt{450}}\right)$$

$$= \ P\ (Z < 1.1785)\ -\ P\ (Z < -1.1785)$$
$$= \ P\ (Z < 1.1785)\ -\ (1 - \ P\ (Z < 1.1785)\ )$$

$$= 2P(Z < 1.1785) - 1$$

$$= (2)(0.8807) - 1 = 0.7614$$

[4 Marks]

Solution 6:

We know that $\bar{X}$ and s^2 are independent if we are sampling from a normal distribution:

$$P[\bar{X} > 6.75 \text{ and } s < 3.75] = P[\bar{X} > 6.75] \, P[s < 3.75]$$

Now, $\bar{X} \sim N\left(8, \frac{3^2}{9}\right) = N(8, 1)$, so

$$P[\bar{X} > 6.75] = P\left(Z > \frac{6.75-8}{1}\right) = P(Z > -1.25) = 0.89435$$

And, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, so

$$P[s < 3.75] = P\left[\frac{8\,s^2}{3^2} < \frac{8\,(3.75)^2}{3^2}\right] = P\left[\chi^2_8 < 12.5\right] = 0.8697$$

Hence, $P[\bar{X} > 6.75 \text{ and } s < 3.75] = (0.89435)(0.8697) = 0.777816$

[4 Marks]

Solution 7:

Type I Error:

P(Reject H_0 when H_0 is True)

P($\bar{X}$ > 110 when μ = 100) <0.01

We have $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$,Hence Z= $\frac{(\bar{X}-\mu)}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

P($\bar{X}$ > 110 when μ = 100) = P(Z > $\frac{(110-100)}{\frac{20}{\sqrt{n}}}$) =P(Z >$\frac{\sqrt{n}}{2}$)

Setting this probability equal to 0.01 , we get $\frac{\sqrt{n}}{2}$=2.3263

Hence *n*=21.65.So the required minimum sample size is 22.

[4 Marks]

Solution 8:

i) The mode of X is the value of x at which $f(x)$is maximum.

For ease of differentiation let us consider *y = log f (x)* for the gamma distribution

$y = log\, f(x) = K + (\alpha - 1)\, log\, x - \lambda x$; (*K* is a constant)

$$\frac{dy}{dx} = 0 + \frac{(\alpha - 1)}{x} - \lambda$$

Setting the above equal to zero gives:

$$x = \frac{(\alpha - 1)}{\lambda}$$

To check for maximum, differentiating second time:

$$\frac{d^2 y}{d x^2} = -\frac{(\alpha - 1)}{x^2} - 0 = -\frac{(\alpha - 1)}{x^2}$$

Substituting $x = \frac{(\alpha - 1)}{\lambda}$ in above, we get: $\frac{d^2 y}{d x^2} = -\frac{\lambda^2}{(\alpha - 1)} < 0$

Hence, the mode is: $\frac{(\alpha - 1)}{\lambda}$, $\alpha > 1$ [4]

ii) Given that $X \sim Gamma\ (40, 0.1)$

Using the relationship $X \sim Gamma\ (\alpha, \lambda) \Rightarrow 2\lambda X \sim \chi^2_{2\alpha}$

$P[X > 500] = P[2\lambda X > 2(0.1)(500)] = P[\chi^2_{80} > 100]$

Using the χ^2 tables on page 169, and interpolating, gives a value of 0.0679

[2]

iii) The mean and variance of the gamma distribution are:

$$E[X] = \frac{\alpha}{\lambda} = \frac{40}{0.1} = 400 \text{ and } Var[X] = \frac{\alpha}{\lambda^2} = \frac{40}{(0.1)^2} = 4000$$

Because the gamma distribution can be represented as the sum *iid e*xponential random variables, in view of the Central Limit Theorem that, for large α, the gamma distribution can be approximated by a normal distribution.

Using the normal approximation to the gamma gives:
$X \sim Gamma\ (40, 0.1) \approx N\ (400, 4000)$

Hence:
$P[X > 500] = P\left(Z > \frac{500-400}{63.25}\right) = P(Z > 1.581)$

Using interpolation from tabulated values for 1.58 and 1.59 in page 160
$P(Z > 1.581) = 1 - P(Z < 1.581) = 1 - 0.9431 = 0.0569$ [3]

[9 Marks]

<u>Solution 9:</u>

The mean and variance of the Pareto distribution are given by:

$$E[X] = \frac{\lambda}{\alpha - 1} = \frac{4}{3 - 1} = 2$$

$$Var[X] = \frac{\alpha\lambda^2}{(\alpha - 1)^2(\alpha - 2)} = \frac{3(4)^2}{(3 - 1)^2(3 - 2)} = 12$$

From page 16 of the *Tables* $Var(Y) = Var[E(Y \mid X)] + E[Var(Y \mid X)]$:

$$Var(Y) = Var[2X + 5] + E[X^2 + 3]$$

$Var\,(Y) = 4\,Var\,[X] + E\,[X^2] + 3$

We know that $E\,[X^2] = Var[X] + (E\,[X])^2 = 12 + 2^2 = 16$
$Var\,(Y) = 4\,(12) + 16 + 3 = 67$

Hence, the standard deviations is $\sqrt{67} = 8.1854$

[5 Marks]

Solution 10:

Let $X_1, X_2, \ldots X_n$ be a random sample from a population with *pdf:*

$$f(x/\theta) = \theta x^{\theta-1};\ 0 < x < 1, \theta > 0.$$

This is seen as Beta (α, β) distribution with $\alpha = \theta$ and $\beta = 1$. [Actuarial tables P14]

i) $E[X] = \frac{\theta}{\theta+1}$

Alternatively, by simple integration

$E(X) = \int_0^1 x\ \Theta\, x^{\Theta-1}\, dx = \int_0^1 \Theta\, x^{\Theta}\, dx = \Theta \left[\frac{x^{\Theta+1}}{\Theta+1}\right]_0^1 = \frac{\Theta}{\Theta+1}$

The Method of moments estimator is obtained by equating population mean to the sample mean.

This gives, $\overline{X} = \frac{\theta}{\theta+1}$. Hence the method of moments estimator is $\widetilde{\theta} = \frac{\bar{X}}{1-\bar{X}}$.

[2]

ii) The Likelihood :

$L(\theta/\underline{x}) = \prod f(x_i) = \prod \theta x_i^{\theta-1} = \theta^n \prod x_i^{\theta-1}$

$\log L(\theta/\underline{x}) = n \log\theta + (\theta-1) \sum log\ x_i$

$\frac{\delta \log L(\theta/\underline{x})}{\delta\theta} = \frac{n}{\theta} + \sum log\ x_i$

$\frac{\delta \log L(\theta/\underline{x})}{\delta\theta} = 0$ implies that $\theta = -\frac{n}{\sum log\ x_i}$

$\frac{\delta^2 \log L(\theta/\underline{x})}{\delta\theta^2} = -\frac{n}{\theta^2} < 0$

Hence, MLE is $\hat{\theta} = -\frac{n}{\sum log\ x_i}$ [4]

iii) The Cramer Rao lower bound for the varianceof an unbiased estimator is

$$CRLB = -\frac{1}{E\left[\frac{\delta^2 \log L(\theta/\underline{x})}{\delta\theta^2}\right]}$$

$$= \frac{1}{\frac{n}{\theta^2}} = \frac{\theta^2}{n}$$ [2]

iv) a) For the given data: $\overline{x}$= 0.304 Hence, Moment estimate is $\frac{\overline{x}}{1-\overline{x}} = \frac{0.304}{1-0.304}$ =0.43678

[1]

b) To calculate M L estimate:

log (0.32)= -1.13943
log (0.19)= -1.66073
log (0.25)= -1.38629
log (0.34)= -1.07881
log (0.45)= -0.8675
Total = -6.13277

Hence $\hat{\theta} = -\frac{5}{-6.13277} = 0.815292$ [2]

c) A maximum likelihood estimate of the mean of the population is :

$$\frac{\hat{\theta}}{1+\hat{\theta}} = \frac{0.815292}{1.8152926} = 0.449125$$ [1]

[12 marks]

<u>Solution 11:</u>

The *p d f* of exponential distribution is given by

$$f(x) = \lambda e^{-\lambda x} ; x > 0 , \lambda > 0$$

Using method of moments: Equating sample mean and population mean → $\bar{x} = \frac{1}{\lambda}$

Hence $\hat{\lambda} = \frac{1}{\bar{x}}$

Mean $\overline{x} = \frac{10*50+30*30+50*20+75*10+105*5}{50+30+20+10+5} = \frac{3675}{115} = 31.95652$

$\hat{\lambda} = \frac{1}{31.95652}$

The *c d f* of exponential distribution is :$F(x) = 1-e^{-\lambda x}; x > 0$

Here, $F(x) = 1 - e^{\frac{-x}{31.95652}}$

$F(20) = 1 - e^{\frac{-20}{31.95652}} = 0.4652$

$F(40) = 1 - e^{\frac{-40}{31.95652}} = 0.7140$

$F(60) = 1 - e^{\frac{-60}{31.95652}} = 0.8470$

$F(90) = 1 - e^{\frac{-90}{31.95652}} = 0.9402$

$F(120) = 1 - e^{\frac{-120}{31.95652}} = 0.9766.$

Probability	*Class*	*Expected* E_i
0.4652	$0 \le x \le 20$	53.498
0.2488	$20 \le x \le 40$	28.612
0.1330	$40 \le x \le 60$	15.295
0.0932	$60 \le x \le 90$	10.718
0.0364	$90 \le x \le 120$	* 4.186
0.0234	$x \ge 120$	* 2.691

*Combine classes so that $E_i \ge 5$

Class	*Observed:* O_i	*Expected:* E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
$0 \le x \le 20$	50	53.498	-3.498	12.236	0.229
$20 \le x \le 40$	30	28.612	1.388	1.927	0.067
$40 \le x \le 60$	20	15.295	4.705	22.137	1.447
$60 \le x \le 90$	10	10.718	-0.718	0.516	0.048
$90 \le x$	5	6.877	-1.877	3.523	0.512
				Sum	**2.303**

H_0: Exponential distribution is appropriate
H_1: Exponential distribution is not appropriate

The degrees of freedom is: 5-1-1 = 3 (5 classes, 1 constraint $\sum E = \sum O$ and estimation of one parameter λ)

$\chi^2_{5\%,3} = 7.815$

As calculated value is lower than tabulated value of χ^2, there is no evidence at 5% significant level to suggest that an exponential distribution is not appropriate.

[9 Marks]

Solution 12:

i) From the given data we have:

$\bar{x}$ = 5.5 $\bar{y}$ =72.9

S_{xx}=$\sum$x^2 –n $\bar{x}^2$ =385-10 × $(\frac{55}{10})^2$ = 385-302.5 =82.5

S_{yy}=$\sum$y^2 –n $\bar{y}^2$ =53,363-10 × $(\frac{729}{10})^2$ = 53,363-53,144.1=218.90

S_{xy}=$\sum$xy –n$\bar{x}$ $\bar{y}$ =3881-10 ×5.5 ×72.9= 3881-4009.5=-128.5

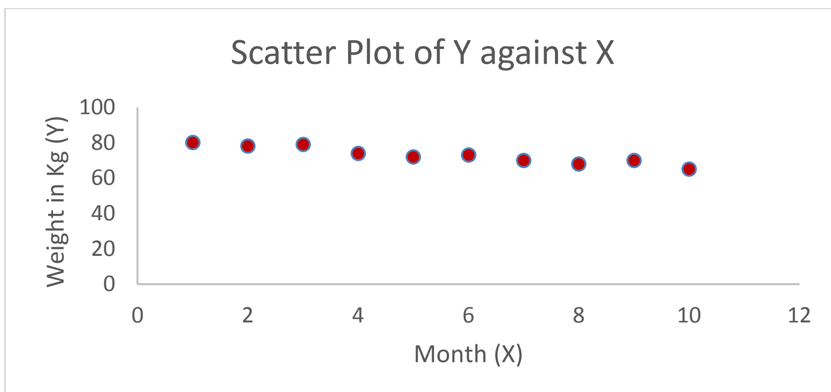
$\hat{\beta}$ = $\frac{S_{xy}}{S_{xx}}$ =$\frac{-128.5}{82.5}$ =-1.55758

$\hat{\alpha}$ = $\bar{y}$ -$\hat{\beta}\bar{x}$ =72.9 –(-1.55758) ×5.5 =81.46667

$\hat{y}$ =81.46667 -1.55758 x

[3]

ii)



The model seems to be appropriate. [3]

iii) $\hat{\sigma}^2 = \frac{1}{(n-2)} \times (S_{yy} - \frac{S_{xy}^2}{S_{xx}})$ $=\frac{1}{8} \times (218.90 - \frac{-128.5^2}{82.5})$ =2.34394 [1]

iv) H_0: β =0 *vs* H_1: β < 0

The test statistic is

$$\frac{\hat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \sim t_{n-2}$$

If H_0 is true then $\frac{\hat{\beta}-0}{\sqrt{\frac{2.3439}{82.5}}} \sim t_8$

Observed value of t_8 is $\frac{-1.55758-0}{\sqrt{\frac{2.3439}{82.5}}}$ = -9.2407

This is much less than –5.041, the lower 0.05% point of the t_8 distribution.

So, we reject H_0 at the 0.05% level and conclude that there is very strong evidence that β < 0. [4]

v) We know from part (i) that

$\hat{y}$ =81.46667 -1.55758 x

$\hat{y}$=81.46667 –(1.55758×20) =50.3151

Standard error of the estimate

$$\sqrt{\sigma^2\{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2)}{s_{xx}}\}}$$

$$= \sqrt{2.34934\{1+\frac{1}{10}+\frac{(20-5.5)^2)}{82.5}\}}$$

=2.92435

95% confidence interval for $\hat{y}$ *is* 50.3151± $t_{0.025,n-2}$ $SE(\hat{y})$

=50.3151± (2.306×2.92435)

(43.57155, 57.05865) [4]

vi)

From above, estimate of mean weight is 50.3151.

Standard Error of estimate= $\sqrt{\sigma^2\{\frac{1}{n}+\frac{(x_i-\bar{x})^2)}{s_{xx}}\}} = \sqrt{2.34934\{\frac{1}{10}+\frac{(20-5.5)^2)}{82.5}}$

=2.4916

So 95% CI for mean weight is 50.3151± $t_{0.025,n-2}$ SE($\hat{y}$)

$=50.3151 \pm (2.306 \times 2.4916)$

$= (44.5695, 56.0607)$

CI for the predicted weight on day 20 is wider than CI for mean weight on day 20. The difference reflects the fact that we are able to estimate mean weight more precisely than individual weight on day 20.

The reason being, the error in the estimate of mean weight is only due to the fact that the population regression line is being estimated by a sample regression line, whereas the error in the prediction of the one particular weight is due to the error in estimating the mean weight plus the variation in weights. [4]

[19 Marks]

Solution 13:

i) Least Square Estimates:

$\hat{\mu} = 405/15 = 27$

$\hat{\tau}_i = \frac{\sum_j yij}{ni} - \frac{\sum_i \sum_j yij}{n}$

$\hat{\tau}_1 = 135/5 - 27 = 0$

$\hat{\tau}_2 = 160/5 - 27 = 5$

$\hat{\tau}_3 = 110/5 - 27 = -5$ [4]

ii) *Assumption:*

Observation are independent and they are from normal populations with common variance.

Hypothesis:

H_0: Average wheat yield for all varieties of wheat is same

H_1: At least two mean yields are different.

We have

$SST = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{n} = 11385 - \frac{405^2}{15} = 450$

$SSB = \sum \frac{y_{i.}^2}{ni} - \frac{y_{..}^2}{n} = \left(\frac{135^2}{5} + \frac{160^2}{5} + \frac{110^2}{5}\right) - \frac{405^2}{15} = 250$

$SSR = SST - SSB = 450 - 250 = 200$

ANOVA table:

Source of Variation	*DF*	*Sum of Squares*	*Mean Squares*
Between Treatments	2	250	125
Residuals	12	200	16.667
Total	14	450	

The variance ratio $F = \frac{125}{16.667} = 7.5$
The critical value of table $F_{(2,12)\ 0.05}$ is 3.885

Since the computed value of the test statistic is greater than the critical value, it is significant.

Hence the value of test statistic F is significant and we reject H_0 at 5% level of significance and conclude that varieties of wheat differ significantly as yields are different. [7]

[11 Marks]

****