Tutorial Letter 204/2/2016

Applied Statistics II

STA2601

Semester 2

Department of Statistics

Trial Examination Paper Solutions

BAR CODE

Dear Student
This is the last tutorial letter for 2016. I would like to take this opportunity again of wishing you well in the coming examination and I also wish you success in all your examinations.

Tutorial letters

You should have received the following tutorial letters:

Tutorial letter no.	Contents
101	General information and assignments.
102	Updated information.
103	Installation of SAS JMP 11.
104	Trial paper.
201	Solutions to assignment 1.
202	Solutions to assignment 2.
203	Solutions to assignment 3.
204	Solutions to trial paper (this tutorial letter).

Some hints about the examination:

- For hypothesis testing always

(i) give the null hypothesis to be tested

(ii) calculate the test statistic to be used

(iii) give the critical region for rejection of the null hypothesis

(iv) make a decision (*reject/do not reject*)

(v) give your conclusion.

- Whenever you make a conclusion in hypothesis testing we never ever say "**we accept $\mathbf{H_0}$**." The two correct options are "**we do not reject $\mathbf{H_0}$**" or "**we reject $\mathbf{H_0}$**".

- Always show **ALL** workings and maintain **four decimal places**.

- Always specify the level of significance you have used in your decision. For example H_0 *is rejected at the* 5% *level of significance* / *we do not reject* H_0 *at the* 5% *level of significance*.

- Always determine and state the rejection criteria. For example if $\text{F}_{\text{table value}} = 3.49$. Reject H_0 if f is greater than 3.49.

- Use my presentation of the solutions as a model for what is expected from you.

Solutions of May/June 2016 Final Examination

QUESTION 1

(a) (i) $n(\mu, \frac{\sigma^2}{n})$ variate. (1)

(ii) χ^2_n variate. (1)

(iii) t_{n-1} variate. (1)

(b) (i) $W = \sum_{i=1}^{10} \frac{(X_i - 73)^2}{16}$ is defined as the sum of 10 independent squared $n\,(0;\ 1)$ variates. Using **result 1.2 in our study guide (page 29),** $W \sim \chi^2_n \Longrightarrow W \sim \chi^2_{10}$. Since $W \sim \chi^2_{10}$, it follows from the properties of the chi-square distribution that $E\,(W) = 10$ using **result 1.1 in our study guide (page 28).** (2)

(ii) Since $U_1 \sim \chi^2_5$ and $U_2 \sim \chi^2_4$, then $V = \dfrac{U_1/5}{U_2/4} \sim F_{5;\ 4}$ using definition 1.21. (2)

[7]

QUESTION 2

(a) (i)

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f_X\,(x_i;\ \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} \\
&= \frac{1}{\theta} e^{-x_1/\theta} \times \frac{1}{\theta} e^{-x_2/\theta} \times \ldots \times \frac{1}{\theta} e^{-x_n/\theta} \\
&= \frac{1}{\theta^n} e^{-(x_1/\theta + x_2/\theta + \ldots + x_n/\theta)} \\
&= \theta^{-n} e^{-\sum x_i/\theta} \quad \text{(see definition 2.5)}
\end{aligned}$$

$$\Longrightarrow \ln L(\theta) = -n \ln \theta - \sum x_i \theta^{-1}$$

$$\Longrightarrow \frac{\partial \ln L(\theta)}{\partial \theta} = \frac{-n}{\theta} - \frac{\sum x_i}{\theta^2}(-1)$$

If we set $\dfrac{\partial \ln L(\theta)}{\partial \theta} = 0$ we get

$$\begin{aligned} \frac{\sum x_i}{\theta^2} &= \frac{n}{\theta} \\ \Longrightarrow \widehat{\theta} &= \frac{\sum X_i}{n} \\ &= \overline{X} \end{aligned}$$

Thus $\widehat{\theta} = \overline{X}$ (the maximum likelihood estimator(m.l.e.) of θ)

(6)

(ii) To show that the m.l.e. is an unbiased estimator, we have to show that $E(\widehat{\theta}) = \theta$.

$E(\widehat{\theta}) = E\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n}\sum E(X_i) = \frac{1}{n}n\theta = \theta$ (q.e.d.). (4)

(b) (i)

$$\begin{aligned} S^{2\prime} &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\ &= \frac{\sigma^2}{n}\left[\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2\right] \text{ (multiplying both sides by } \sigma^2\text{)} \\ &\Longrightarrow E(S^{2\prime}) = \frac{\sigma^2}{n}E\left[\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2\right] \end{aligned}$$

Now $E\left(\sum_{i=1}^{n}\left[\frac{X_i - \overline{X}}{\sigma}\right]^2\right) = n-1$ since $\sum_{i=1}^{n}\left[\frac{X_i - \overline{X}}{\sigma}\right]^2 \sim \chi^2_{n-1}$.

$$\begin{aligned} \Longrightarrow E\left(S^{2\prime}\right) &= \frac{\sigma^2}{n}E\left[\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2\right] \\ &= \frac{\sigma^2}{n} \times n-1 \\ &= \frac{n-1}{n}\sigma^2 \neq \sigma^2 \end{aligned}$$

Thus, $S^{2\prime}$ is an unbiased estimator for σ^2. (3)

(ii) From result 1.3 we know that $\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$

This also means (from result 1.1) that $Var\left(\sum_{i=1}^{n}\left[\frac{X_i-\overline{X}}{\sigma}\right]^2\right)=2(n-1).\ldots\ldots\ldots\ldots(1)$

Now

$$\begin{aligned}
S^{2\prime} &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\\
&= \frac{\sigma^2}{n}\left[\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2\right] \quad \text{(multiplying both sides by } \sigma^2\text{)}\\
\Longrightarrow Var(S^{2\prime}) &= \left(\frac{\sigma^2}{n}\right)^2 Var\left[\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2\right]\\
&= \frac{\sigma^4}{n^2}\cdot 2\left(n-1\right) \qquad \text{(from equation (1))}\\
&= \frac{2\sigma^4(n-1)}{n^2}
\end{aligned}$$

Similarly

$$\begin{aligned}
S^{2} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\\
&= \frac{\sigma^2}{n-1}\left[\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2\right] \quad \text{(multiplying both sides by } \sigma^2\text{)}\\
\Longrightarrow Var(S^{2}) &= \left(\frac{\sigma^2}{n-1}\right)^2 Var\left[\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2\right]\\
&= \frac{\sigma^4}{(n-1)^2}\cdot 2\left(n-1\right) \qquad \text{(from equation (1))}\\
&= \frac{2\sigma^4}{(n-1)}
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{2\sigma^4}{n-1} &> \frac{2\sigma^4(n-1)}{n^2}\\
\Longrightarrow & \quad var(S^2) > var(S^{2\prime})
\end{aligned}$$

(6)

[19]

QUESTION 3

(a) We have to test

H_0 : The observations come from a normal distribution.

H_1 : The observations do not come from a normal distribution.

(3)

(b) $n = \quad \sum_{i=1}^{n} X_i = 1\,312$

$$\widehat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} == \frac{1\,312}{42} \approx 31.2381$$

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} \\ &= \frac{2\,381.6190}{42} \\ &= 56.7052 \end{aligned}$$

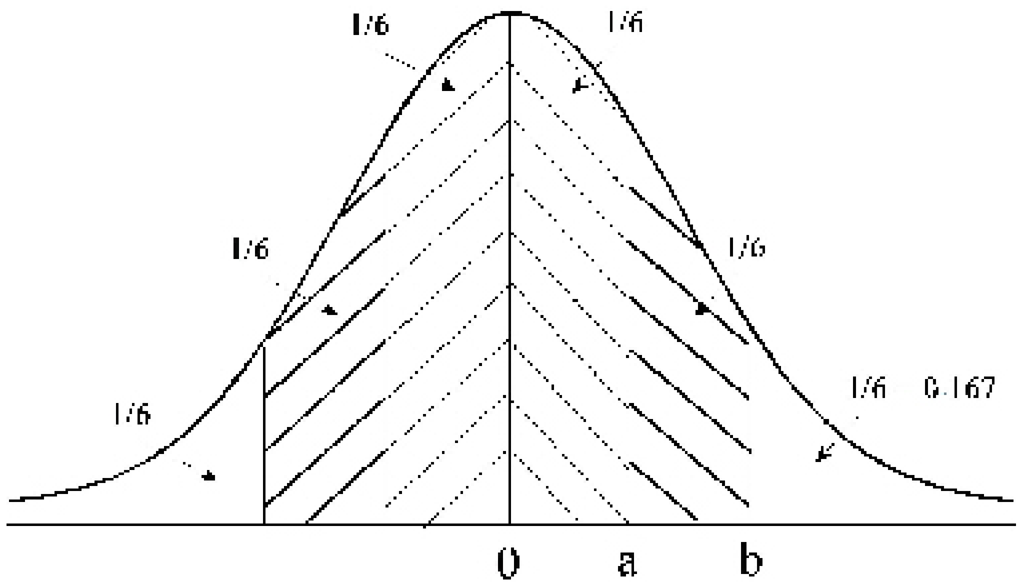
(5)

(c) If we divide the observations into 6 classes with equal expected frequencies, it means that $\pi_i = \frac{1}{6}$ for each interval $\Rightarrow n\pi_i = 7$.

The biggest problem is to *determine the interval limits* in terms of the X-scale such that each interval has a probability of $\frac{1}{6} = 0.167$.

We start with the standardised $n\,(0; 1)$ scale (as always) and transform back to the X-scale by making use of

$$Z = \frac{X - \widehat{\mu}}{\widehat{\sigma}} = \frac{X - 31.2381}{\sqrt{56.7052}}.$$



For the **first interval** we have that $P[Z \leq -b] = 0.167$ and we will find b by working with the

"positive mirror image", i.e. $P[Z \geq b] = 0.167 \quad \Longrightarrow \quad P[Z \leq b] = 0.833$ (this probability is $1 - 0.167$).

From table II we get that $\Phi(0.966) = 0.833 \Rightarrow b = 0.966.$

The first interval is where $Z \leq -0.966$ (Note that $P[Z \leq -0.966] = 0.167$)

$$\Longrightarrow \frac{X - \widehat{\mu}}{\widehat{\sigma}} = \frac{X - 31.2381}{\sqrt{56.7052}} = \frac{X - 31.2381}{7.5303} \leq -0.966$$

$$\begin{aligned} \therefore X &\leq 31.2381 - 0.966(7.5303) \\ &= 23.9638 \text{ (}X\text{-scale)} \end{aligned}$$

Hence, the first interval is where $X \leq 23.9638$

(4)

(d) The normality assumption is not violated because from the JMP graphical output we see that the normal curve does fit the histogram very well. The box plot depicts an almost symmetric distribution. From the normal quantile plot we see no systematic deviation around the line. So we conclude from the graphical output that the sample comes from a normal distribution. and the points seem not to be deviating from the diagonal on the Normal Quantile Plot. (3)

(e) We have to test $H_0 : \mu = 28$ against $H_1 : \mu > 28$.

Method I: Using the critical value approach:

$$\begin{aligned} t_{calc} &= \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{S} \\ &= \frac{\sqrt{42}\,(31.2381 - 28)}{1.1760} \\ &\approx 2.7534 \end{aligned}$$

Test is one-tailed. $\alpha = 0.05$. The critical value is $t_{\alpha;n-1} = t_{0.05;41} = 1.645$. Reject H_0 if $t_{calc} > 1.645$.

Since $2.7534 > 1.645$, we reject H_0 at the 5% level of significance and conclude that the mean weight taken to respond to customer compliants is greater than 28, that is, $\mu > 28$.

Method II: Using the p-value approach

p-value $= 0.0044$. Since $0.0044 < 0.05$, we reject H_0 at the 5% level of significance and conclude that the mean weight taken to respond to customer compliants is greater than 28, that is, $\mu > 28$.

(4)

(f) We want to test:

$H_0 : \sigma^2 = 64$ against $H_1 : \sigma^2 > 64$

Method I: Using the critical value approach:

Assuming μ is unknown, i.e., $\widehat{\mu} = \overline{X}$, then the test statistic is

$$U = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{2\,381.6190}{64} \approx 37.2128$$

$$\alpha = 0.05$$

$$\begin{aligned} \chi^2_{\alpha;\ n-1} &= \chi^2_{0.05;41} \\ &= 55.7585 + \frac{1}{10}\,(67.5048 - 55.7585) \\ &= 55.7585 + \frac{1}{10}\,(11.7463) \\ &= 56.9331 \end{aligned}$$

Reject H_0 if $U > 56.9331$.

Since $37.2128 < 56.9331$, we do not reject H_0 at the 5% level of significance and conclude that $\sigma = 64$.

Method II: Using the p-value approach

p-value $= 0.6397$. Since $0.6397 > 0.05$, we do not reject H_0 at the 5% level of significance and conclude that $\sigma = 64$. (4)

[23]

QUESTION 4

(a) $n_1 = 5$ $\quad \Sigma X_{1i} = 25$ $\quad \Sigma\left(X_{1i} - \overline{X}_1\right)^2 = 14$ $\quad H_0: \ \sigma_1^2 = \sigma_2^2$ against $H_1:$

$n_2 = 7$ $\quad \Sigma X_{2i} = 42$ $\quad \Sigma\left(X_{2i} - \overline{X}_2\right)^2 = 16$

$\sigma_1^2 \neq \sigma_2^2$

$$\begin{aligned} n_1 &= 5 & n_2 &= 7 \\ S_1^2 &= \frac{1}{n_1 - 1}\Sigma\left(X_{1i} - \overline{X}_1\right)^2 & S_2^2 &= \frac{1}{n_2 - 1}\Sigma\left(X_{2i} - \overline{X}_2\right)^2 \\ &= \frac{1}{5-1}(14) & &= \frac{1}{7-1}(16) \\ &= \frac{1}{4}(14) & &= \frac{1}{6}(16) \\ &= 3.5 & &= 2.6667 \end{aligned}$$

The test statistic is

$$\begin{aligned} F &= \frac{\sigma_2^2}{\sigma_1^2} \times \frac{S_1^2}{S_2^2} \\ &= 1 \times \frac{3.5}{2.6667} \\ &\approx 1.3125 \end{aligned}$$

The critical values are $F_{\frac{\alpha}{2};n_1-1;n_2-1} = F_{0.025;4;6} = 6.23$ and $F_{1-\frac{\alpha}{2};n_1-1;n_2-1} = \dfrac{1}{F_{\frac{\alpha}{2};n_2-1;n_1-1}} = \dfrac{1}{F_{0.025;6;4}} = \dfrac{1}{9.20} \approx 0.1087$

Reject H_0 if $F < 0.1087$ or if $F > 6.23$.

Since $0.1087 < 1.3125 < 6.23$, we do not reject H_0 at the 5% level of significance and conclude that the variances are equal, that is, $\sigma_1^2 = \sigma_2^2$. (6)

(b) The test is based on the assumptions that:

- The samples are independent.
- Both samples are from normal populations.

(3)

(c) $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

$$\overline{X}_1 = \frac{1}{n_1}\Sigma X_{1i} = \frac{1}{5}(25) = 5 \qquad \overline{X}_2 = \frac{1}{n_2}\Sigma X_{2i} = \frac{1}{7}(42) = 6$$

The test statistic is

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Now

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{\Sigma\left(X_{1i} - \overline{X}_1\right)^2 + \Sigma\left(X_{2i} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} \\ &= \frac{14 + 16}{5 + 7 - 2} \\ &= \frac{30}{10} \\ &= 3 \\ \Longrightarrow & \quad S_{pooled} = \sqrt{3} \approx 1.7321 \end{aligned}$$

Then

$$
\begin{aligned}
T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \\
&= \frac{(5-6)-(0)}{1.7321\sqrt{\dfrac{1}{5}+\dfrac{1}{7}}} \\
&= \frac{-1}{1.7321\sqrt{0.342857142}} \\
&= \frac{-1}{1.01421391} \\
&\approx 0.9860
\end{aligned}
$$

The critical value is $t_{\alpha/2;n_1+n_2-2} = t_{0.025;10} = 2.228$.Reject H_0 if $T < -2.228$ and $T > 2.228$.

Since $-2.228 < 0.9860 < 2.228$, we do not reject H_0 at the 5% level and conclude that the mean results of the two processes are not significantly different from each other, i.e., $\mu_1 = \mu_2$. (7)

[16]

QUESTION 5

(a) (i) Yes, it may be reasonable to assume that the five groups may be considered as *independent groups* if the respondents in one group do not influence the opinion of the other respondents in the other groups. (2)

(ii) No formal tests for normality are included in the output and the graphical output shows only the "Means Diamonds" which is not a graphical test for normality. To perform the ANOVA we simply have to assume that the five groups may be considered as coming from normal populations. (2)

(iii) We have to test:

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$, against $H_1 : \sigma_p^2 \neq \sigma_q^2$ for at least one $p \neq q$

From **Figure 5**, we conclude that all the tests for the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$ are not significant at the 5% level of significance. Using the Levene's test, p-value $= 0.0885$. Since $0.0885 > 0.05 \Longrightarrow$ we can not reject H_0 at the 5% level of significance. The assumption of *equal variances* is not violated. (3)

(b) (i) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ against
$H_1 : \mu_p \neq \mu_q$ for at least one $p \neq q$.

(2)

(ii) The test statistic is $F = \frac{MSTr}{MSE} \sim F_{k-1;n-k}$

(iii) From the output: Computations for ANOVA we see that $F = 12.5621$ **which is highly significant** with a p–value of $< 0.0001 << 0.05$. We reject H_0 in favour of H_1 at the 5% level of significance and conclude that there is a significant difference in the population **mean ratings** among the five groups, that is, $\mu_p \neq \mu_q$ for at least one $p \neq q$.

(3)

(c) All pairs of means differ significantly except for the pairs *"A-B", "A-E", "B-E", "C-D"* and *"E-C"*. This is graphically confirmed by the "Means Diamonds" where we can see that the pairs have almost identical pictures. There is overlap between the "Means Diamonds". On the "All Pairs Tukey-Kramer" display the pairs of circles overlap completely with one for group A and group B almost being the same. From the output of the formal statistical test we see that the confidence interval for the difference of the mean ratings for the pairs **include zero** and we canot reject the null hypothesis of equal means. The confidence intervals for the pairs *"A-B", "A-E", "B-E", "C-D"* and *"E-C" are* $(-4.31646; 4.98313)$, $(-1.98313; 7.31646)$, $(-2.31646; 6.98313)$, $(-2.31646; 6.98313)$ and $(-0.64979; 8.64979)$ respectively. Confirming this are the p-values which are all greater than 0.05, with values 0.9995, 0.4611, 0.5881, 0.5881 and 0.1166 respectively.

The $Abs(Dif) - LSD$ for the pairs *"A-B", "A-E", "B-E", "C-D"* and *"E-C"* are -4.3165, -1.9831, -2.3165, -2.3165 and -0.6498 respectively and are all negative showing that pairs of means are not significantly different from each other. The groups *A, B* and *E* share the letter A, *E* and *C* share the letter B and *C* and *D* share the letter C.

All the other intervals for the difference of the means are (positive value; positive value) which **excludes zero** and means **we reject** $\mu_p = \mu_q \Longrightarrow \mu_p \neq \mu_q$.

(5)

(d) One can use the advertisement *A, B* and *E* and avoid *C* and *D* beacuse they had low mean ratings. Thus underselling and correctly selling the pen's characteristtics to the public results in higher ratings than overselling it. (3)

[20]

QUESTION 6

(a) The summary statistics are

$$n = 7 \qquad \Sigma x_i = 35 \qquad \Sigma\,(x_i - \overline{x})^2 = 28$$

$$\Sigma y_i\,(x_i - \overline{x}) = 235 \qquad \Sigma y_i = 620 \qquad \Sigma\,(y_i - \overline{y})^2 = 2335.7143$$

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\Sigma y_i\,(x_i - \overline{x})}{\Sigma\,(x_i - \overline{x})^2} = \frac{235}{28} \approx 8.392857 \\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} = 88.571429 - (8.392857)\,(5) \\
&= 46.607144
\end{aligned}$$

So the least squares regression equation for Y on X is:

$$Y = 46.607144 + 8.392857X.$$

(6)

(b) $x = 6\,500 = 6.5$ thousands. The predicted sales volume is

$$\begin{aligned}
\widehat{Y}_i &= 46.607144 + 8.392857X \\
&= 46.607144 + 8.392857\,(6.5) \\
&= 46.607144 + 54.5535705 \\
&\approx 101.1607
\end{aligned}$$

Thus the estimated sales is $R34\,101\,160.70$. (2)

(c) The variance of the estimate is $Var\widehat{Y}(x) = \sigma^2 \left[\frac{1}{n} + (x - \overline{x})^2/d^2\right]$ since $x = 6.5$ is a value within the original domain (See p. 219 study guide.)

Now $s^2 = 72.68$

So the approximate (or estimated) value for this variance when $x = 6.5$ is:

$$\begin{aligned}
\widehat{Var\widehat{Y}}(x) &= s^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{d^2}\right] \\
&= 72.68 \left[\frac{1}{7} + \frac{(6.5 - 5)^2}{28}\right] \\
&= 72.68 \left[\frac{1}{7} + \frac{9}{112}\right] \\
& 72.68 \left[\frac{25}{112}\right] \\
&= 16.2232
\end{aligned}$$

$\therefore$ Standard error of estimate is $\sqrt{16.2232} \approx 4.0278$ (3)

(d) We have to test $H_0: \ \beta_1 = 0$ against

$H_1: \ \beta_1 \neq 0.$

Method I: Using the critical value approach:

From the output:

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - B_1}{s/d} \\ &= \frac{8.3929 - 0}{1.6111} \\ &\approx 5.21 \end{aligned}$$

$\alpha = 0.05$ $\alpha/2 = 0.025$ $t_{\alpha/2;n-2} = t_{0.025;5} = 2.571$. Reject H_0 if $T < -2.571$ or if $T > 2.571$ or if $|T|| > 2.571$.

Since $5.21 > 2.571$, we reject H_0 in favour of H_1 at the 5% level significance and conclude that $\beta_1 \neq 0$. This means that the regression line is significant to explain the variability in y. (Only when $\beta_1 = 0$, does it imply that regression is meaningless.)

Method II: Using the p-value approach

p-value $= 0.0034 << 0.05$. We reject H_0 in favour of H_1 at the 5% level of significance and conclude that $\beta_1 > 0$. This means that the regression line is significant to explain the variability in y. (Only when $\beta_1 = 0$, does it imply that regression is meaningless.)

(4)

[15]

[100]