# Tutorial letter 204/2/2017

## Applied Statistics II

## STA2601

## Semester 2

## Department of Statistics

**Trial Examination Paper Solutions**

**Dear Student**
This is the last tutorial letter for 2017 semester 1. 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 13. |
| 104 | Trial paper. |
| 201 | Solutions to assignment 1. |
| 202 | Solutions to assignment 2. |
| 203 | Solutions to assignment 3. |
| 204 | Solutions to trial papers (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 $F_{\text{table value}} = 3.49$. 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 2017 Paper One Final Examination**

## QUESTION 1

(a) $E(T) = \theta$ (study guide page 41) (1)

(b) Variance (study guide page 43) (1)

(c) F-distribution, $(5, 9)$ degrees of freedom (study guide page 32, definition 1.21) (2)

(d) Independence (study guide page 121) (1)

(e) Type I error (study guide page 28) (1)

**[6]**

## QUESTION 2

(a) (i)

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} P(X_i = r_i) \\ &= \frac{\lambda^{r_1} e^{-\lambda}}{\left(1 - e^{-\lambda}\right) r_1!} \frac{\lambda^{r_2} e^{-\lambda}}{\left(1 - e^{-\lambda}\right) r_2!} \cdots \frac{\lambda^{r_n} e^{-\lambda}}{\left(1 - e^{-\lambda}\right) r_n!} \\ &= \frac{\left(e^{-\lambda}\right)^n \lambda^{r_1 + r_2 + \ldots + r_n}}{\left(1 - e^{-\lambda}\right)^n r_1! r_2! \ldots r_n!} \\ &= e^{-\lambda n} \left(1 - e^{-\lambda}\right)^{-n} \lambda^{\sum_{i=1}^{n} r_i} \cdot \left(\prod_{i=1}^{n} r_i!\right)^{-1}. \end{aligned}$$

(4)

(ii) So

$$\ell nL(\lambda) = -n\lambda - n\ell n\left(1 - e^{-\lambda}\right) + \sum_{i=1}^{n} r_i \ell n(\lambda) - \sum_{i=1}^{n} \ell n(r_i!)$$

Since

$$\begin{aligned} \frac{\partial \ell n\left(1 - e^{-\lambda}\right)}{\partial \lambda} &= \frac{1}{\left(1 - e^{-\lambda}\right)} \frac{\partial}{\partial \lambda}\left(1 - e^{-\lambda}\right) \\ &= \frac{-e^{-\lambda}(-1)}{\left(1 - e^{-\lambda}\right)} \\ &= \frac{e^{-\lambda}}{\left(1 - e^{-\lambda}\right)} \end{aligned}$$

it follows that

$$\frac{\partial \ell nL(\lambda)}{\partial \lambda} = -n - \frac{ne^{-\lambda}}{\left(1 - e^{-\lambda}\right)} + \frac{\sum_{i=1}^{n} r_i}{\lambda} + 0, \text{ that is}$$

$$\frac{\partial \ell nL(\lambda)}{\partial \lambda} = -n - \frac{ne^{-\lambda}}{\left(1 - e^{-\lambda}\right)} + \frac{\sum_{i=1}^{n} r_i}{\lambda}$$

(4)

(b) $E(X_i) = 7\theta \qquad i = 1,\ 2,\ \ldots,\ n$

The least square estimator is

$$\begin{aligned} Q(\theta) &= \sum_{i=1}^{n} [X_i - E(X_i)]^2 \\ &= \sum_{i=1}^{n} (X_i - 7\theta)^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial Q}{\partial \theta} &= \sum_{i=1}^{n} 2(X_i - 7\theta)(-7) \\ &= -14 \sum_{i=1}^{n} (X_i - 7\theta) \end{aligned}$$

Set $\frac{\partial Q}{\partial \theta} = 0$

$$\begin{aligned} 0 &= -14\sum_{i=1}^{n}(X_i - 7\theta) \\ &= \sum_{i=1}^{n}(X_i - 7\theta) \\ &= \sum_{i=1}^{n} X_i - 7n\theta \\ 7n\theta &= \sum_{i=1}^{n} X_i \end{aligned}$$

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^{n} X_i}{7n} \\ \hat{\theta} &= \frac{1}{7}\overline{X} \end{aligned}$$

(6)

**[14]**

**QUESTION 3**

(a) Test for skewness:

$H_0$ : The distribution is normal $(\Rightarrow \beta_1 = 0)$.

$H_1$ : $\beta_1 \neq 0$.

(Please note: The alternative must be two-sided. There is no indication of a one-sided test.)

The critical value is $0.587$. Reject $H_0$ if $\beta_1 < -0.587$ or $\beta_1 > 0.587$ or $|\beta_1| > 0.587$.

$$\text{Now } \beta_1 = \frac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left(\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}\right)^3} = \frac{\frac{1}{40}(234.75)}{\left(\sqrt{\frac{1}{40}(1\,485.5)}\right)^3}$$

$$= \frac{5.86875}{\left(\sqrt{37.1375}\right)^3}$$

$$= \frac{5.86875}{(6.09405448)^3}$$

$$= \frac{5.86875}{226.3179482}$$

$$\simeq 0.0259$$

Since $-0.587 < 0.0259 < 0.587$, we cannot reject $H_0$ at the $10\%$ level of significance. It seems as if the sample comes from a symmetrical distribution. (7)

(b) (i) Divide the observations into $5$ classes with equal expected frequencies. This means that $\pi_i = \frac{1}{5} = 0.2$ for each interval $\Rightarrow n\pi_i = 40(0.2) = 8$.

Basically the problem is to determine the interval limits in terms of the $X$-scale such that each interval has a probability of $0.2$.

We start with the standardised $n\,(0;\,1)$ scale (as always) and transform back to the $X$-scale by making use of the transformation $Z = \dfrac{X-\mu}{\sigma} = \dfrac{X-20}{6}$.

Another "obstacle" is that table II (stoker) is only tabulated for probabilities $\geq 0.500$ (in other words for positive $Z$-values). So we need a little "manipulation" of the table in order to find $Z$ if

$$\underbrace{P(Z \leq z) = \Phi(z)}_{\text{by definition}} < 0.500$$

From the first interval we have that $P\,(Z \leq a) = 0.2$ and we must determine $a$.

From table II (under $\Phi\,(Z) = 0.8$) we find that $Z = 0.842$.

$\Longrightarrow P\,(Z \geq 0.842) = 0.2 = P\,(Z \leq -0.842)$

$$\begin{aligned}
\text{But } P\,(Z \leq -0.842) &= P\left(\frac{X-\mu}{\sigma} \leq -0.842\right) \\
&= P\left(\frac{X-20}{6} \leq -0.842\right) \\
&= P\,(X \leq 20 - 0.842 \times 6) \\
&= P\,(X \leq 20 - 5.052) \\
&= P\,(X \leq 14.948)
\end{aligned}$$

The first interval is where $X \leq 14.95$

(4)

(ii) Expected frequency is $n\hat{\pi}_i = \frac{1}{5} \times 40 = 8$.

| Table of observed and expected frequencies | | |
|---|---|---|
| **Class interval** | $O_i$ | $\hat{e}_i = n\hat{\pi}_i$ |
| $X < 14.95$ | 9 | 8 |
| $14.95 \leq X < 18.48$ | 9 | 8 |
| $18.48 \leq X < 21.52$ | 6 | 8 |
| $21.51 \leq X < 25.05$ | 7 | 8 |
| $X \geq 25.05$ | 9 | 8 |
| **Totals** | **40** | **40** |

(1)

(iii) Goodness of fit test:

We have to test

$H_0$ : The observations come from a $n\left(20, 6^2\right)$ distribution

$H_1$ : The observations do not come from a $n\left(20, 6^2\right)$ distribution.

We use the test statistic

$$Y^2 = \sum_{i=1}^{6} \frac{(N_i - n\pi_i)^2}{n\pi_i} \qquad \text{(page 98 study guide)}$$

$$\therefore Y^2 = \tfrac{1}{8}\left((9-8)^2 + (9-8)^2 + (6-8)^2 + (7-8)^2 + (9-8)^2\right)$$

$$= \tfrac{1}{8}(1+1+4+1+1)$$

$$= \frac{8}{8}$$

$$= 1$$

The critical value is

$$\chi^2_{\alpha;k-r-1} = \chi^2_{0.05;5-0-1} = \chi^2_{0.05;4} = 9.48773$$

Reject $H_0$ if $Y^2 \geq 9.48773$.

Since $1 < 9.48773$, we cannot reject $H_0$ at the 5% level of significance and conclude that the data follow a normal distribution.

FOR YOUR INFORMATION: For the degrees of freedom $k-r-1$ where $k$ is the number of classes and $r$ is the number of estimated parameters.

| Test | Value of $r$ | Parameter unknown |
|---|---|---|
| $n\left(20, 6^2\right)$ | $r=0$ | |
| $n\left(20, \sigma^2\right)$ | $r=1$ | $\sigma$ unknown |
| $n\left(\mu, 6^2\right)$ | $r=1$ | $\mu$ unknown |
| $n\left(\mu, \sigma^2\right)$ | $r=2$ | both $\mu$ and $\sigma$ unknown |

(7)

(c) (i) The assumptions are:

- observations are independent
- the data follows a normal distribution.

(2)

(ii) We have to test $H_0 : \mu = 18$ against $H_1 : \mu < 18$.

**Method 1: Using the critical value approach**

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{s} = \frac{\sqrt{40}\,(19.75 - 18)}{6.17169} \approx 1.7933$$

The critical value is

$$\begin{aligned} t_{\alpha;n-1} &= t_{0.05;39} \\ &= 1.690 + \frac{4}{5}(1.684 - 1.690) \\ &= 1.690 + \frac{4}{5}(-0.006) \\ &= 1.690 - 0.0048 \\ &\approx 1.6852 \end{aligned}$$

We will reject $H_0$ if $T \leq -1.685$.

Since $1.7934 > -1.685$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that $\mu = 18$, that is, the mean average number of orders is $18$.

**Method II: Using the p-value approach**

$p$-value $= 0.9597$. Since $0.9597 > 0.05$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that $\mu = 18$, that is, the mean average number of orders is $18$.

(3)

(iii) We have to test $H_0 : \sigma = 5$

against $H_1 : \sigma \neq 5$

**Method 1: Using the critical value approach**

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

$$\begin{aligned} U &= \frac{(n-1)\,S_X^2}{\sigma^2} \\ &= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} \\ &= \frac{1\,485.5}{5^2} \\ &\approx 59.42 \end{aligned}$$

The critical values are

$$\begin{aligned} &= \chi^2_{1-\alpha/2;n-1} && ; && = \chi^2_{\alpha/2;n-1} \\ &= \chi^2_{0.975;39} && ; && = \chi^2_{0.025;39} \\ &= 16.7908 + \tfrac{9}{10}(24.4331 - 16.7908) && ; && = 46.9792 + \tfrac{9}{10}(59.3417 - 46.9792) \\ &= 16.7908 + 0.9(7.6423) && ; && = 46.9792 + 0.9(12.3625) \\ &= 16.7908 + 6.87807 && ; && = 46.9792 + 11.12625 \\ &= 23.6689 && ; && = 58.1055 \end{aligned}$$

Reject $H_0$ if $U < 23.6689$ or $U > 58.1055$

Since $59.42 > 58.1055$, we reject $H_0$ at the $5\%$ level of significance and conclude that $\sigma \neq 5$.

**Method II: Using the p-value approach**

$p$-value $= 0.0382$. Since $0.0382 < 0.05$, we reject $H_0$ at the $5\%$ level of significance and conclude that $\sigma \neq 5$.

(3)

(iv) We are $95\%$ confident that the true average number of orders received is between 17.78 and 21.72, that is, if we select many random samples of the sample size, and if we calculate a confidence interval for each of these samples, then in about $95\%$of these cases, the population mean will lie within the interval $18$ to $22$. (1)

(d) We have to test: $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A < \mu_B$

$n_X = 40 \quad \overline{X} = 19.75 \quad S_X^2 = 38.09$

$n_Y = 30 \quad \overline{Y} = 22 \quad S_Y^2 = 42.25$

The test statistic is

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

Now

$$\begin{aligned} S_p^2 &= \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2} \\ &= \frac{(40 - 1)\,38.09 + (30 - 1)42.25}{40 + 30 - 2} \\ &= \frac{1\,485.51 + 1\,225.25}{68} \\ &= \frac{2\,710.76}{68} \\ &\approx 39.864118 \\ \Longrightarrow\ & S_{pooled} = \sqrt{39.864118} \approx 6.3138 \end{aligned}$$

Then

$$\begin{aligned} T &= \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \\ &= \frac{(19.75 - 22) - (0)}{6.3138\sqrt{\frac{1}{40} + \frac{1}{30}}} \\ &= \frac{-2.25}{6.3138\sqrt{0.058333333}} \\ &= \frac{-2.25}{1.524927575} \\ &\approx -1.4755 \end{aligned}$$

The critical value is

$$\begin{aligned} t_{\alpha;n_X+n_Y-2} &= t_{0.05;68} \\ &= 1.671 + \frac{8}{40}(1.660 - 1.671) \\ &= 1.671 + \frac{1}{5}(-0.011) \\ &= 1.671 - 0.0022 \\ &\approx 1.6688 \end{aligned}$$

We will reject $H_0$ if $T \leq -1.669$.

Since $-1.4755 > -1.669$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that $\mu_X = \mu_Y$, i.e., mean number of orders of population B is the same as the mean number of orders of population A.

(8)

**[36]**

**QUESTION 4**

(a) (i) The Mosaic Plot shows that the sample sizes for flawed and perfect are almost in the ratio $1:12$. The proportions of shift 1, 2 and 3 seem to be equal in proportion across all types of shift conditions as evidenced by the horizontal lines that are almost in alignment. The hypothesis of no association might not be rejected.

(3)

(ii) $H_0$: There are no differences in quality between the three shifts.

$H_1$: There are differences in quality between the three shifts.

(2)

(iii) The test statistic is $Y^2 = \sum_{k=1}^{k} \frac{\left(\text{Observed} - \text{Expected}\right)^2}{\text{Expected}}$ and the value is $Y^2 = 2.351$.

(2)

(iv) The critical value is $\chi^2_{0.05;2} = 5.99147$. Since $2.351 < 5.99147$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that there are no differences in quality between the three shifts.

OR

Alternatively the $p$-value is $= 0.3087$. Since $0.3087 > 0.05$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that there are no differences in quality between the three shifts.

(2)

(b) $n_1 = 17 \qquad S_1^2 = (1.12)^2 = 1.2544 \qquad n_2 = 12 \qquad S_2^2 = (2.36)^2 = 5.5696$

The $95\%$ confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$P\left(F_{1-\frac{\alpha}{2};n_2-1;n_1-1} < \frac{\sigma_1^2}{\sigma_2^2}\frac{S_2^2}{S_1^2} < F_{\frac{\alpha}{2};n_2-1;n_1-1}\right) = 1-\alpha$$

$$\left[\frac{F_{1-\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2};\frac{F_{\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}\right]$$

$\alpha = 0.05, \alpha/2 = 0.025$

$$\begin{aligned} F_{1-\frac{\alpha}{2};n_2-1;n_1-1} &= F_{0.975;11;16} = \frac{1}{F_{0.025;16;11}} \approx \frac{1}{3.33} \approx 0.3003 \\ F_{\frac{\alpha}{2};n_2-1;n_1-1} &= F_{0.025;11;16} = \frac{1}{2}(2.99 + 2.89) = 2.94 \end{aligned}$$

$\therefore$The 95% confidence interval is

$$\left[\frac{F_{1-\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2};\frac{F_{\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}\right]$$
$$\left[\frac{0.3003}{5.5696/1.2544};\frac{2.94}{5.5696/1.2544}\right]$$
$$\left[\frac{0.3003}{4.44005102};\frac{2.94}{4.44005102}\right]$$
$$[0.0676;\ 0.6622]\,.$$

We are 95% confident that the confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ lies between $0.0676$ to $0.6622$.

(6)

**[15]**

**QUESTION 5**

| Group | Short | Moderate | Large |
|---|---|---|---|
| $n$ | 6 | 6 | 6 |
| $\sum X_{ij}$ | 6 | 24 | 6 |
| $\overline{X}_i$ | 1 | 4 | 1 |
| $\sum \left(X_{ij} - \overline{X}_i\right)^2$ | 8 | 8 | 14 |

(a)

$$S_1^2 = \frac{1}{n_1-1}\sum\left(X_{1j}-\overline{X}_2\right)^2 \qquad S_2^2 = \frac{1}{n_2-1}\sum\left(X_{2j}-\overline{X}_2\right)^2$$

$$= \frac{1}{6-1}(8) \qquad = \frac{1}{6-1}(8)$$

$$= \frac{1}{5}(8) \qquad = \frac{1}{5}(8)$$

$$= 1.6 \qquad = 1.6$$

$$S_3^2 = \frac{1}{n_3-1}\sum\left(X_{3j}-\overline{X}_3\right)^2$$

$$= \frac{1}{6-1}(14)$$

$$= \frac{1}{5}(14)$$

$$= 2.8$$

From the computations above it, follows that $S_1^2 = 1.6$; $S_2^2 = 1.6$ and $S_3^2 = 2.8$.

(3)

(b) (i) Ordinary average = $\frac{1.6+1.6+2.8}{3} = \frac{6.0}{3} = 2.0$ (2)

(ii) $MSE = \frac{SSE}{kn-k}$.

For this $ANOVA$ problem, we have $k = 3$ (there are four groups) and $n = 5$ (the number of observations in each sample).

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij}-\overline{X}_i\right)^2 \qquad \textbf{OR} \quad SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij}-\overline{X}_i\right)^2$$

$$= 8+8+14 \qquad = 66 \quad \textbf{wrong figure given to students}$$

$$= 30$$

$$\therefore MSE = \frac{30}{3(6)-3} \qquad \therefore MSE = \frac{66}{3(6)-3}$$

$$= \frac{30}{15} \qquad = \frac{66}{15}$$

$$= 2. \qquad \text{The result in (i) = result in (ii).} \qquad = 4.4.$$

This makes perfect sense! $MSE$ is like a pooled variance or an average variance, because the assumption of $ANOVA$ is that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ and if these variances are unknown, we estimate it by pooling.

(4)

(c) We have to test

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

(ii) Using the critical value approach:

The critical value is $F_{2;15} = 3.68$. Reject $H_0$ if $F > 3.68$.

**Using the p-value approach:**

We reject $H_0$ if $p$–value $< 0.05$.

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

From the output: Computations for ANOVA we see that $F = 9.0$ which is significant with a $p$–value of $0.0027$. Since $0.0027 < 0.05$ we reject $H_0$ in favour of $H_1$ at the $5\%$ level of significance and conclude that $u_p \neq \mu_q$ for at least one pair $p \neq q$, that is, the mean amount of attitude change for the size of discrepancy are not the same.

(4)

(d) If the discrepancy is small or large, there is not much change in attitude as compared to when the size of the discrepancy is moderate. (2)

(e) Manually, we should have computed for each pair of means, a test statistic

$$T_{pq} = \frac{\overline{X}_p - \overline{X}_q}{\mathrm{S}_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

where we have samples of equal sizes if we want to incorporate the principle of the Bonferroni equality.

The Turkey–Kramer HSD that are shown in the JMP out perform individual comparisons that make adjustments for multiple test.

Confidence intervals that do not include zero imply that the pairs of means differ significantly. The pairs that do not include zero are "*Moderate - Large*" and "*Moderate - Small*". The confidence interval for the pairs are $(0.87917 : 5.120826)$; and $( 0.87917; 5.120826)$; respectively. These are the intervals that do not include zero and it means we reject the null hypothesis and conclude that $\mu_{Moderate} \neq \mu_{\mathrm{L\,arg\,}e}$ and $\mu_{Moderate} \neq \mu_{Short}$. This is also supported by the fact

that the $p$-values for the differences between the means are $0.0060$ and $0.0060$ respectively. All $p$-values are $<< 0.05$ (highly significant), leading to the rejection of the null hypothesis of equal means.

The pairs that do not have the same letter connecting them means that the pairs are significantly different from each other.

Confirming this is the **Abs(Dif)-HSD** which are $0.8792$ and $0.8792$ for the pairs "*Moderate - Large*" and "*Moderate - Small*" respectively. Since all of them are positive thus, the means are significantly different. (Recall a negative value of **Abs(Dif)-HSD** means the groups are not significantly different from each other.)

(4)

**[19]**

## QUESTION 6

(a) Yes. One can fit a simple linear regression since there is a strong positive relationship. (2)

(b) $\hat{\beta}_1 = 0.0090$. For every kilometre covered, *delivery time* increases by $0.0090$ days. (1)

(c) $X = 600$.

The predicted delivery time for a customer situated 600 kilometres from the company is

$$\begin{aligned} \widehat{Delivery\ time} &= 4.0177 + 0.0090(600) \\ &= 4.0177 + 5.4 \\ &= 9.4177 \end{aligned}$$

$\Longrightarrow$ It will be $9.42$ days. (1)

(d) The $95\%$ confidence interval for the slope $\beta_1$ is

$$\widehat{\beta}_1 \pm t_{\alpha/2;n-2} \times \frac{s}{d}$$

$\widehat{\beta}_1 = 0.0090 \qquad t_{\alpha/2;n-2} = t_{0.025;8} = 2.306$

Thus, the $95\%$ confidence interval for the slope $\widehat{\beta}_1$ is

$$\begin{array}{lcl} \widehat{\beta}_1 & \pm & t_{\alpha/2;n-2} \times \dfrac{s}{d} \\ 0.0090 & \pm & 2.306 \times 0.002384 \\ 0.0090 & \pm & 0.0055 \\ (0.0090 - 0.0055 & ; & 0.0090 + 0.0055) \\ (0.0035 & ; & 0.0145) \end{array}$$

(4)

(e) $R^2 = 0.638995 \Longrightarrow 63.90$ of the variability in *delivery time* is being explained / or accounted for by the least squares line. (2)

**[10]**

**[100]**

## Solutions of May/June 2017 Paper Two Final Examination

### QUESTION 1

(a) Chisquare. (1)

(b) Let X and Y be two random variables with correlation coefficient $\rho$. If X and Y are independent then $\rho = 0$ (i.e X any Y are uncorrelated). Thus the connection will be that $\rho = 0$. (3)

(c) The power of the test is 1 - P(Type II error), that is, $1 - \beta$ where $\beta$= P(Type II error). (2)

**[6]**

**QUESTION 2**

(a) The method of obtaining least square estimators of $\theta_1,\ ...,\ \theta_k$ are found by minimising

- $Q\left(\theta_1,\ ...,\ \theta_k\right) = \sum_{i=1}^{n} \left(X_i - E\left(X_i\right)\right)^2$
- Then derive $\frac{\partial Q}{\partial \theta_j}; \quad j = 1,\ ...,\ k$
- Then set $\frac{\partial Q}{\partial \theta_j} = 0; \quad j = 1,\ ...,\ k$ thus obtaining $k$ equations with $k$ unknowns, which are solved to obtain $\hat{\theta}_1,\ ...,\ \hat{\theta}_k$. (4)

(b) $E\left(X_i\right) = \theta \qquad i = 1,\ 2,\ \ldots,\ n$

The least square estimator is

$$\begin{aligned} Q\left(\theta\right) &= \sum_{i=1}^{n} \left[X_i - E\left(X_i\right)\right]^2 \\ &= \sum_{i=1}^{n} \left(X_i - \theta\right)^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial Q}{\partial \theta} &= \sum_{i=1}^{n} 2\left(X_i - \theta\right)\left(-1\right) \\ &= -2\sum_{i=1}^{n} \left(X_i - \theta\right) \end{aligned}$$

Set $\frac{\partial Q}{\partial \theta} = 0$

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}(X_i - \theta) \\ &= \sum_{i=1}^{n}(X_i - \theta) \\ &= \sum_{i=1}^{n} X_i - n\theta \\ n\theta &= \sum_{i=1}^{n} X_i \\ \hat{\theta} &= \frac{\sum_{i=1}^{n} X_i}{n} \\ \hat{\theta} &= \overline{X} \end{aligned}$$

(4)

(c)

$$\begin{aligned} f(x;\alpha;\beta) &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \qquad \text{for } x > 0 \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{for } x \leq 0 \end{aligned}$$

(i) The likelihood function

$$\begin{aligned} L(x;\beta) &= \prod_{i=1}^{n} f_X(X_i;\beta) \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} (X_i)^{\alpha-1} e^{-X_i/\beta} \\ &= \Gamma(\alpha)^{-1}\beta^{-\alpha} X_1^{\alpha-1} e^{-X_1/\beta} \cdot \Gamma(\alpha)^{-1}\beta^{-\alpha} X_2^{\alpha-1} e^{-X_2/\beta} \ldots\ldots \Gamma(\alpha)^{-1}\beta^{-\alpha} X_n^{\alpha-1} e^{-X_n/\beta} \\ &= \Gamma(\alpha)^{-n}\beta^{-\alpha n} (X_1.X_2 \ldots X_n)^{\alpha-1} e^{-X_1/\beta} \cdot e^{-X_2/\beta} \ldots e^{-X_n/\beta} \\ &= \Gamma(\alpha)^{-n}\beta^{-\alpha n} \left(\prod_{i=1}^{n} X_i\right)^{\alpha-1} e^{-\sum X_i/\beta}. \end{aligned}$$

(4)

(ii)

$$\begin{aligned} LogL\left(\beta\right) &= Log\left(\Gamma\left(\alpha\right)^{-n}\beta^{-\alpha n}\left(\prod_{i=1}^{n}X_i\right)^{\alpha-1}e^{-\sum X_i/\beta}\right) \\ &= -n\,\log\Gamma\left(\alpha\right)-\alpha n\,\log\,\beta+\left(\alpha-1\right)\sum\log\,X_i-\sum X_i/\beta \end{aligned}$$

(3)

(iii) To find M.L.E. we must set $\frac{\partial LogL\left(\beta\right)}{\partial\beta}=0$.

$$\begin{aligned} \frac{\partial LogL\left(\beta\right)}{\partial\beta} &= \frac{-\alpha n}{\beta}-(-1)\sum X_i\beta^{-2} \\ \Longrightarrow -\frac{\alpha n}{\beta}+\frac{\sum X_i}{\beta^2} &= 0 \\ \frac{\sum X_i}{\beta} &= \alpha n \\ \therefore \sum X_i &= \alpha n\beta \\ \therefore \widehat{\beta} &= \frac{\sum X_i}{\alpha n}. \end{aligned}$$

(4)

**[19]**

**QUESTION 3**

(a) (i) **Test for skewness:**

$H_0$ : The distribution is normal $\left(\Rightarrow \beta_1=0\right)$.

$H_1$ : $\beta_1\neq 0$.

(Please note: The alternative must be two-sided. There is no indication of a one-sided test.)

The critical value is $0.389$. Reject $H_0$ if $\beta_1<-0.389$ or $\beta_1>0.389$ or $|\beta_1|>0.389$.

$$\text{Now } \beta_1 = \frac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left(\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}\right)^3} = \frac{9.6}{\left(\sqrt{25}\right)^3}$$

$$= \frac{9.6}{(5)^3}$$

$$= \frac{9.6}{125}$$

$$= 0.0768$$

Since $0.0768 < 0.389$ we do not reject $H_0$ at the $10\%$ level of significance level and conclude that the distribution is symmetric.

(7)

(ii) **Test for kurtosis:**

We have to test:

$H_0$ : The distribution is normal $(\Rightarrow \beta_2 = 3)$.

$H_1$ : $\beta_2 \neq 3$.

Since we have a sample size greater than 50, the test is based on $\beta_2$ (page 111 in the study guide).

The size of the sample, $n = 100$. We reject $H_0$ at the $10\%$ significance level if $B_2 <$ lower $5\%$ point or $B_2 >$ upper $5\%$ point in table B.

The critical values are $2.35$ and $3.77$. Reject $H_0$ if $B_2 < 2.35$ or $B_2 > 3.77$.

Now the value of the test statistic is

$$\beta_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right]^2}$$

$$= \frac{832.2}{[25]^2}$$

$$= \frac{832.2}{625}$$

$$= 1.33152$$

$$\approx 1.3315$$

Since $1.3315 < 2.35$, we reject $H_0$ at the $10\%$ level of significance and conclude that the distribution does not have the kurtosis of a normal distribution.

(7)

(iii) No, the sample failed one test and we conclude that the distribution does not originate from a normal distribution.

(1)

(b) (i) The assumptions are:

- observations are independent
- the data follows a normal distribution

Now based on the assumption of **independent observations** and the assumption that the ages of defaulters have a **normal distribution** (i.e., the sample comes from a normal population) we may assume that

$$T = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{S} \sim t_{n-1}.$$

**Are they met?** The defaulters, that is, the people were drawn randomly, thus the assumption of **independent observations is met.**

The normal quantile plot shows that the points at both ends are not following a diagonal. They seem to slightly deviate from the line. Secondly the histogram and box plot shows that data is slightly positively skewed almost symmetric (Its subjective).

We need a proper test. The Shapiro-Wilk test for normality shows that the null hypothesis ($H_0$: Data comes from a normal distribution) would not be rejected ($p$-value $= 0.1566$), indicating that we may assume that the data does come from a normal distribution.

(4)

(ii) We have to test $H_0 : \mu \geq 55$ against $H_1 : \mu < 55$.

**Method 1: Using the critical value approach**

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{s} = \frac{\sqrt{30}\,(52.0667 - 55)}{17.1885} \approx -0.9347$$

The critical value is $t_{\alpha;n-1} = t_{0.05;29} = 1.699$

We will reject $H_0$ if $T \leq -1.699$.

Since $-0.9347 > -1.699$, we do not reject $H_0$ at the $5\%$ level of significance and conclude that $\mu \geq 55$ (older people are the ones defaulting), .that is, the average age of those who default is at least 55 years.

**Method II: Using the p-value approach**

$p$-value $= 0.1788$. Since $0.1788 > 0.05$, we do not reject $H_0$ at the 5% level of significance and conclude that $\mu \geq 55$ (older people are the ones defaulting), .that is, the average age of those who default is at least 55 years. (4)

(iii) Since $\sigma^2$ will be known, we use the z-distribution and the test statistic will be

$$Z = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{\sigma}$$

(2)

(iv) We are 95% confident that the true value of $\mu$ will be from $45.6484$ to $58.4849$. No. This confidence interval is for a two-tailed test. and the test in (ii) is a one-tailed test. (4)

**[29]**

**QUESTION 4**

(a) We want to test:

$H_0 : \pi_1 = 0.20, \pi_2 = 0.50, \pi_3 = 0.30$

$H_1$ : At least one of the proportion is significantly different from the one specified.

$$N = 1\,000 \Longrightarrow \quad n\pi_1 = 200 \qquad n\pi_2 = 500$$
$$\text{and} \qquad\qquad n\pi_3 = 300$$

The test statistic is

$$\begin{aligned} Y^2 &= \sum_{i=1}^{3} \frac{(N_i - n\pi_i)^2}{n\pi_i} \\ &= \frac{(250-200)^2}{200} + \frac{(480-500)^2}{500} + \frac{(270-300)^2}{300} \\ &= \frac{(50)^2}{200} + \frac{(-20)^2}{500} + \frac{(-30)^2}{300} \\ &= 12.5 + 0.8 + 3 \\ &= 16.3 \end{aligned}$$

Now $\chi^2_{\alpha;k-r-1} = \chi^2_{0.05;2} = 5.99417$.. Reject $H_0$ if $Y^2 \geq 5.99147$.

Since $16.3 > 5.99417$,we reject $H_0$ and conclude at the $5\%$ level that the proportions are significantly different from the one specified. (8)

(b) We have to test $H_0 : \sigma_1^2 = \sigma_2^2$

against $H_1 : \sigma_1^2 > \sigma_2^2$

$n_1 = 11 \qquad S_1^2 = 12 \qquad n_2 = 13 \qquad S_2^2 = 4$

The test statistic is

$$\begin{aligned} F &= \frac{\sigma_2^2}{\sigma_1^2} \times \frac{S_1^2}{S_2^2} \\ &= 1 \times \frac{12}{4} \\ &= 3 \end{aligned}$$

The critical values is $F_{\alpha;n_1-1;n_2-1} = F_{0.05;10;12} = 2.91$. Reject $H_0$ if $F > 2.91$.

Since $3 > 2.91$, we reject $H_0$ at the $5\%$ level of significance and conclude that the reaction times of males are more variable than the reaction time of females, i.e. $\sigma_1^2 > \sigma_2^2$.

(7)

(c) (i) We have to test $H_0 : \mu = 0$ against $H_0 : \mu < 0$.

**Method 1: Using the critical value approach**

$$T = -4.71429$$

The critical value is $t_{\alpha;n-1} = t_{0.05;19} = 1.729$. Reject $H_0$ if $T \leq -1.729$

Since $-4.71429 < -1.7291$, we reject $H_0$ in favour of $H_1$ at the 5% level of significance and conclude that $\mu < 0$.

**Method II: Using the p-value approach**

$p$-value $< 0.0001$. Since $0.0001 << 0.05$, we can reject $H_0$ in favour of $H_1$ at the 1% level of significance and conclude that $\mu < 0$. (4)

(ii)

| | | |
|---|---|---|
| $n = 20$ | $\overline{Y} = 3.7$ | std error$= 0.35$ |
| $\alpha = 0.05$ | $\alpha/2 = 0.025$ | $t_{\alpha/2;(n-1)} = t_{0.025;19} = 2.093$ |

The 95% confidence interval for the difference of two mean is

$$\begin{aligned}
\overline{Y} &\pm t_{\alpha/2;(n-1)} \times \frac{S_y}{\sqrt{n}} \\
\overline{Y} &\pm t_{\alpha/2;(n-1)} \times \text{std error} \\
-1.65 &\pm 2.093 \times 0.35 \\
-1.65 &\pm 0.7326 \\
(-1.65 - 0.7326) &; -1.65 + 0.7326) \\
(-2.3826) &; -0.9174)
\end{aligned}$$

(3)

**[22]**

**QUESTION 5**

(a) We have to test:

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

Using the Bartlett's test, $p$-value $= 0.6067$. Since $0.6067 > 0.05 \Longrightarrow$ we can not reject $H_0$ at the 5% level of significance. The assumption of equal variances is not violated.

(4)

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

(2)

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

From the output: Computations for ANOVA we see that $F = 14.0061$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 running time among the four treatments, that is, $\mu_p \neq \mu_q$ for at least one $p \neq q$. The **mean running times** for the different types of additives differ significantly (at any level of significance). This implies that $\mu_p \neq \mu_q$ for at least one pair $p \neq q$.

(4)

(c) Confidence intervals that include zero imply that the pairs of means are not significantly different from each other. Most pairs of means differ significantly except for the pair "*Control*" and "*Additive C*" and the pair "*Additive A*" and "*Additive B*". This is graphically confirmed by the "Means Diamonds" where we can see that "*Control*" and "*Additive C*" have almost identical pictures and their two circles overlap to a large extent on the "All Pairs Tukey-Kramer" display. The same is true for the pair "*Additive A*" and "*Additive B*" (their two circles overlap almost completely).

From the output of the formal statistical test we see that the confidence interval for the difference of the mean running time ("*Additive C*" - "*Control*") = $(-0.120679 : 0.1873455)$. We also see that the confidence interval for the difference of the mean running time ("*Additive A*" - "*Additive B*") = $(-0.120679 : 0.1873455)$. These are the only intervals which includes **zero** and implies we cannot reject $\mu_{Additive\ C} = \mu_{Control}$ and we cannot reject $\mu_{Additive\ A} = \mu_{Additive\ B}$.

All the other intervals for the difference of the means are (positive value; positive value) which excludes zero and means we reject $\mu_p = \mu_q \Longrightarrow \mu_p \neq \mu_q$.The Abs(Dif) LSD of *Additive C* and *Control* is negative $(-0.12068)$, they share the same letter A and the Abs(Dif) LSD of *Additive B* and *Additive A* is negative $(-0, 12068)$, they share the same letter B. However, the pairs $\mu_{Additive\ C}$ and $\mu_{Additive\ A}$ , $\mu_{Control}$ and $\mu_{Additive\ A}$ , $\mu_{Additive\ C}$ and $\mu_{Additive\ B}$ , $\mu_{Control}$ and $\mu_{Additive\ B}$are significantly different since their confidence intervals do not include zero. Their Abs(Dif) LSD are positive.

(5)

**[12]**

**QUESTION 6**

(a) $\hat{\beta}_0 = 790$, and $\hat{\beta}_1 = -7.18$ . Thus, the least squares regression line is

$\widehat{Y} = 790 - 7.18x \Longrightarrow \widehat{\text{Rate of flow}} = 790 - 7.18\text{Age}.$ (3)

(b) $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

**Method I: Using the critical value approach**:

From the output:

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - B_1}{s/d} \\ &= \frac{-7.18 - 0}{2.0793} \\ &\approx -3.45 \end{aligned}$$

$\alpha = 0.05$ $\alpha/2 = 0.025$ $t_{\alpha/2;n-2} = t_{0.025;8} = 2.306$. Reject $H_0$ if $T < -2.306$ or if $T > 2.306$ or if $|T| > 2.306$.

Since $-3.45 < -2.306$, 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 $= 0.0087 << 0.05$. We reject $H_0$ in favour of $H_1$ at the $5\%$ level of 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.)

(4)

(c) The $95\%$ confidence interval for $\beta_1$ is

$$\widehat{\beta}_1 \pm t_{\alpha/2;n-2} \times SEb_1$$

$\alpha = 0.05$ $\alpha/2 = 0.025$ $t_{\alpha/2;n-2} = t_{0.025;8} = 2.306$
$\widehat{\beta}_1 = -7.18$ $SEb_1 = 2.0793$

The confidence interval for $\widehat{\beta}_1$ is

$$\begin{array}{lcl} \widehat{\beta}_1 & \pm & t_{\alpha/2;n-2} \times SEb_1 \\ -7.18 & \pm & 2.306 \times 2.0793 \\ -7.18 & \pm & 4.7949 \\ (-7.18 - 4.7949 & ; & -7.18 - 4.7949) \\ (-11.9749 & ; & -2.3851) \end{array}$$

(4)

(d) The predicted rate of flow of blood in the kidney of a 70-year-old person is

$$\begin{array}{rcl} \widehat{\text{Rate of flow}} & = & 790 - 7.18(70) \\ & = & 790 - 502.6 \\ & = & 287.4 \end{array}$$

$\Longrightarrow$ It will be $287.4$. (1)

**[12]**

**[100]**