Tutorial Letter 204/1/2016

Applied Statistics II

STA2601

Semester 1

Department of Statistics

Trial Examination Paper Solutions

BAR CODE

Dear Student
This is the last tutorial letter for 2016 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 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 $F_{\text{table value}} = 3.49$. 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 October/November 2015 Final Examination

QUESTION 1

(a) (i)

$$\begin{aligned} E\left(T_1\right) &= E\left(\frac{X_1+X_2+X_2}{3}\right) \\ &= \left(\frac{E\left(X_1\right)+E\left(X_2\right)+E\left(X_3\right)}{3}\right) \\ &= \left(\frac{\mu+\mu+\mu}{3}\right) \\ &= \mu \end{aligned}$$

$$\begin{aligned} E\left(T_2\right) &= E\left(\frac{X_1+2X_2+2X_3}{5}\right) \\ &= \left(\frac{E\left(X_1\right)+2E\left(X_2\right)+2E\left(X_3\right)}{5}\right) \\ &= \left(\frac{\mu+2\mu+2\mu}{5}\right) \\ &= \mu \end{aligned}$$

Thus T_1 and T_2 are unbiased estimators of μ. (5)

(ii) The most efficient of the two estimators will be the one with the smallest variance. Thus, the variance of each estimator should be computed. Do you recall that if X and Y are stochastically independent variables then $Var(aX+bY)=a^2Var(X)+b^2Var(Y)$.

Now

$$\begin{aligned} Var(T_1) &= Var\left(\frac{X_1+X_2+X_2}{3}\right) \\ &= \frac{1}{9}Var\left(X_1\right)+\frac{1}{9}Var\left(X_2\right)+\frac{1}{9}Var\left(X_3\right) \\ &= \frac{1}{9}\sigma^2+\frac{1}{9}\sigma^2+\frac{1}{9}\sigma^2 \\ &= \frac{1}{3}\sigma^2 \\ &\approx 0.3333\sigma^2 \end{aligned}$$

$$\begin{aligned} Var(T_2) &= Var\left(\frac{X_1+2X_2+2X_3}{5}\right) \\ &= \frac{1}{25}Var\left(X_1\right)+\frac{4}{25}Var\left(X_1\right)+\frac{4}{25}Var\left(X3\right) \\ &= \frac{1}{25}\sigma^2+\frac{4}{25}\sigma^2+\frac{4}{25}\sigma^2 \\ &= \frac{9}{25}\sigma^2 \\ &= 0.36\sigma^2 \end{aligned}$$

$\therefore\ Var(T_1) < Var(T_2) \implies T_1$ is a more efficient estimator of μ. In other words we would prefer T_1 because it has a smaller variance (which is a desirable property for an estimator since the $\sqrt{\text{variance}}$ is often called the "error of the estimation"). (6)

(b) $E(W_i) = c_i^2\theta \qquad i = 1, 2, \cdots, k$

$$\begin{aligned} Q(\theta) &= \sum_{i=1}^{k} (W_i - E(W_i))^2 \\ &= \sum_{i=1}^{k} \left(W_i - c_i^2\theta\right)^2 \end{aligned}$$

Now

$$\begin{aligned} \frac{dQ}{d\theta} &= 2\sum_{i=1}^{k} \left(W_i - c_i^2\theta\right) \times -c_i^2 \\ &= -2\sum_{i=1}^{k} \left(c_i^2 W_i - c_i^4\theta\right) \\ &= -2\left(\sum_{i=1}^{k} c_i^2 W_i - \theta\sum_{i=1}^{k} c_i^4\right) \end{aligned}$$

Then set $\frac{dQ}{d\theta} = 0$

$$\implies 0 = -2\left(\sum_{i=1}^{k} c_i^2 W_i - \theta\sum_{i=1}^{k} c_i^4\right)$$

$$0 = \sum_{i=1}^{k} c_i^2 W_i - \theta\sum_{i=1}^{k} c_i^4$$

Making θ subject of the formula

$$\begin{aligned} \theta\sum_{i=1}^{k} c_i^4 &= \sum_{i=1}^{k} c_i^2 W_i \\ \widehat{\theta} &= \frac{\sum_{i=1}^{k} c_i^2 W_i}{\sum_{i=1}^{k} c_i^4}. \end{aligned}$$

(5)

[16]

QUESTION 2

(a) Small. (1)

(b) Type II error. (1)

(c) Independence. (1)

(d) Variance. (1)

(e) Skewness, symmetric**.** (2)

(f) $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - \overline{X}_i\right),$ $MSE = \dfrac{SSE}{(kn - k)}$

$$F = \frac{MSTr}{MSE} = \frac{n \sum_{i=1}^{k} \left(\overline{X}_i - \overline{X}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - \overline{X}_i\right)^2 / (kn-k)}$$ (4)

[10]

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 0.492. Reject H_0 if $\beta_1 < -0.492$ or $\beta_1 > 0.492$ or $|\beta_1| > 0.492$.

$$\text{Now } \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)^3} = \frac{\frac{1}{60}(-5\,832)}{\left(\sqrt{\frac{1}{60}(2\,160)}\right)^3}$$

$$= \frac{-97.2}{\left(\sqrt{36}\right)^3}$$

$$= \frac{-97.2}{(6)^3}$$

$$= \frac{-97.2}{216}$$

$$\approx -0.45.$$

Since $-0.492 < -0.45 < 0.492$ we cannot reject H_0 at the 10% level of significance and conclude that the distribution is symmetric. (7)

(ii) **Test for kurtosis:**

We have to test:

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

H_1 : $\beta_2 \neq 3$.

The critical values are $3.99 + \frac{10}{25}(3.87 - 3.99) = 3.942$ and $2.15 + \frac{10}{25}(2.27 - 2.15) = 2.198$. Reject H_0 if $\beta_2 < 2.198$ or $\beta_2 > 3.942$

Now the value of the test statistic is

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

$$= \frac{\frac{1}{60}(272\,160)}{\left[\frac{1}{60}(2\,160)\right]^2}$$

$$= \frac{4\,536}{[36]^2}$$

$$= \frac{4\,536}{1\,296}$$

$$= 3.5$$

Since $2.198 < 3.5 < 3.942$, we do not reject H_0 at the 10% level of significance and conclude that the distribution does have the kurtosis of a normal distribution. (7)

(iii) Yes, the distribution does originate from a normal distribution since it passed both tests. (1)

(b) We have to test $H_0 : \mu = 60$ against $H_1 : \mu < 60$.

$n = 60 \qquad \sum X = 3\,480 \qquad \sum \left(X_i - \overline{X}\right)^2 = 2\,160$

$\overline{X} = \dfrac{\sum X}{n} = \dfrac{3\,480}{60} = 58.$

$$\begin{aligned} S_X^2 &= \frac{1}{n-1} \sum \left(X_i - \overline{X}\right)^2 \\ &= \frac{1}{60-1} (2\,160) \\ &= \frac{1}{59} (2\,160) \\ &= 36.61016949 \\ \therefore S &= \sqrt{36.61016949} \\ &\approx 6.0506 \end{aligned}$$

$$\begin{aligned} t_{calc} &= \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{S} \\ &= \frac{\sqrt{60}\,(58-60)}{6.0506} \\ &= \frac{\sqrt{60}\,(-2)}{6.0506} \\ &= \frac{-15.49193338}{6.0506} \\ &\approx -2.5604 \end{aligned}$$

Test is one-tailed. $\alpha = 0.05$. The critical value is $t_{\alpha;n-1} = t_{0.01;59} = 2.423 + \dfrac{19}{20}(2.39 - 2.423) \approx 2.392$. Reject H_0 if t_{calc} is less than -2.392.

Since $-2.5604 < -2.392$, we reject H_0 at the 1% level of significance and conclude that $\mu < 60$. (7)

(c) $H_0 : \sigma^2 = 40$ against $H_1 : \sigma^2 < 40$

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

$$\begin{aligned} U &= \frac{\Sigma \left(X_i - \overline{X}\right)^2}{\sigma^2} \\ &= \frac{2\,160}{40} \\ &= 54 \end{aligned}$$

$$\alpha = 0.01$$

$$\begin{aligned} \chi^2_{1-\alpha;\; n-1} &= \chi^2_{0.99;59} \\ &= 29.7067 + \frac{9}{10}(37.4848 - 29.7067) \\ &\approx 36.707 \end{aligned}$$

Reject H_0 if $U < 36.707$.

Since $54 > 36.707$, we do not reject H_0 at the 1% level of significance and conclude that $\sigma^2 < 40$. (6)

[28]

QUESTION 4

(a) Yes, it may be reasonable to assume that the four machines may be considered as independent groups if the breakdown of one machine does not influence the other machines. (2)

(b) 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 Levene's test, p-value $= 0.9594$. Since $0.9594 > 0.05 \Longrightarrow$ we can not reject H_0 at the 5% level of significance. The assumption of equal variances is not violated.

(3)

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

(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 $F = 19.9948$ 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 time among the four machines, that is, $\mu_p \neq \mu_q$ for at least one $p \neq q$.

(4)

(d) No. The $Abs(Dif) - LSD$ for the pair 42 is positive which is 1.1182. The confidence interval is $(1.11816; 3.481841)$ with a p-value of 0.0006 does not include zero. They are not sharing the same letter. We conclude that the means are significantly different from each other. Thus we conclude that $\mu_2 \neq \mu_4$. (3)

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

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

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 include zero imply that the pairs of means do not differ significantly. All pairs include zero except 43 and 32. This is also supported by the fact that the p-values for the differences between the means are 0.0678 and 0.5067 respectively. The p-values are > 0.05, leading to the non rejection of the null hypothesis of equal means. Thus the pair of means are not significantlky from each other, that is, $\mu_3 = \mu_4$ and $\mu_2 = \mu_3$.

The groups that do not have the same letter connecting them means that the groups are significantly different from each other. The pair 43 share the letter A and the pair 32 share the letter B.

Confirming this is the **Abs(Dif)-LSD** of the two pairs which are all negative, that is, -0.0858 (43) and -0.7858 (32) respectively. (Recall a negative value of **Abs(Dif)-LSD** means the groups are not significantly different from each other.)

(4)

[16]

QUESTION 5

(a) (i) ANOVA Test.

We still have to test $H_0: \quad \mu_1 = \mu_2$ against

$H_1: \quad \mu_1 \neq \mu_2.$

(This does not look familiar to the usual $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ of section 7.6 because we have only two groups!)

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

For this data set $k = 2$ and $n = 12$

Method I: Using the critical value approach:

From the output: Computations for ANOVA we see that $F = 18.8571$.

$F_{\alpha;k-1;kn-k} = F_{0.05;1;22} = 4.30$. Reject H_0 if $F \geq 4.30$.

Since $18.8571 > 4.30$, we reject H_0 in favour of H_1 at the 5% level of significance. The two means differ significantly.

Method II: Using the p-value approach

p-value $= 0.0003 << 0.05$. We reject H_0 in favour of H_1 at the 5% level of significance and conclude that the two means differ significantly.

(3)

(ii) The assumptions needed for this test to be valid are:

1) Independent samples
2) Equal population variances, i.e. $\sigma_1^2 = \sigma_2^2$
3) Samples are from normal populations.

(2)

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

$n = 25$ $R = 0.5$

$$\begin{aligned} T &= \sqrt{n-2}\frac{R}{\sqrt{1-R^2}} \\ &= \sqrt{25-2}\frac{0.5}{\sqrt{1-(0.5)^2}} \\ &= \sqrt{23}\frac{0.5}{\sqrt{0.75}} \\ &= \frac{4.795831523 \times 0.5}{0.866025403} \\ &= \frac{2.397915762}{0.866025403} \\ &= 2.768874621 \\ &\approx 2.7689 \end{aligned}$$

$\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$ and $t_{0.025;23} = 2.069$. We reject H_0 if $T < -2.069$ or if $T > 2.069$ or $|T| > 2.069$.

Since $2.7689 > 2.069$, we reject H_0 at the 5% level of significance and conclude that the correlation is significantly different from zero, that is, $\rho \neq 0$.

(7)

(c) H_0 : There is no difference in opinion between psychologists and psychiatrists regarding behavioural modification as a useful technique.

H_1 : Psychiatrists feel less favourable towards the usefulness of behavioural modification

For this 2×2 table for the exact test is

Is behavioural modification useful?				
	Yes	**No**	**Row total**	
Psychologists	5	2	7	
Psychiatrists	1	$4 (= x)$	5	$\leftarrow k$
Column total	6	6	12	
		$\uparrow$ n		

Now $k = 5$, $n = 6$ and $x = 4$ [Note: $k = 5$, $n = 6$ and $x = 1$ thus $P(X \leq 1) = 0.121$]

The alternative "Psychiatrists feel less favourable towards the usefulness of behavioural modification" would imply a large value of x to reject H_0 i.e., $P(X \geq x) = \alpha$.

Now $x = 4$ and

$$\begin{aligned} P\left(X \geq x\right) &= 1 - P(X < x - 1) \\ P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - 0.879 \\ &= 0.121 \ \text{(from table D study guide p131)} \end{aligned}$$

Since $0.121 > 0.05$, we do not reject H_0 at the 5% level of significance and conclude that there is no difference in opinion between psychologists and psychiatrists regarding behavioural modification as a useful technique. (6)

[18]

QUESTION 6

(a) The estimates β_0, β_1 and σ^2 of the model are 4.6979, 1.9705 and 81.4 respectively. (3)

(b) $x = 15$. The predicted sales volume is

$$\begin{aligned} \widehat{Y}_i &= 4.6979 + 1.9705X \\ &= 4.6979 + 1.9705\,(15) \\ &= 4.6979 + 29.5575 \\ &= 34.2554 \end{aligned}$$

Thus the estimated sales volume is $R34\,255\,400$. (1)

(c) From Figure 6 it follows that

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{10}(Y_i - \widehat{Y}_i)^2 = 81.4.$$

we also know that $\widehat{Y} = 34.2554$.

A 95% confidence interval for the expected sales vloume, y for a pharmacy that purchases 15% of its prescription directly from the supplier is given by

$$\widehat{\beta}_0 + \widehat{\beta}_1 X \pm t_{\alpha/2;n-2} \times S\sqrt{\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{\sum(X_i - \overline{X})^2}}$$

where $t_{\alpha/2;n-2} = t_{0.025;8} = 2.306$.

$$\sqrt{\frac{1}{n}+\frac{\left(X-\overline{X}\right)^2}{\sum(X_i-\overline{X})^2}} = \sqrt{\frac{1}{10}+\frac{(15-33.8)^2}{3\,407.6}}$$
$$= \sqrt{0.1+0.103721094}$$
$$= \sqrt{0.203721094}$$
$$= 0.4514$$

So,

$$\begin{aligned} & \widehat{\beta}_0+\widehat{\beta}_1 X && \pm \quad t_{\alpha/2;n-2}\times S\sqrt{\frac{1}{n}+\frac{\left(X-\overline{X}\right)^2}{\sum(X_i-\overline{X})^2}} \\ = \; & 34.2554 && \pm \quad 2.306\times\sqrt{81.4}\,(0.4514) \\ = \; & 34.2554 && \pm \quad 9.3915 \\ = \; & (34.2554-9.3915 && ; \quad 34.2554+9.3915) \\ = \; & (24.8639 && ; \quad 43.6469) \end{aligned}$$

(4)

(d) $H_0:\beta_1=0$ $\qquad$ $H_1:\beta_1>0$

Method I: Using the critical value approach:

From the output:

$$T = \frac{\hat{\beta}_1-B_1}{s/d} = \frac{1.9705-0}{0.154548} \approx 12.75$$

$\alpha=0.05$ $\qquad$ $t_{\alpha;n-2}=t_{0.05;8}=1.86$. Reject H_0 if T is greater than 1.86.

Since $12.75>1.86$, we reject H_0 in favour of H_1 at the 5% level significance and conclude that $\beta_1>0$.

Method II: Using the p-value approach

p-value $<0.0001<<0.05$. We reject H_0 in favour of H_1 at the 5% level of significance and conclude that $\beta_1>0$. (4)

[12]

[100]