Tutorial Letter 204/1/2015

Applied Statistics II

STA2601

Semester 1

Department of Statistics

Trial Examination Paper Solutions

BAR CODE

Dear Student
This is the last tutorial letter for 2015 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.
105	Hints to assignment 1.
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 2014 Final Examination

QUESTION 1

(a) The random variables $X_1, X_2, \ ..., \ X_n$ constitute a *random sample* from the distribution with probability density function (pdf) $f_X(x)$ if $X_1, X_2, \ ..., \ X_n$ are **independent** random variables, each with pdf $f_X(x)$. (1)

(b) The function $T = \sum_{i=1}^{n} (X_i - \mu)^2$ of a random sample $X_1, X_2, \ ..., \ X_n$ is called a statistic if the following condition holds: if μ **is known** (1)

(c) If you fail to reject the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ on the basis of sample data, when in fact there is a difference between the variances of two populations, you have made a **type II** error. (1)

(d) If the exceedance probability $P\left(\overline{X} \geq \overline{x} \mid H_0 \text{ is true}\right) = 0.0001$ and the alternative hypothesis is $H_1 : \mu > \mu_0$, it means that $\overline{x}$ is **highly** significant. (1)

(e) The Bonferroni inequality can be applied to any **simultaneous** inference problem. (1)

[5]

QUESTION 2

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

- $Q(\theta_1, \ ..., \ \theta_k) = \sum_{i=1}^{n} (X_i - E(X_i))^2$
- 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 $\hat{\theta}_1, \ ..., \ \hat{\theta}_k$. (3)

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

The least square estimator is

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

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

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) $f_X(x) = cX^{c-1}$ for $x > 1$.

The maximum likelihood is

$$\begin{aligned} L(c) &= \prod_{i=1}^{n} f(X_i;\ c) \\ &= \prod_{i=1}^{n} cX_i^{c-1} \\ &= cX_1^{c-1} \times cX_2^{c-1} \times \cdots \times cX_n^{c-1} \\ &= c^n \prod_{i=1}^{n} X_i^{c-1} \end{aligned}$$

$$\begin{aligned}\therefore \log L(c) &= n\log c + (c-1)\log\prod_{i=1}^{n} X_i \\ &= n\log c + c\log\prod_{i=1}^{n} X_i - \log\prod_{i=1}^{n} X_i \\ \frac{d\log L(c)}{dc} &= \frac{n}{c} + \log\prod_{i=1}^{n} X_i\end{aligned}$$

Setting $\frac{d\log L(c)}{dc} = 0$

$$\begin{aligned}\therefore \quad \frac{n}{c} + \log\prod_{i=1}^{n} X_i &= 0 \\ \frac{n}{c} &= -\log\prod_{i=1}^{n} X_i \\ n &= -c\log\prod_{i=1}^{n} X_i \\ c\log\prod_{i=1}^{n} X_i &= -n \\ \widehat{c} &= \frac{-n}{\log\prod_{i=1}^{n} X_i}\end{aligned}$$

(6)

[13]

QUESTION 3

(a) From the histogram we conclude that the distribution "looks" fairly symmetrical since the normal curve superimposed covers the histogram symmetrically (this is subjective). The normal quantile plot seem to suggest that there is no systematic deviation from the straight line and the box plot is almost symmetrical. Thus data seems to be normally distributed. (3)

(b) **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 0.587. Reject H_0 if $\beta_1 < -0.587$ or $\beta_1 > 0.587$ or $|\beta_1| > 0.587$

$$\text{Now } \beta_1 = \frac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left(\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}\right)^3} = \frac{\frac{1}{40}(17.82375)}{\left(\sqrt{\frac{1}{40}(250.975)}\right)^3}$$

$$= \frac{0.44559375}{\left(\sqrt{6.274375}\right)^3}$$

$$= \frac{0.44559375}{(2.504870256)^3}$$

$$= \frac{0.44559375}{15.71649531}$$

$$\simeq 0.0284.$$

Since $-0.587 < 0.0284 < 0.587$ we do not reject H_0 at the 10% level of significance. It seems as if the sample comes from a symmetrical distribution.

Test for kurtosis:

We have to test:

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

H_1 : $\beta_2 \neq 3$.

With interpolation we find the interval values from table C page 114 study guide. The critical values are:

$$\text{Lower 5\% critical value} = 0.7440 + \tfrac{4}{5}(0.7470 - 0.736)$$

$$= 0.7440 + 0.8(0.003)$$

$$= 0.7440 + 0.0024$$

$$\approx 0.7464$$

$$\text{Upper 5\% critical value} = 0.8578 + \tfrac{4}{5}(0.8540 - 0.8578)$$

$$= 0.8578 + 0.8(-0.0038)$$

$$= 0.8578 - 0.00304$$

$$\approx 0.8548$$

We reject H_0 if $A < 0.7464$ or $A > 0.8548$

Now the value of the test statistic is

$$\begin{aligned} A &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left|X_i-\overline{X}\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}} \\ &= \frac{\frac{1}{40}(79.1)}{\sqrt{\frac{1}{40}(250.975)}} \\ &= \frac{1.9775}{\sqrt{6.274375}} \\ &= \frac{1.9775}{2.504870256} \\ &\approx 0.7895 \end{aligned}$$

Since $0.7464 < 0.7895 < 0.8548$, we cannot reject H_0 at the 10% level of significance and conclude that the kurtosis of the sample is not significantly different from the kurtosis of a normal distribution. (14)

(c) We have to test $H_0 : \mu = 8$ against $H_1 : \mu \neq 8$

Method 1: Critical value approach

$$\begin{aligned} t_{calc} &= \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \\ &= \frac{6.975-8}{\frac{2.53678}{\sqrt{40}}} \\ &\approx -2.5555 \end{aligned}$$

The critical value is

$$\begin{aligned} t_{\alpha/2;n-1} &= t_{0.025;39} \\ &= 2.03+\frac{4}{5}(2.021-2.03) \\ &= 2.03+0.8(-0.009) \\ &= 2.03-0.0072 \\ &\approx 2.0228 \end{aligned}$$

Reject H_0 if $t_{calc} > 2.0228$ or $t_{calc} < -2.0228$ or if $|t_{calc}| > 2.0228$.

Since $-2.5555 < -2.0228$, the null hypothesis is rejected at the 0.05 level of significance and we conclude that the mean length of time on hold is not 8 minutes, i.e., $\mu \neq 8$.

Method II: p-value approach

Since $p = 0.0146 < 0.05$, we reject H_0 at the 0.05 level of significance and conclude that the true average length of time on hold is not equal to 8 minutes, i.e., $\mu \neq 8$. (4)

(d) The 95% confidence interval for μ is 6.1637 to 7.7863. Yes 8 is not contained in the interval. (2)

[23]

QUESTION 4

(a) (i) $\alpha = 0.05 \qquad \alpha/2 = 0.025 \qquad t_{=\alpha/2;(n_1+n_2-2)} = t_{0.025;60} = 2$

The 95% two-sided confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{array}{lcl} \left(\overline{X}_1 - \overline{X}_2\right) & \pm & t_{\alpha/2;(n_1+n_2-2)} \times S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \\ (50.55 - 56.65) & \pm & 2 \times 14.62\sqrt{\frac{1}{31}+\frac{1}{31}} \\ -6.1 & \pm & 29.24\sqrt{0.064516129} \\ -6.1 & \pm & 7.427 \\ (-6.1 - 7.427 & ; & -6.1 + 7.427) \\ (-13.527 & ; & 1.327) \end{array}$$

(5)

(ii) Looking at the confidence interval, one can conclude that the means are the same since zero is contained in the interval. Thus the two innovative teaching methods yield the same result. (2)

(b) (i) We are testing $H_0 : \mu = 50$ against $H_1 : \mu \neq 50$.

The power of the test is a function of Φ which is defined as $\Phi = \dfrac{\delta}{\sqrt{2}}$

$$\begin{aligned}\delta &= \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\\ &= \frac{\sqrt{13}(50+0.75\,\sigma-50)}{\sigma}\\ &= \pm\sqrt{13}\,(0.75)\\ &= 2.7042\end{aligned}$$

$$\Longrightarrow \Phi = \frac{\delta}{\sqrt{2}} = \frac{2.7042}{\sqrt{2}} \approx 1.9122$$

From table F we read of the power as 69% (i.e., $1-\beta = 0.69$) (4)

(ii) β =P(Type II error) $= 1 - 0.69 = 0.31$ (1)

(c) We have to test:

H_0 : $\pi_1 = 0.20$; $\pi_2 = 0.50$ and $\pi_3 = 0.30$.
H_1 : At least one of the proportions is different from the specified values

	Observed frequency	**Expected frequency**
Bad-tempered	250	$1\,000(0.20) = 200$
Even-tempered	480	$1\,000(0.50) = 500$
Good-tempered	270	$1\,000(0.30) = 300$
Total	**1 000**	**1 000**

The test statistic is:

$$\begin{aligned}Y^2 &= \sum_{i=1}^{3}\frac{(N_i-e_i)^2}{e_i}\\ &= \frac{(250-200)^2}{200}+\frac{(480-500)^2}{500}+\frac{(270-300)^2}{300}\\ &= \frac{(50)^2}{200}+\frac{(-20)^2}{500}+\frac{(-30)^2}{300}\\ &= 12.5+0.8+3.0\\ &= 16.3\end{aligned}$$

$Y^2 \sim \chi^2_{\alpha;k-1}$ (see p.153) and we have $k-1 = 3-1 = 2$. Thus the critical value is $\chi^2_{0.05;2} = 5.99147$. Reject H_0 if $Y^2 \geq 5.99147$.

Since the test statistic $Y^2 = 16.3 > 5.99147$, we reject the null hypothesis at the 5% level. The postulate about the proportions seems to be incorrect. (7)

(d) $H_0 : \rho = 0.5$ against $H_1 : \rho > 0.5$

$n = 52 \qquad r = 0.65$

$$\begin{aligned} U &= \frac{1}{2}\log_e \frac{1+r}{1-R}r \\ &= \frac{1}{2}\log_e \frac{1-0.65}{1+0.65} \\ &= \frac{1}{2}\log_e \frac{1.65}{0.35} \\ &= \frac{1}{2}\log_e 4.714285714 \\ &\approx 0.7753 \end{aligned} \qquad \begin{aligned} \eta &= \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \\ &= \frac{1}{2}\log_e \frac{1-0.50}{1+0.50} \\ &= \frac{1}{2}\log_e \frac{0.5}{1.5} \\ &= \frac{1}{2}\log_e 3 \\ &\approx 0.5493 \end{aligned}$$

Note: You can read the values from Table X Stoker.

The test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{52-3}(0.7753-0.5493) \\ &= \sqrt{49} \times (0.226) \\ &\approx 1.582 \end{aligned}$$

$\alpha = 0.05$ and $Z_{0.05} = 1.645$. Reject H_0 if $Z > 1.645$

Since $1.582 < 1.645$, we do not reject H_0 at the 5% level of significance and conclude that $\rho = 0.5$. (7)

[26]

QUESTION 5

(a) It is reasonable to assume that the four samples are independent. The outcome of one group of immigrants cannot influence the outcome of another group because they are not related and they were randomly selected. (2)

(b) To use JMP for this purpose we should have obtained Normal Quantile Plots or histograms for each separate group. (1)

(c) This assumption can be formally tested:

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

From the tests that the variances are equal (Figure 4) we see that all the p–values $>> 0.05$ which means that we can not reject H_0. The assumption of equal variances is not violated. (2)

(d) We have to test:

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

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

From Figure 3: Graphical representation and computations for ANOVA we see that $F = 4.6107$ which is significant with a p–value $= 0.0177$ and we reject H_0 at the 5% level of significance and conclude that $u_p \neq \mu_q$ for at least one pair $p \neq q$, i.e., at least one pair of mean is significantly different from each other. (4)

[9]

QUESTION 6

(a) Plot of *Y* versus *X*

(4)

(b) $n = 12$ $\quad \Sigma X = 228$ $\quad \Sigma X^2 = 5\,256$
$\Sigma XY = 9\,324$ $\quad \Sigma Y = 540$ $\quad \Sigma Y^2 = 25\,522$
$\overline{X} = 19$ $\quad \overline{Y} = 45$

$$\begin{aligned} b &= \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} \\ &= \frac{12(9\,324) - (228)(540)}{12(5\,256) - (228)^2} \\ &= \frac{111\,888 - 123\,120}{63\,072 - 51\,984} \\ &= \frac{-11\,232}{11\,088} \\ &\approx -1.013 \end{aligned}$$

$$
\begin{aligned}
a &= \overline{Y} - b\overline{X} \\
&= 45 - (-1.013)(19) \\
&= 45 + 19.247 \\
&= 64.247
\end{aligned}
$$

$\therefore \widehat{Y} = 64.247 - 1.013X.$ (5)

(c)

X	Y	$\widehat{Y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 X$	$e_i = \left(Y_i - \widehat{Y_i}\right)^2$
8	59	56.143	8.16
6	58	58.169	0.03
11	56	53.104	8.39
22	53	41.961	121.86
14	50	50.065	0.00
17	45	47.026	4.10
18	43	46.013	9.08
24	42	39.935	4.26
19	39	45	36.00
23	38	40.948	8.69
26	30	37.909	62.55
40	27	23.727	10.71

(8)

(d) $x = 19$. The percentage of wormy fruits is

$$\begin{aligned} \widehat{Y}_i &= 64.247 - 1.013(19) \\ &= 64.247 - 19.247 \\ &= 45 \end{aligned}$$

$\Longrightarrow 45\%$

(2)

(e) $\alpha = 0.01 \qquad t_{\alpha;n-2} = t_{0.01;11} = 2.718$

Reject H_0 if t_{calc} is less than -2.178.

Now $\hat{\beta}_1 = -1.013 \qquad s/d = \dfrac{\sqrt{27.384}}{\sqrt{924}} = 0.172152152$

The test statistic is

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - 0}{s/d} \\ &= \frac{-1.013}{0.172152152} \\ &\approx -5.8843 \end{aligned}$$

Since $-5.8843 < -2.178$, we reject H_0 at the 1% level of significant and conclude that $\beta_1 \neq 0$, i.e., the slope is significant. (5)

[24]

[100]