Tutorial Letter 204/2/2014

Applied Statistics II

STA2601

Semester 2

Department of Statistics

Trial Examination Paper Solutions

BAR CODE

Dear Student
This is the last tutorial letter for 2014. I would like to take this opportunity again of wishing you well in the coming examinations 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	Errata in tutorial letter 101.
104	Installation of SAS JMP 11.
105	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 May/June 2014 Final Examination

QUESTION 1

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

(ii) $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ (study guide page 50) (1)

(b) The steps are:

- Step 1: Find $L(\theta) = \prod_{i=1}^{n} f_X(X_i;\ \sigma)$
- Step 2: Find $Log\ L(\theta)$
- Step 3: Find $\frac{d \log L(\theta)}{d\theta}$, set it to zero and solve for θ.

(4)

[6]

QUESTION 2

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

(5)

(b) So

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

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

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

$$\frac{\partial \ell n L(\lambda)}{\partial \lambda} = -n - \frac{ne^{-\lambda}}{\left(1-e^{-\lambda}\right)} + \frac{\sum_{i=1}^{n} r_i}{\lambda} \qquad (5)$$

[10]

QUESTION 3

(a) If $U = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2}$ then $U \sim \chi^2_{n-1}$ (result 1.2).

Then

$$\begin{aligned}1-\alpha &= P\left(\chi^2_{1-\frac{1}{2}\alpha;n-1} < U < \chi^2_{\frac{1}{2}\alpha;n-1}\right) \\ &= P\left[\chi^2_{1-\frac{1}{2}\alpha;n-1} < \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\sigma^2} < \chi^2_{\frac{1}{2}\alpha;n-1}\right] \\ &= P\left[\frac{1}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \frac{\sigma^2}{\Sigma\left(X_i-\overline{X}\right)^2} < \frac{1}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] \\ &= P\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \sigma^2 < \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]\end{aligned}$$

Thus the $100\,(1-\alpha)\,\%$ two-sided confidence interval for σ^2 is given by

$$\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}};\quad \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] \qquad (5)$$

(b) (i) $n=30 \qquad \overline{X}=41 \qquad s=6 \qquad \alpha=0.10$

Now $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$

$(n-1)s^2=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$

Thus

$$\begin{aligned}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 &= (n-1)s^2\\ &= (30-1)6^2\\ &= 29(36)\\ &= 1\,044\end{aligned}$$

$\chi^2_{\alpha/2;n-1}=\chi^2_{0.05;29}=42.5569$

$\chi^2_{1-\alpha/2;n-1}=\chi^2_{0.95;29}=17.7083$

The 90% confidence interval for σ^2 now becomes

$$\begin{aligned}\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}};\quad \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] &= \left[\frac{1\,044}{42.5569};\quad \frac{1\,044}{17.7083}\right]\\ &= [24.53186205;\quad 58.95540509]\\ &= [24.5319;\quad 58.9554]\end{aligned}$$

(5)

(ii) If 30 is contained in the interval, then the hypothesis $H_0:\sigma^2=30$ will not be rejected. (1)

(c) H_0 : There is no association between hemisphere and site of lesion.

H_1 : There is an association between hemisphere and site of lesion.

For this 2×2 table for the exact test is

		Hemisphere			
		Right	**Left**	**Total**	
Region	**Anterior**	2	5	7	
	Posterior	4	$1^{*}(=x)$		$\longleftarrow \quad k$
	Total	6	6 $\uparrow$ k	12	$\longrightarrow \quad N$

Now $k = 5, n = 6$ and $x = 1$

In this case

$$\begin{aligned} P\left(X \geq x\right) &= 1 - P(X < x - 1) \\ P(X \geq 1) &= 1 - P(X \leq 0) \\ &= 1 - 0.008 \\ &= 0.992 \end{aligned}$$

and $P(X \leq x) = P(X \leq 1) = 0.121.$

We can only reject H_0 in favour of the two-sided alternative if x is too large or too small and if it represents a "rare event", in other words only if

$$P\left(X \leq x\right) \leq \frac{\alpha}{2} \text{ or if } P\left(X \geq x\right) \leq \frac{\alpha}{2}$$

Since the test is two tailed we take the smaller of the two probabilities, i.e., we take 0.121. Since $0.121 > \frac{\alpha}{2} = 0.025$, we do not reject H_0 at the 5% level of significance and conclude that there is no association between hemisphere and lesion.[Note if first column is chosen for n then $n = 6, k = 5$ and $x = 4$

In this case

$$\begin{aligned} P\left(X \geq x\right) &= 1 - P(X < x - 1) \\ P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - 0.879 \\ &= 0.121 \end{aligned}$$

and $P(X \leq x) = P(X \leq 4) = 0.992$. Thus the probability will still be 0.121 and same conclusion holds]. (8)

[19]

QUESTION 4

(a) (i) Let $Y_i =$ score before $-$ score after.

We have to test: $H_0 : \mu_Y = 0$ against $H_1 : \mu_Y < 0$

Method 1: Using the critical value approach

The test statistic is

$$\begin{aligned} T &= \frac{\overline{Y} - \mu}{\frac{S_y}{\sqrt{n}}} \\ &= \frac{-2 - 0}{0.36992} \\ &\approx -5.4066. \end{aligned}$$

The critical value is $t_{\alpha;(n-1)} = t_{0.01;19} = 2.539$. Reject H_0 if $T \leq -2.539$.

Since $-5.5066 < -2.539$, we reject H_0 in favour of H_1 at the 1% level of significance and conclude that $\mu_Y < 0$.

Method II: Using the p-value approach

p-value < 0.0001. Since $0.0001 < 0.05$, we reject H_0 in favour of H_1 at the 1% level of significance and conclude that $\mu_Y < 0$. (6)

(ii) If the same questionnaire is used to test mental activity before and after admission of the drug, there might be a “carry over effect” (i.e., remembering something from the first to the second application). This implies that the true effect of the heroin will be “masked” and might seem less dramatic than it could be when different tests that do not have a “carry over effect” are used. (2)

(b) (i) The chickens will be ready for slaughtering if they weigh more than 2kg on average . This implies one sided testing.

We have to test $H_0 : \mu = 2$ against $H_1 : \mu > 2$. (2)

(ii) The assumption of independent observations and the assumption that masses have a normal distribution (i.e. that the sample comes from a normal population). We may assume that

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

(2)

(iii) **Method 1: Using the critical value approach:**

$$\begin{aligned} T &= \frac{\sqrt{n}\left(\bar{X}-\mu_0\right)}{S} \\ &= \frac{\sqrt{9}\,(1.9667-2)}{0.14361} \\ &\approx -0.6963 \end{aligned}$$

The critical value is $t_{\alpha;n-1} = t_{0.01;8} = 2.896$. Reject H_0 if $T \geq 2.896$.

Since $-0.6963 < 2.896$, we do not reject H_0 at the 1% level of significance. It seems as if the chickens are not ready for slaughtering.

Method II: Using the p-value approach:

p-value $= 0.7470$. Since $0.747 > 0.05$, we cannot reject H_0 at the 1% level of significance. It seems as if the chickens are not ready for slaughtering. (4)

(iv) If we know that $\sigma = 0.10$, we will use the test statistic Z

Method 1: Using the critical value approach:

$$\begin{aligned} Z &= \frac{\sqrt{n}\left(\bar{X}-\mu_0\right)}{\sigma} \sim n\,(0;1) \\ &= \frac{\sqrt{9}\,(1.9667-2)}{0.10} \end{aligned}$$
$$\approx -1$$

The critical value is $Z_\alpha = Z_{0.01} = 2.326$. Reject H_0 if $Z > -2.326$.

Since $-1 < 2.326$, we do not reject H_0. at the 1% level of significance. It seems as if the chickens are not ready for slaughtering.

Method II: Using p-value approach: p-value $= 0.8413$. Since $0.8413 > 0.05$, we do not reject H_0 at the 1% level of significance. It seems as if the chickens are not ready for slaughtering. (4)

[20]

QUESTION 5

(a) The estimates for the variances of the four groups are:

$$\hat{\sigma}_1^2 = S_1^2 = \frac{\sum_{j=1}^{21}\left(X_{1j}-\overline{X}_1\right)^2}{n_1-1} = \frac{5.5500}{20} = 0.2775$$
$$\hat{\sigma}_2^2 = S_2^2 = \frac{\sum_{j=1}^{21}\left(X_{2j}-\overline{X}_2\right)^2}{n_2-1} = \frac{6.8844}{20} = 0.34422$$
$$\hat{\sigma}_3^2 = S_3^2 = \frac{\sum_{j=1}^{21}\left(X_{3j}-\overline{X}_3\right)^2}{n_3-1} = \frac{6.1450}{20} = 0.30725$$
$$\hat{\sigma}_4^2 = S_4^2 = \frac{\sum_{j=1}^{21}\left(X_{4j}-\overline{X}_4\right)^2}{n_4-1} = \frac{8.4718}{20} = 0.42359$$

From the computations above it, follows that $S_1^2 = 0.2775$; $S_2^2 = 0.34422$; $S_3^2 = 0.30725$ and $S_4^2 = 0.42359$. (4)

(b) (i) The ordinary average of the four estimated variance is

$$\begin{aligned}\frac{\hat{\sigma}_1^2+\hat{\sigma}_2^2+\hat{\sigma}_3^2+\hat{\sigma}_4^2}{4} &= \frac{0.2775+0.34422+0.30725+0.42359}{4}\\ &= \frac{1.35256}{4}\\ &= 0.33814\end{aligned}$$

(2)

(ii)

$$\begin{aligned}MSE &= \frac{SSE}{kn-k}\\ &= \frac{\sum_{i=1}^{4}\sum_{j=1}^{21}\left(X_{ij}-\overline{X}_i\right)^2}{kn-k}\\ &= \frac{5.5500+6.8844+6.1450+8.4718}{80}\\ &= \frac{27.0512}{80}\\ &= 0.33814\end{aligned}$$

This is the same value as the average of the four estimated variances computed in (a). (4)

(c) Yes, the one group cannot influence the other groups. (2)

(d) We have to test:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against

$H_1 : \mu_p \neq \mu_q$ for at least one $p \neq q$.

We have $k = 4$ random samples of size $n = 21$ each.

	$\sum_{j=1}^{21} X_{ij}$	$\overline{X}_i$	$\left(\overline{X}_i - \overline{X}\right)^2$
$i = 1$	75.60	3.6	$(3.6 - 3.875)^2 = 0.075625$
$i = 1$	65.10	3.1	$(3.1 - 3.875)^2 = 0.600625$
$i = 1$	96.60	4.6	$(4.6 - 3.875)^2 = 0.525625$
$i = 1$	88.20	4.2	$(4.2 - 3.875)^2 = 0.105625$
	325.5	$\overline{\overline{X}} = 3.875$	1.3075

Now

$$\begin{aligned} SST_r &= n \sum_{i=1}^{k} \left(X_i - \overline{\overline{X}} \right)^2 \\ &= 21(1.3075) \\ &= 27.4575 \end{aligned}$$

From question (b) above we already know that MSE = 0.33814

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

$$\begin{aligned} MST_r = \frac{SST_r}{k-1} &= \frac{27.4575}{3} \\ &= 9.1525 \end{aligned}$$

$$\begin{aligned} F = \frac{MST_r}{MSE} &= \frac{n \sum_{i=1}^{k} \left(X_i - \overline{\overline{X}} \right)^2 /(k-1)}{\sum_{i=1}^{4} \sum_{j=1}^{21} \left(X_{ij} - \overline{X}_i \right)^2 /(kn-k)} \\ &= \frac{9.1525}{0.33814} \\ &\approx 27.0672 \end{aligned}$$

The critical value is $F_{\alpha;k-1;kn-1} = F_{0,05;3;80} = 2.72$. Reject H_0 if $F > 2.72$

Since $7.0672 > 2.72$, H_0 is rejected at the 5% level of significance. There is a significant difference in the mean cholesterol score of the four different groups. (11)

[23]

QUESTION 6

(a) The sample correlation coefficient is

$$\begin{aligned} R &= \frac{\Sigma X_i Y_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{n}}{\sqrt{\left(\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right)\left(\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}\right)}} \\ &= \frac{\Sigma\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\left[\Sigma\left(X_i - \overline{X}\right)^2\right]\left[\Sigma\left(Y_i - \overline{Y}\right)^2\right]}} \\ &= \frac{1\,141}{\sqrt{[12.9]\,[102\,940]}} \\ &= \frac{1\,141}{\sqrt{1\,327\,926}} \\ &= \frac{1\,141}{1\,152.356716} \\ &\approx 0.9901 \end{aligned}$$

(3)

(b) $H_0 : \rho = 0.8$ against $H_1 : \rho > 0.8$

$$\begin{aligned} U &= \tfrac{1}{2}\log_e \frac{1+R}{1-R} & \eta &= \tfrac{1}{2}\log_e \frac{1+\rho}{1-\rho} \\ &= \tfrac{1}{2}\log_e \frac{1+0.9901}{1-0.9901} & &= \tfrac{1}{2}\log_e \frac{1+0.8}{1-0.8} \\ &= \tfrac{1}{2}\log_e \frac{1.9901}{0.0099} & &= \tfrac{1}{2}\log_e \frac{1.8}{0.2} \\ &= \tfrac{1}{2}\log_e 201.020202 & &= \tfrac{1}{2}\log_e 9 \\ &\approx 2.6517 & &\approx 1.0986 \end{aligned}$$

Note: Instead of using the above formula you can read these values or interpolate them from table X (when $r = 0.8$ $\eta = 1.0986$). Now, the test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{10-3}(2.6517-1.0986) \\ &= \sqrt{7} \times 1.5531 \\ &\approx 4.1091. \end{aligned}$$

$\alpha = 0.01$, and $Z_\alpha = Z_{0.01} = 2.326$. Reject H_0 if $Z > 2.326$

Since $4.1091 > 2.326$, we reject H_0 at the 1% level of significance and conclude that $\rho > 0.8$. (6)

(c) No, pairs of observations (e.g. age and cost) were taken on a number of computer equipment. The equipment were selected at random and the two variables measured on each equipment. (2)

(d)

$$\begin{aligned} \text{Hence, } \hat{\beta}_1 &= \frac{\sum Y_i (X_i - \overline{X})}{\sum (X_i - \overline{X})^2} \\ &= \frac{1\,141}{12.9} \\ &\approx 88.4496. \end{aligned}$$

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= \frac{2\,560}{10} - 88.4496\left(\frac{24}{10}\right) \\ &= 256 - 88.4496(2.4) \\ &= 256 - 212.27904 \\ &= 43.72096 \\ &\approx 43.721 \end{aligned}$$

$\widehat{\text{Cost}} = 43.7210 + 88.4496\text{Age}$ (4)

(e) $x = 5$. The predicted cost is

$$\begin{aligned}\widehat{Y}_i &= 43.721 + 88.4496X \\ &= 43.721 + 88.4496\,(5) \\ &= 43.721 + 442.248 \\ &= 485.969\end{aligned}$$

(1)

(f) The confidence interval for the predicted cost is

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

$\widehat{\beta}_0 + \widehat{\beta}_1 = 485.969$ $\qquad$ $t_{\alpha/2;n-2} = t_{0.025;8} = 2.306$

$S = \sqrt{252.37375} \approx 15.8863$ $\qquad$ $d^2 = \sum\left(X_i - \overline{X}\right)^2 = 12.9$

$X = 5$ $\qquad$ and $\qquad$ $\overline{X} = 2.4$

Thus, the 95% confidence interval for the predicted cost for equipment that is 5 years old is

$$\begin{array}{lcl}\widehat{\beta}_1 + \widehat{\beta}_0 & \pm & t_{\alpha/2;n-2} \times S\sqrt{1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}} \\ 485.969 & \pm & 2.306 \times 5.8863\sqrt{1 + \frac{1}{10} + \frac{(5-2.4)^2}{12.9}} \\ 485.969 & \pm & 36.6338078 \times \sqrt{1.1 + 0.54031007} \\ 485.969 & \pm & 36.6338078 \times \sqrt{1.624031007} \\ 485.969 & \pm & 46.6852 \\ (485.969 - 46.6852 & ; & 485.969 + 46.6852) \\ (439.2838 & ; & 532.6542)\end{array}$$

(6)

[22]

[100]