Tutorial letter 204/1/2018

Applied Statistics II

STA2601

Semester 1

Department of Statistics

Solutions to Trial Examination

Dear Student
This is the last tutorial letter for 2018 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 or 14.
104	Errata to tutorial letter 101
105	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 Oct/Nov 2017 Final Examination

QUESTION 1

(a) True. The variance of any distribution is the sum of the squares of the deviation. (1)

(b) False. A type I error is committed when H_0 is rejected when in actual fact it is true. (1)

(c) False. Correlation does not mean causation, i.e., correlation does not imply causality. Thus, one can not draw cause and effect conclusions based on correlation. (2)

(d) False. The assumptions underlying a one-way analysis of variance are

- observations are independent (given),
- data comes from a normal population, and.
- equal population variances.

(2)

[6]

QUESTION 2

(a) $X_i \sim n(\mu; \sigma^2)$ for $i = 1;\ 2$ and 3.

$E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for $i = 1;\ 2$ and 3

$$\begin{aligned} E(T_1) &= E\left[\frac{X_1 + X_2 + X_3}{3}\right] \\ &= \tfrac{1}{3}[E(X_1) + E(X_2) + E(X_3)] \\ &= \tfrac{1}{3}[\mu + \mu + \mu] \\ &= \tfrac{1}{3}[3\mu] \\ &= \mu \end{aligned}$$

$$
\begin{aligned}
E(T_2) &= E\left[\frac{X_1 + 2X_2 + 2X_3}{5}\right] \\
&= \tfrac{1}{5}\left[E(X_1) + 2E(X_2) + 2E(X_3)\right] \\
&= \tfrac{1}{5}[\mu + 2\mu + 2\mu] \\
&= \tfrac{1}{5}[5\mu] \\
&= \mu
\end{aligned}
$$

T_1 and T_2 are both unbiased estimators of μ. (6)

(b) (i) $f_W(w_i;\ \theta) = \dfrac{1}{\theta}e^{-w_i/\theta}$ for $w \geq 0$ and $\dfrac{1}{\theta} > 0$

The maximum likelihood estimator is

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f_W(w_i;\ \theta) \quad \text{(see definition 2.5)} \\
&= \prod_{i=1}^{n} \frac{1}{\theta}e^{-w_i/\theta} \\
&= \frac{1}{\theta}e^{-w_1/\theta} \times \frac{1}{\theta}e^{-w_2/\theta} \times \ldots \times \frac{1}{\theta}e^{-w_n/\theta} \\
&= \frac{1}{\theta^n}e^{-\sum w_i/\theta} \\
&= \theta^{-n}e^{-\sum w_i/\theta}
\end{aligned}
$$

$$
\begin{aligned}
\therefore \ln L(\theta) &= -n\ln\theta - \frac{\sum w_i}{\theta} \\
\Longrightarrow \frac{\partial \ln L(\theta)}{\partial\theta} &= \frac{-n}{\theta} - \frac{\sum w_i}{\theta^2} \times -1 \\
&= \frac{-n}{\theta} + \frac{\sum w_i}{\theta^2}
\end{aligned}
$$

Setting $\frac{\partial \ln L(\theta)}{\partial \theta} = 0$ we get

$$\begin{aligned} \frac{-n}{\theta} + \frac{\sum w_i}{\theta^2} &= 0 \\ \frac{\sum w_i}{\theta^2} &= \frac{n}{\theta} \\ \frac{\sum w_i}{n} &= \widehat{\theta} \\ \Longrightarrow \quad \widehat{\theta} &= \overline{W} \text{ (the maximum likelihood estimator (m.l.e.) of } \theta) \end{aligned}$$

$\therefore \widehat{\theta} = \frac{\sum_{i=1}^{n} W_i}{n} = \overline{W}$ is the maximum likelihood estimator of θ. (7)

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

$$\begin{aligned} E(\widehat{\theta}) &= E(\overline{W}) \\ &= E\left(\frac{1}{n}\sum W_i\right) \\ &= \frac{1}{n}\sum E(W_i) \\ &= \frac{1}{n}\sum \theta \\ &= \frac{1}{n} \times n\theta \\ &= \theta \text{ (q.e.d.)} \end{aligned}$$

(3)

[16]

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.)

With interpolation we find the critical value (from table A page 110 study guide) to be

$$\begin{aligned} \text{Critical value} &= 0.711 + \frac{26-25}{30-25}(0.662 - 0.711) \\ &= 0.711 + \frac{1}{5}(-0.049) \\ &= 0.711 + (-0.0098) \\ &\approx 0.701 \end{aligned}$$

Reject H_0 if $\beta_1 < -0.701$ or $\beta_1 > 0.701$ or $|\beta_1| > 0.701$

$$\text{Now } \beta_1 = \frac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left(\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}\right)^3} = \frac{\frac{1}{26}(-53.1893)}{\left(\sqrt{\frac{1}{26}(89.0622)}\right)^3}$$

$$= \frac{-2.045742308}{\left(\sqrt{3.425469231}\right)^3}$$

$$= \frac{-2.045742308}{(1.850802321)^3}$$

$$= \frac{-2.045742308}{6.339866402}$$

$$\approx -0.3227$$

Since $-0.701 < -0.3227 < 0.701$ we do not reject H_0 at the 10% level of significance level and conclude that this distribution is symmetric.

(7)

(b) (i) The normal quantile plot shows that the points at the left end and in the middle are not following the diagonal. They seem to slightly deviate from the line. The box plot shows a longer tail to the left suggesting that data is negatively skewed. However, the normal curve on the histogram seems to fit the data well (Its subjective). We need proper test.

(3)

(ii) We have to test $H_0 : \mu = 17.5$ against $H_1 : \mu \neq 17.5$.

From the output $\overline{X} = 15$ and $s = 1.88746$.

Method 1: Using the critical value approach

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{s} = \frac{\sqrt{26}\,(15 - 17.5)}{1.88746} \approx -6.7538$$

The critical value is $t_{\alpha/2;n-1} = t_{0.025;25} = 2.06$. We will reject H_0 if $T \leq -2.06$, or if $T > 2.06$ or if $|T| > 2.06$.

Since $-6.7538 < -2.06$ we reject H_0 at the 5% level of significance and conclude that $\mu \neq 17.5$, i.e., the mean is significantly different from 17.5.

Method II: Using the p-value approach

p-value < 0.0001. Since $0.0001 < 0.05$, we reject H_0 at the 5% level of significance and conclude that $\mu \neq 17.5$, i.e., the mean is significantly different from 17.5.

(5)

(iii) From the output, the 95% confidence interval for μ is 14.2376 to 15.7624.

(2)

(iv) Yes. The interval supports the conclusion in part b(ii). Since the 95% confidence interval is the same as testing a two sided test at the 5% level. Now we are 95% confident that $14.2376 \leq \mu \leq 15.7624$. The two tailed 5% test can be compared to a 95% confidence interval. In this case the value 17.5 does not lie in the interval and thus we reject H_0 at the 5% level of significance and conclude that $\mu \neq 17.5$, i.e., the mean is significantly different from 17.5.

(2)

(iv) We have to test $H_0 : \sigma^2 = 9$

against $H_1 : \sigma^2 \neq 9$

Method 1: Using the critical value approach

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

$$U = \frac{(n-1)\,s^2}{\sigma^2} = \frac{25\,(1.88746)^2}{9} \approx 9.8958$$

The critical values are $\chi^2_{1-\alpha/2;n-1} = \chi^2_{0.975;25} = 13.1197$ and $\chi^2_{\alpha/2;n-1} = \chi^2_{0.025;25} = 40.6465$

Reject H_0 if $U < 13.1197$or $U > 40.6465$

Since $9.8958 < 13.1197$, we reject H_0 at the 5% level of significance and conclude that $\sigma^2 \neq 9$.

Method II: Using the p-value approach

p-value $= 0.0062$. Since $0.0062 < 0.05$, we reject H_0 at the 5% level of significance and conclude that $\sigma^2 \neq 9$.

(4)

(vi) The assumption made was that the mean μ is unknown and hence the test statistic $U = \dfrac{\Sigma \left(X_i - \overline{X}\right)^2}{\sigma^2}$ was used at χ^2_{n-1}. (2)

[25]

QUESTION 4

(a) The sample correlation coefficient r is

$$\begin{aligned} r &= \frac{\Sigma \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\Sigma \left(x_i - \overline{x}\right)^2 \Sigma \left(y_i - \overline{y}\right)^2}} \\ &= \frac{1\,141}{\sqrt{(12.9)\,(102\,940)}} \\ &= \frac{1\,141}{\sqrt{1\,327\,926}} \\ &= \frac{1\,141}{1\,152.356716} \\ &\approx 0.9901 \end{aligned}$$

(3)

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

$n = 10$ $R = 0.99$

$$\begin{aligned} U &= \frac{1}{2}\log_e \frac{1+r}{1-r} & \eta &= \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \\ &= \frac{1}{2}\log_e \frac{1+0.99}{1-0.99} & &= \frac{1}{2}\log_e \frac{1+0.8}{1-0.8} \\ &= \frac{1}{2}\log_e \frac{1.99}{0.01} & &= \frac{1}{2}\log_e \frac{1.8}{0.2} \\ &= \frac{1}{2}\log_e 199 & &= \frac{1}{2}\log_e 9 \\ &\approx 2.6467 & &\approx 1.0986 \end{aligned}$$

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

The test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{10-3}(2.6467-1.0986) \\ &= \sqrt{7} \times 1.5481 \\ &\approx 4.0959 \end{aligned}$$

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

Since $4.0959 > 2.326$, we reject H_0 at the 1% level of significance and conclude that $\rho > 0.8$, that is, the correlation is more than 0.8.

(7)

(c) Consider the simple linear regression $\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X$

Then

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\Sigma\,(x_i - \overline{x})\,(y_i - \overline{y})}{\Sigma\,(x_i - \overline{x})^2} \\
&= \frac{\Sigma\,(x_i - \overline{x})\,(y_i - \overline{y})}{d^2} \\
&= \frac{1\,141}{12.9} \\
&\approx 88.4496
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\
&= \frac{2\,560}{10} - 88.4496\left(\frac{24}{10}\right) \\
&= 256 - 88.4496\,(2.4) \\
&= 256 - 212.27904 \\
&= 43.7210
\end{aligned}$$

The estimated regression equation is $\widehat{Cost} = 43.7210 + 88.4496 Age$.

(7)

(d) $x_i = 3.5$

The expected cost is

$$\begin{aligned}
\widehat{Cost} &= 43.7210 + 88.4496 Age \\
&= 43.7210 + 88.4496\,(3.5) \\
&= 43.7210 + 309.5736 \\
&= 353.2946 \\
&\approx R353.29
\end{aligned}$$

(1)

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

$\widehat{\beta}_0 + \widehat{\beta}_1 X = 353.2946 \qquad t_{\alpha/2;n-2} = t_{0.05;8} = 1.86$

$d^2 = \Sigma\left(x_i - \overline{x}\right)^2 = 12.9 \qquad MSE = s^2 = 252 \Longrightarrow s = \sqrt{252} = 15.8745$

Now

$$\begin{aligned} SE &= S\sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{d^2}} \\ &= 15.8745\sqrt{1 + \frac{1}{10} + \frac{(3.5 - 2.4)^2}{12.9}} \\ &= 15.8745\sqrt{1 + 0.1 + 0.093798449} \\ &= 15.8745\sqrt{1.19379845} \\ &\approx 17.3447 \end{aligned}$$

The 90% confidence interval for the expected cost for a machine that is 3.5 years old is

$$\begin{aligned} \widehat{\beta}_0 + \widehat{\beta}_1 X & \pm t_{\alpha/2;n-2} \times S\sqrt{1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{d^2}} \\ 353.2946 & \pm 1.86 \times 17.3447 \\ 353.2946 & \pm 32.2611 \\ (353.2946 - 32.2611) & ; \quad 353.2946 + 32.2611) \\ (321.0335 & ; \quad 385.5557) \end{aligned}$$

(4)

(f) The X-values used in the construction of the regression line are 0.75 to 4.25. In this case, estimates will be outside the range of X-values used in the construction of the regression line. The limits might become unreliable as the relationship between X and Y outside this range is not known and may be different from the one found in the specified range.

(2)

[24]

QUESTION 5

(a) $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$

$n_1 = 20 \quad \overline{X}_1 = 57.4 \quad S_1 = 8.124$

$n_2 = 25 \quad \overline{X}_2 = 63.4 \quad S_2 = 7.874$

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

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{(20 - 1)8.124^2 + (25 - 1)7.874^2}{20 + 25 - 2} \\ &= \frac{19\,(65.999376) + 24\,(61.999876)}{43} \\ &= \frac{1\,253.988144 + 1\,487.997024}{43} \\ &= \frac{2\,741.985168}{43} \\ &= \approx 63.7671 \\ \Longrightarrow \quad & S_{pooled} = \sqrt{63.7671} \approx 7.9854 \end{aligned}$$

The test statistic is

$$\begin{aligned} T &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(57.4 - 63.4) - (0)}{7.9854\sqrt{\frac{1}{20} + \frac{1}{25}}} \\ &= \frac{-6}{7.9854\sqrt{0.09}} \\ &= \frac{-6}{2.39562} \\ &\approx -2.5046 \end{aligned}$$

Test is one tailed. The critical value is $t_{\alpha;(n_1+n_2-2)} = t_{0.05;43}$

Interpolating $t_{0.05;40} = 1.684$ and $t_{0.05;60} = 1.671$.

$$\begin{aligned} t_{0.05;43.} &= 1.684 + \frac{3}{20}(1.671 - 1.684) \\ &= 1.684 + \frac{3}{20}(-0.013) \\ &= 1.684 - 0.00195 \\ &\approx 1.682 \end{aligned}$$

Reject H_0 if $T < -1.682$.

Since $-2.5046 < -1.682$, we reject H_0 at the 5% level and conclude that $\mu_1 < \mu_2$, that is, women are on average socially more skillful than men.

In order to perform the tests we assumed that:

- the observations in each sample are independent and also the two samples are mutually independent.
- the observations are normally distributed.
- the two population variances are equal.

(11)

(b) (i) If $\mu = 4$, a 90% confidence interval for σ^2 is

$$\left[\frac{\Sigma (X_i - \mu)^2}{\chi^2_{\frac{1}{2}\alpha;n}} < \sigma^2 < \frac{\Sigma (X_i - \mu)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}} \right]$$

Then

$$\sum_{i=1}^{n} X_i = 60; \qquad \sum_{i=1}^{10} X_i^2 = 380; \qquad \mu = 4$$

$$\begin{aligned} \Sigma (X_i - \mu)^2 &= \Sigma X_i^2 - 2\mu \Sigma X_i + n\mu^2 \\ &= 380 - 2(4)(60) + 20(4)^2 \\ &= 380 - 480 + 320 \\ &= 220 \end{aligned}$$

$$\begin{aligned} \chi^2_{\frac{1}{2}\alpha;n} &= \chi^2_{0.05;20} = 10.8508 \\ \chi^2_{1-\frac{1}{2}\alpha;n} &= \chi^2_{0.95;20} = 30.4104 \end{aligned}$$

Thus, the 95% one-sided confidence interval for σ is

$$\left[\frac{\Sigma\,(X_i-\mu)^2}{\chi^2_{\frac{1}{2}\alpha;n}} < \sigma^2 < \frac{\Sigma\,(X_i-\mu)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}}\right]$$

$$\left[\frac{220}{31.4104} < \sigma^2 < \frac{220}{10.8508}\right]$$

$$\left[7.004 < \sigma^2 < 20.275\right]$$

(5)

(ii) Since 9 is contained in the interval, then the null hypothesis, $H_0 : \sigma^2 = 9$ will not be rejected. (1)

[17]

QUESTION 6

(a) We have to test:

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

Using the Bartlett's test, p-value $= 0.8810$. Since $0.8810 > 0.05 \Longrightarrow$ we can not reject H_0 at the 5% level of significance. The three groups have equal population variances. Thus, the assumption of equal variances is not violated.

(3)

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

(iii) From the output: Computations for ANOVA we see that $F = 9.1304$ which is significant with a p–value of 0.0039. Since $0.0039 < 0.05$, we reject H_0 in favour of H_1 at the 5% level of significance and conclude that $\mu_p \neq \mu_q$ for at least one $p \neq q$, that is, there are significant differences in typing performance among the three keyboard designs.

(4)

(c) Keyboard design C and keyboard design B share the same letter A and are not significantly different from each other.

The keyboard design which are significantly different from each other have **Abs(Dif)-HSDs** that are positive. The pairs are AB and AC which are 0.6966 and 1.6966 respectively. Since they are positive, the means are significantly different. (Recall a negative value of **Abs(Dif)-HSD** means the groups are not significantly different from each other.)

Confidence intervals that do not include zero imply that the pairs of means differ significantly. The pairs AB and AC do not include zero. The confidence interval for the pairs are $(0.6966; 7.3034)$ and $(1.6966; 8.3034)$. These are the only intervals that do not include zero and it means we reject the null hypothesis of equal means and conclude that $\mu_A \neq \mu_B$ and $\mu_A \neq \mu_C$. The p-values are 0.0183 and 0.0043 respectively which are less than 0.05 and thus leading to the rejection of the null hypothesis of equal means.

(5)

[12]

[**100**]