Tutorial letter 204/1/2017

Applied Statistics II

STA2601

Semester 1

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 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 2016 Final Examination

QUESTION 1

(a) True. According to *definition 1.16* of the study guide, X is a continuous variable with $E(X) = 100$ and $Var(X) = 64$.

(b) True. The normal density function **is symmetric about the mean** μ and $P(X \leq \mu) = \int_{-\infty}^{\mu} f_X(x)\,dx = 0.5$. If $\mu = 73$ the rest follows as

$$\int_{-\infty}^{73} f_X(x)\,dx = 0.5 = P(X \leq 73).$$

(c) False. The most efficient several unbiased estimators is the one with the smallest variance.

(d) True.

(e) False. The power of a test is found by computing $1 - P$ (type II error).

(f) False, if X and Y do not have a bivariate normal distribution, then $\rho = 0$ does not necessarily imply independence, *study guide page 135*.

(g) True.

[10]

QUESTION 2

(a) $f_X x = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(X-\mu)^2/\sigma^2}$ $\qquad -\infty < X < \infty.$

Now $f_X x = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}X^2/\sigma^2}$ since $\mu = 0$.

The maximum likelihood is

$$\begin{aligned} L\left(\sigma^2\right) &= \prod_{i=1}^{n} f\left(X_i;\sigma^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}X_i^2/\sigma^2} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}X_1^2/\sigma^2} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}X_2^2/\sigma^2} \times \cdots \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}X_n^2/\sigma^2} \\ &= \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum X_i^2/\sigma^2} \end{aligned}$$

$$\begin{aligned} \log L\left(\sigma^2\right) &= -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum X_i^2/\sigma^2 \\ \frac{d\log L\left(\lambda\right)}{d\lambda} &= -\frac{n}{\sigma} - 0 + \sum X_i^2/\sigma^3 \end{aligned}$$

Setting $\frac{d\log L\left(\lambda\right)}{d\lambda} = 0$

$$\begin{aligned} -\frac{n}{\sigma} - 0 + \sum X_i^2/\sigma^3 &= 0 \\ \sum X_i^2/\sigma^3 &= \frac{n}{\sigma} \\ n\sigma^3 &= \sigma\sum X_i^2 \\ n\sigma^2 &= \sum X_i^2 \\ \sigma^2 &= \frac{\sum X_i^2}{n} \\ \Rightarrow \quad \widehat{\sigma}^2 &= \frac{\sum X_i^2}{n} \end{aligned}$$

(8)

(b)

$$\begin{aligned} T &= \frac{\sum X_i^2}{n} \\ E(T) &= E\left(\frac{\sum X_i^2}{n}\right) \\ &= \frac{1}{n}\sum E\left(X_i^2\right) \end{aligned}$$

Now $E(X_i) = 0$ and $Var(X_i) = \sigma^2$

$$\begin{aligned} Var(X_i) &= E\left(X_i^2\right) - (E(X_i))^2 \\ \sigma^2 &= E\left(X_i^2\right) - 0^2 \\ \sigma^2 &= E\left(X_i^2\right) \end{aligned}$$

Thus,

$$\begin{aligned} E(T) &= \frac{1}{n}\sum E\left(X_i^2\right) \\ &= \frac{1}{n}\sum \sigma^2 \\ &= \frac{1}{n} n\sigma^2 \\ &= \sigma^2 \end{aligned}$$

(4)

(c) If $U = \dfrac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2}$ 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

$$\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}};\quad \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right] \qquad (4)$$

[16]

QUESTION 3

(a) (i) We are testing $H_0: \mu = 50$ against $H_1: \mu \neq 50$ and we assume that $\mu_0 = 50 + 0.75\sigma$

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

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

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

From table F:

For $n = 13, \quad \upsilon = 12, \quad \Phi = 1.9122 = 1.9$ the power is 0.69 at the 5% level of significance.

(4)

(ii) Let the probability of a Type II error $= \beta$.

$\beta = 1-$ power $= 1 - 0.69 = 0.31$ (for a Type I $= \alpha = 0.05$).

(1)

(b) H_0 : There is no association between type of personality and the type of car.

H_1 : Introverts avoid sports models.

For this 2×2 table for the exact test is

Type of car	**Type of personality**		
	Introvert	**Extrovert**	**Total**
Sport models	1^*	4	5 $\longleftarrow$ k
Sedan	4	3	7
Total	5	7	12 $\longrightarrow$ N
	$\uparrow$		
	n		

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

The alternative "introverts avoid sports model would imply a small value of x to reject H_0 i.e., $P(X \leq x) = \alpha$.

Now $x = 1$ and

$$\begin{aligned} P\,(X \leq x) &= P(X \leq 1) \\ &= 0.247 \quad \text{(from table D study guide p131)} \end{aligned}$$

Since $0.247 > 0.05$, we do not reject H_0 at the 5% level of significance and conclude that there is no association between type of personality and type of car.

(8)

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

$n = 36 \quad r = 0.4$

Method I: Using the critical value approach

$$\begin{aligned} T &= \sqrt{n-2}\frac{R}{\sqrt{1-R^2}} \\ &= \sqrt{36-2}\frac{0.4}{\sqrt{1-(0.4)^2}} \\ &= \sqrt{34}\frac{0.4}{\sqrt{0.84}} \\ &= \frac{5.830951895 \times 0.4}{0.916515139} \\ &= \frac{2.332380758}{0.916515139} \\ &= 2.544836041 \\ &\approx 2.5448 \end{aligned}$$

$\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$ and the critical value is

$$\begin{aligned} t_{\alpha/2;n-2} &= t_{0.025;34} \\ &= 2.042 + \frac{4}{5}(2.030 - 2.042) \\ &= 2.042 + \frac{4}{5}(-0.012) \\ &= 2.042 - 0.0096 \\ &\approx 2.032 \end{aligned}$$

We reject H_0 if $T < -2.032$ or if $T > 2.032$ or $|T| > 2.032$.

Since $2.5448 > 2.032$, 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)

(d) (i) The confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2;(n_1+n_2-2)} \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{(400 - 1)257^2 + (450 - 1)251^2}{400 + 450 - 2} \\ &= \frac{399\,(66\,049) + 449\,(63\,001)}{848} \\ &= \frac{26\,353\,551 + 28\,287\,449}{848} \\ &= \frac{54\,641\,000}{848} \\ &\approx 64\,435.14151 \\ \Longrightarrow\ & S_{pooled} = \sqrt{64\,435.14151} \approx 253.8408 \end{aligned}$$

$t_{\alpha/2;(n_1+n_2-2)} = t_{0.005;848} = 2.576$

So the 99% confidence interval is is

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2;(n_1+n_2-2)} \times S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(1\,252 - 1\,330) \pm 2.576 \times 253.8408\sqrt{\frac{1}{400} + \frac{1}{450}}$$

$$-78 \pm 653.8939008\sqrt{0.004722222}$$

$$-78 \pm 44.9346$$

$$(-78 - 44.9346; -78 + 44.9346)$$

$$(-122.9346; -33.0654)$$

(10)

(ii) Since the confidence interval does not include the value 0, we can reject $H_0 : \mu_1 - \mu_2 = 0$ in favour of $H_1 : \mu_1 - \mu_2 \neq 0$ (two-sided) and conclude that the mean usage per household has changed between the two years.

(2)

[32]

QUESTION 4

(a) 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 time elapsed 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 smokers were drawn randomly, thus the assumption of **independent observations is met.**

The normality assumption is violated because from the JMP graphical output we see that the normal curve does not fit the histogram very well and there also seem to be a systematic deviation around the line in the Normal Quantile Plot. The histogram show that data is

negatively skewed and this is also supported by the boxplot with a longer tail to the left (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 $= 0.3686$), indicating that we may assume the data does come from a normal distribution.

(4)

(b) We have to test $H_0 : \mu = 45$ against $H_1 : \mu \neq 45$.

Method 1: Using the critical value approach:

$$\begin{aligned} T &= \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{S} \\ &= \frac{\sqrt{20}\,(59.3 - 45)}{9.83602} \\ &\approx 6.5018 \end{aligned}$$

The critical value is $t_{\alpha/2;n-1} = t_{0.025;19} = 2.093$

We will reject H_0 if $T \geq 2.093$ or $T \leq -2.093$ or if $|T| \geq 2.093$.

Since $6.5018 > 2.093$, we reject H_0 at the 5% level of significance and conclude that the mean is significantly different from 45.

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 the mean is significantly different from 45.

(4)

(c) If we know that $\sigma = 10$, we will use the test statistic Z

Method 1: Using the critical value approach:

$$\begin{aligned} Z &= \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{\sigma} \sim n\,(0;\,1) \\ &= \frac{\sqrt{20}\,(59.3 - 45)}{10} \\ &\approx 6.3952 \end{aligned}$$

The critical value is $Z_{\alpha/2} = Z_{0.025} = 1.96$

We will reject H_0 if $Z \geq 1.96$ or $Z \leq -1.96$ or if $|Z| \geq 1.96$.

Since $6.3952 > 1.96$, we reject H_0. at the 5% level of significance and conclude that the mean is significantly different from 45.

Method II: Using 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 the mean significantly different from 45.

(4)

(d) We are 95% confident that the true value of μ lies between 54.6966 and 63.9034, 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 54.6966 to 63.9034.

Since the 95% confidence interval is the same as testing a two sided test at the 5% level. Now we are 95% confident that $54.6366 \leq \mu \leq 63.9034$. The two tailed 5% test can be compared to a 95% confidence interval.

(4)

[16]

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 Levene's test, p-value $= 0.9765$. Since $0.9765 > 0.05 \Longrightarrow$ 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 $F = 21.0922$ 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 treatments, that is, $\mu_p \neq \mu_q$ for at least one $p \neq q$. The mean REMs sleep time depends on the concentration of ethanol.

(4)

(c) Confidence intervals that do not include zero imply that the pairs of means are significantly different from each other. All pairs do not include zero except the pairs $(2g/kg - 4g/kg)$ and $(1g/kg - 2g/kg)$. The confidence interval for the pairs are $(-2.2864$ to $32.6064)$; and $(-3.8264$ to $31.0664)$; respectively. These are the only intervals that include zero and it means we do not reject the null hypothesis and conclude that $\mu_{2g/kg} = \mu_{4g/kg}$ and $\mu_{1g/kg} = \mu_{2g/kg}$. This is also supported by the fact that the p-value for the differences between the means are 0.1005 and 0.1564 respectively. All are > 0.05, leading to the non rejection of the null hypothesis of equal means. (4)

[14]

QUESTION 6

(a) $\hat{\beta}_0 = 11$, $\hat{\beta}_1 = 3$ and $\sigma^2 = 9$. Thus, the least squares regression line is

$\widehat{Y} = 11 + 3x \Longrightarrow$ Increase in height $= 11 + 3$Fertilizer. (4)

(b)

	Fertilizer (grams/tree) x_i	Increase in height (cm) y_i	$y_i - \overline{y}$	$(y_i - \overline{y})^2$
Group 1	1	14	-9	81
	1	11	-12	144
	1	16	-7	49
Group 2	3	23	0	0
	3	19	-4	16
	3	20	-3	9
Group 3	5	20	-3	9
	5	30	7	49
	5	27	4	16
Group 4	7	35	12	144
	7	31	8	64
	7	30	7	49
Total	48	276	0	630

$$\sum_{i=1}^{12} (y_i - \overline{y})^2 - b_1^2 \sum_{i=1}^{12} (x_i - \overline{x})^2$$
$$= 630 - (3)^2\, 60$$
$$= 630 - 540$$
$$= 90$$

(4)

(c) Replace $X = 6$ in the regression equation in (a) then

$$\begin{aligned} Y &= 11 + 3(6) \\ &= 11 + 18 \\ &= 29 \end{aligned}$$

$\therefore$ The expected growth (increase in height) for $x = 6$ is 29

(1)

(d) $R^2 = 0.8571 \Longrightarrow 85.71\%$ of the variability in *increase in height* is being explained / or accounted for by the least squares line.

(1)

(e) Yes, the residual do not form any pattern. They form a null plot around zero. (2)

[12]

[**100**]