Tutorial letter 202/2/2017

Applied Statistics II

STA2601

Semester 2

Department of Statistics

Solutions to Assignment 02

QUESTION 1

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

$$\begin{aligned}
\text{Critical value} &= 0.2 + \frac{436-400}{450-400}(0.188-0.2)) \\
&= 0.2 + \frac{36}{50}(-0.012) \\
&= 0.2 + (-0.00864) \\
&\simeq 0.1914
\end{aligned}$$

Reject H_0 if $\beta_1 < -0.1914$ or $\beta_1 > 0.1914$ or $|\beta_1| > 0.1914$

$$\text{Now } \beta_1 = \frac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left(\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}\right)^3} = \frac{2\,648.266}{\left(\sqrt{124.942}\right)^3}$$

$$\begin{aligned}
&= \frac{2\,648.266}{(11.17774575)^3} \\
&= \frac{2\,648.266}{1\,396.569909} \\
&\simeq 1.8963.
\end{aligned}$$

Since $-1.8963 > 0.1914$ we reject H_0 at the 10% level of significance level and conclude that this distribution is not symmetric.

(7)

(ii) **Test for kurtosis:**

We have to test:

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

H_1 : The distribution is leptokurtic $\left(\Rightarrow \beta_2 > 3\right)$.

$n = 436$. From table B (page 111 study guide) the upper 5% critical value is

$$\begin{aligned}\text{Critical value} &= 3.41 + \frac{436-400}{450-400}(3.39-3.41)) \\ &= 3.41 + \frac{36}{50}(-0.02) \\ &= 3.41 + (-0.0144) \\ &\approx 3.3956\end{aligned}$$

We will reject H_0 at the 5% level of significance (one-sided) if $\beta_2 > 3.3956$

Now the value of the test statistic is

$$\begin{aligned}\beta_2 &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]^2} \\ &= \frac{119\,812.018}{[124.942]^2} \\ &= \frac{119\,812.018}{15\,610.50336} \\ &\approx 7.6751\end{aligned}$$

Since $7.6751 > 3.3956$, we reject H_0 at the 5% level of significance and conclude that the distribution is leptokurtic.

(7)

(iii) No, the distribution of sentence lengths does not originate from a normal distribution since it failed both tests (not symmetric and does not have the kurtosis of a normal distribution (i.e., it is leptokurtic)).

(1)

(b) H_0 : The sentence length distribution of the epistle to the Romans follows a Sichel distribution.

H_1 : The sentence length distribution of the epistle to the Romans does not follow a Sichel distribution.

Class interval	Observed frequency	Expected frequency	$\frac{(N_i - e_i)^2}{e_i}$
$1 - 5$	67	78	1.5513
$6 - 10$	144	132	1.0909
$11 - 15$	87	90	0.10
$16 - 20$	42	50	1.28
$21 - 25$	43	34	2.3824
$26 - 30$	14	12	0.3333
$31 - 35$	12	13	0.0769
$36 - 40$	6	10	1.60
$41 - 45$	7	5	0.80
$46 - 50$	9	6	1.50
> 50	5	6	0.1667

Test statistic:

$$
\begin{aligned}
Y^2 &= \sum_{i=1}^{11} \frac{(N_i - e_i)^2}{e_i} \\
&= 1.5513 + 1.0909 + 0.1 + ... + 0.1667 \\
&= 10.8815.
\end{aligned}
$$

We have $k - 1 = 10$. The critical value $\chi^2_{0.05;10} = 18.307$. Reject H_0 if $Y^2 \geq 18.307$

Since the test statistic $Y^2 = 10.8\dot{8}15 < 18.307$ we cannot reject the null hypothesis at the 5% level. We must conclude that a Sichel distribution is probably a good fit for this dataset of sentence lengths.

(15)

[30]

QUESTION 2

(a) Start the *JMP* program.

> Enter *Type of parent* in the first column and label it *Type of parent*.

(make sure to change the scale to nominal)

> Enter *Colour of the down* in the second column and label it *Colour of the down*.

(make sure to change the scale to nominal)

> Enter the frequency in the third column and label it *Count*.

Your data should look like this.

Type of parent	**Colour of the down**	**Count**
A	Coloured	210
A	White	50
B	Coloured	146
B	White	54
C	Coloured	34
C	White	6

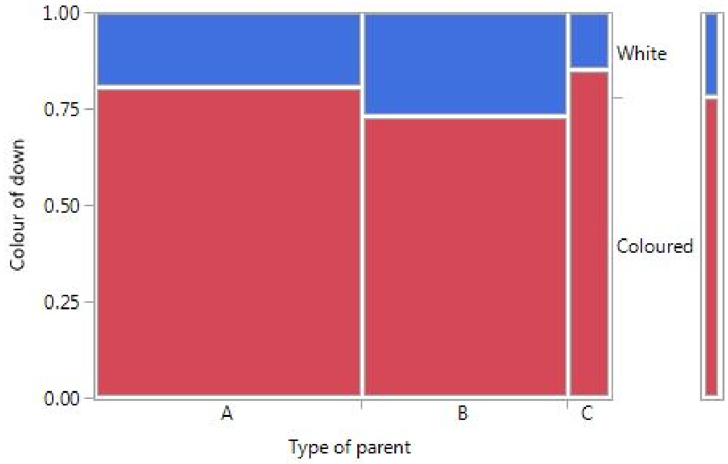
This is a chi-square test of association. To fit the model:

> Choose Analyze>Fit Y by X with *Type of parent* as X, Factor and *Colour of the down* as Y, Response and *Count* as $Freq$.

> Click Ok.

Contingency Analysis of Colour of down By Type of parent

Mosaic Plot



Freq: Count

Contingency Table

Colour of down

Type of parent	Count Total % Col % Row %	Coloured	White	Total
	A	210	50	260
		42.00	10.00	52.00
		53.85	45.45	
		80.77	19.23	
	B	146	54	200
		29.20	10.80	40.00
		37.44	49.09	
		73.00	27.00	
	C	34	6	40
		6.80	1.20	8.00
		8.72	5.45	
		85.00	15.00	
	Total	390	110	500
		78.00	22.00	

Tests

N	DF	-LogLike	RSquare (U)
500	2	2.6103568	0.0099

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.221	0.0735
Pearson	5.218	0.0736

The mosaic output shows that the proportion of coloured-down chicks was almost four times the proportion of whites in parent type A and parent type C. There were almost three times the number of coloured-down chicks as compared to whites in parent type B. However, the proportions seem to be almost the same as evidenced by the horizontal lines (alignments). The hypothesis of no association might not be rejected.

(11)

(b) H_0: There is no association between type of parent and the colour of the down of the chicks.

H_1: There is an association between type of parent and the colour of the down of the chicks

(2)

(c) The test statistic is $Y^2 = \sum_{k=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ and the value is $Y^2 = 5.218$.

(2)

(d) Yes, the row percentage seems to be similar. For coloured-down chicks it is 81%, 73% and 85% for parent type A, B, and C, respectively. One might expect the null hypothesis not to be rejected.

(3)

(e) The critical value is $\chi^2_{0.05;2} = 5.99147$. Since $5.218 < 5.99147$, we do not reject H_0 at the 5% level of significance and conclude that there is no association between parent type and colour down of chicks.

Alternatively the p-value is $= 0.0736$. Since $0.0736 > 0.05$, we do not reject H_0 at the 5% level of significance and conclude that there is no association between parent type and colour down of chicks.

(2)

[20]

QUESTION 3

(a) H_0 : There is no association between gender and final examination result.

H_1 : Females performed better than males.

The 2×2 table for the exact test is

		Gender			
		x			
		$\uparrow$			
	Male	Female			
Fail	4	1	5	$\longleftarrow$	n
Pass	4	3	7		
	8	4	12	$\longrightarrow$	N
		$\uparrow$			
		k			

We choose $k = 4$, $n = 5$ and $x = 1$.

The alternative (females performed better than males) would imply a small value of x to reject H_0, i.e. so small that $\mathbf{P}\left(\mathbf{X} \leq \mathbf{x}\right) \leq \boldsymbol{\alpha}$.

Now $x = 1$ and $P\left(X \leq 1\right) = 0.424$ (From table D).

Now $0.424 > 0.05 = \alpha$, we do not reject H_0 at the 5% level of significance and conclude that there is no association between gender and final examination result.

(10)

(b) (i) $H_0 : \rho = 0.3$ against $H_1 : \rho \neq 0.3$

$n = 35$ $R = 0.48$

$$\begin{aligned} U &= \frac{1}{2}\log_e \frac{1+R}{1-R} & \eta &= \frac{1}{2}\log_e \frac{1+\rho}{1-\rho} \\ &= \frac{1}{2}\log_e \frac{1+0.48}{1-0.48} & &= \frac{1}{2}\log_e \frac{1+0.3}{1-0.3} \\ &= \frac{1}{2}\log_e \frac{1.48}{0.52} & &= \frac{1}{2}\log_e \frac{1.3}{0.7} \\ &= \frac{1}{2}\log_e 2.846153846 & &= \frac{1}{2}\log_e 1.857142857 \\ &\approx 0.5230 & &\approx 0.3095 \end{aligned}$$

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

The test statistic is

$$\begin{aligned} z &= \sqrt{n-3}(U-\eta) \\ &= \sqrt{35-3}(0.5230-0.3095) \\ &= \sqrt{32} \times 0.2135 \\ &\approx 1.2077 \end{aligned}$$

$\alpha = 0.01$, $\alpha/2 = 0.005$ and $Z_{0.005} = 2.576$. Reject H_0 if $Z > 2.576$ or $Z < -2.576$ or $|Z| > 2.576$

Since $1.2077 < 2.576$, we do not reject H_0 and conclude that $\rho = 0.3$ at the 1% level of significance.

(8)

(ii) $\alpha = 0.05$, $\alpha/2 = 0.025$ and $Z_{0.025} = 1.96$.

The 95% confidence for η is

$$\begin{aligned} U - \frac{1.96}{\sqrt{n-3}} &< \eta < U + \frac{1.96}{\sqrt{n-3}} \\ 0.5230 - \frac{1.96}{\sqrt{35-3}} &< \eta < 0.5230 + \frac{1.96}{\sqrt{35-3}} \\ 0.5230 - \frac{1.96}{\sqrt{32}} &< \eta < 0.5230 + \frac{1.96}{\sqrt{32}} \\ 0.5230 - 0.3465 &< \eta < 0.5230 + 0.3465 \\ 0.1765 &< \eta < 0.8695 \end{aligned}$$

Now $\dfrac{e^{0.1765} - e^{-0.1765}}{e^{0.1765} + e^{-0.1765}} = \dfrac{1.1930 - 0.8382}{1.1930 + 0.8382} = \dfrac{0.3548}{2.0312} \approx 0.1747 \approx 0.17$

and $\dfrac{e^{0.8695} - e^{-0.8695}}{e^{0.8695} + e^{-0.8695}} = \dfrac{2.3857 - 0.4192}{2.3857 + 0.4192} = \dfrac{1.9665}{2.8049} \approx 0.7011 \approx 0.70$

i.e., 95% confidence interval for ρ is $(0.17; 0.70)$.

OR alternatively

Using Table X we have

for $\eta = 0.1717 : \rho = 0.17$ and $\eta = 0.1820 : \rho = 0.18$

Using linear interpolation for $\eta = 0.1765$

$$\begin{aligned} \rho &= 0.17 + \frac{0.1765 - 0.1717}{0.1820 - 0.1717}(0.18 - 0.17) \\ &= 0.17 + \frac{0.0048}{0.0103} \times 0.01 \\ &= 0.17 + 0.004660194 \\ &= 0.174660194 \\ &\approx 0.17 \end{aligned}$$

for $\eta = 0.8673 : \rho = 0.7$ and $\eta = 0.8872 : \rho = 0.71$

Once more using linear interpolation for $\eta = 0.8695$

$$\begin{aligned}
\rho &= 0.7 + \frac{0.8695 - 0.8673}{0.8872 - 0.8673}(0.71 - 0.7) \\
&= 0.7 + \frac{0.0022}{0.0199} \times 0.01 \\
&= 0.7 + 0.001105527 \\
&= 0.701105527 \\
&\approx 0.70
\end{aligned}$$

Thus, the 95% confidence interval for ρ is $(0.17; 0.70)$.

(7)

(c) $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 < \rho_2$

$r_1 = 0.5 \quad n_1 = 103$
$r_2 = 0.8 \quad n_2 = 52$

$$\begin{aligned}
U_1 &= \frac{1}{2}\log_e \frac{1+r_1}{1-r_1} & U_2 &= \frac{1}{2}\log_e \frac{1+r_2}{1-r_2} \\
&= \frac{1}{2}\log_e \frac{1+0.5}{1-0.5} & &= \frac{1}{2}\log_e \frac{1+0.8}{1-0.8} \\
&= \frac{1}{2}\log_e \frac{1.5}{0.5} & &= \frac{1}{2}\log_e \frac{1.8}{0.2} \\
&= \frac{1}{2}\log_e 3 & &= \frac{1}{2}\log_e 9 \\
&\approx 0.5493 & &\approx 1.0986
\end{aligned}$$

(or just read the values for U_1 and U_2 from table X)

The test statistic is

$$
\begin{aligned}
z &= \frac{U_1 - U_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}} \\
&= \frac{0.5493 - 1.0986}{\sqrt{\dfrac{1}{103 - 3} + \dfrac{1}{53 - 3}}} \\
&= \frac{-0.5493}{\sqrt{\dfrac{1}{100} + \dfrac{1}{50}}} \\
&= \frac{-0.5493}{\sqrt{0.03}} \\
&= \frac{-0.5493}{0.17320508} \\
&\approx -3.1714
\end{aligned}
$$

$\alpha = 0.05$ and $Z_{0.05} = 1.645$. We reject H_0 if $Z < -1.645$.

Since $-3.1714 < -1.645$, we reject H_0 at the 5% level of significance and conclude that $\rho_1 < \rho_2$, i.e., the correlation coefficient for population 1 is significantly smaller than that for population 2.

(10)

[35]

QUESTION 4

(a) If μ is unknown, a 95% confidence interval for σ^2 is

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

Then

$$
\sum_{i=1}^{9} X_i = 1\,125; \qquad \sum_{i=1}^{9} X_i^2 = 140\,665; \qquad \overline{X} = 1\,125/9 = 125
$$

$$\begin{aligned}\Sigma\left(X_i-\overline{X}\right)^2 &= \sum_{i=1}^{9} X_i^2 - n\overline{X}^2\\ &= 140\,665 - 9\,(125)^2\\ &= 140\,665 - 140\,625\\ &= 40\end{aligned}$$

$$\begin{aligned}\chi^2_{\frac{1}{2}\alpha;n-1} &= \chi^2_{0.025;8} &= 17.5346\\ \chi^2_{1-\frac{1}{2}\alpha;n-1} &= \chi^2_{0.975;8} &= 2.17973\end{aligned}$$

Thus, if μ is unknown, a 95% confidence interval for σ^2 is

$$\begin{aligned}&\left[\frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{1}{2}\alpha;n-1}} < \sigma^2 < \frac{\Sigma\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{1}{2}\alpha;n-1}}\right]\\ &\left[\frac{40}{17.5346} < \sigma^2 < \frac{40}{2.17973}\right]\\ &\left[2.2812 < \sigma^2 < 18.3509\right]\\ &[2.28;\, 18.35]\,.\end{aligned}$$

(9)

(b) If $\mu = 125$, a 95% confidence interval for σ^2 is

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

$$\Sigma\,(X_i-\mu)^2 = 40$$

$$\begin{aligned}\chi^2_{\frac{1}{2}\alpha;n} &= \chi^2_{0.025;9} &= 19.0228\\ \chi^2_{1-\frac{1}{2}\alpha;n} &= \chi^2_{0.975;9} &= 2.70039\end{aligned}$$

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

$$\left[\frac{\Sigma\,(X_i-\mu)^2}{\chi^2_{\frac{1}{2}\alpha;n}} < \sigma^2 < \frac{\Sigma\,(X_i-\mu)^2}{\chi^2_{1-\frac{1}{2}\alpha;n}}\right]$$
$$\left[\frac{40}{19.0228} < \sigma^2 < \frac{40}{2.70039}\right]$$
$$\left[2.1027 < \sigma^2 < 14.8127\right]$$
$$[2.10;\ 14.81]\,.$$

(6)

[15]

[**100]**