Tutorial letter 202/1/2017

Applied Statistics II

STA2601

Semester 1

Department of Statistics

Solutions to Assignment 02

QUESTION 1

(a)

	x_i	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^3$	$(x_i - \overline{x})^4$
	0.48	−0.13	0.0169	−0.002197	0.00028561
	0.96	0.35	0.1225	0.042875	0.01500625
	0.52	−0.09	0.0081	−0.000729	0.00006561
	0.36	−0.25	0.0625	−0.015625	0.00390625
	0.49	−0.12	0.0144	−0.001728	0.00020736
	⋮	⋮	⋮	⋮	⋮
	⋮	⋮	⋮	⋮	⋮
	0.74	0.13	0.0169	0.002197	0.00028561
	0.55	−0.06	0.0036	−0.000216	0.00001296
	0.49	−0.12	0.0144	−0.001728	0.00020736
	0.58	−0.03	0.0009	−0.000027	0.00000081
	0.64	0.03	0.0009	0.000027	0.00000081
Total	36.6	0	1.4572	0.116088	0.104689

$n = 60$ $\qquad \sum X = 36.6$ $\qquad \overline{X} = 0.61$

$\sum \left(X_i - \overline{X}\right)^2 = 1.4572$ $\qquad \sum \left(X_i - \overline{X}\right)^3 = 0.116088$ $\qquad \sum \left(X_i - \overline{X}\right)^4 = 0.104689$

(9)

(b) (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.)

The critical value is 0.492. Reject H_0 if $\beta_1 < -0.492$ or $\beta_1 > 0.492$ or $|\beta_1| > 0.492$.

$$\text{Now } \beta_1 = \frac{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^3}{\left(\sqrt{\frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2}\right)^3} = \frac{\frac{1}{60}(0.116088)}{\left(\sqrt{\frac{1}{60}(1.4572)}\right)^3}$$

$$= \frac{0.0019348}{\left(\sqrt{0.024286666}\right)^3}$$

$$= \frac{0.0019348}{(0.1558418)^3}$$

$$= \frac{0.0019348}{0.003784877}$$

$$\simeq 0.5112$$

Since $0.5112 > 0.492$ we reject H_0 at the 10% level of significance level and conclude that the 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 : $\beta_2 \neq 3$.

Since we have a sample size greater than 50, the test is based on β_2 (page 111 in the study guide).

The size of the sample, $n = 60$, thus 60 is between 50 and 75, we need to interpolate the critical values.

From table B (page 111 in the study guide):

The upper 5% percentage point for B_2 is

$$3.99 + \frac{(60 - 50)}{75 - 60}(3.87 - 3.99) = 3.99 + \frac{10}{25}(-0.12) = 3.942.$$

The lower 5% percentage point for A is

$$2.15 + \frac{(60 - 50)}{75 - 60}(2.27 - 2.15) = 2.15 + \frac{10}{25}(0.12) = 2.198.$$

We reject H_0 at the 10% significance level if $B_2 <$ lower 5% point or $B_2 >$ upper 5% point in table C.

The critical values are 3.942 and 2.198. Reject H_0 if $B_2 < 2.198$ or $B_2 > 3.942$.

Now the value of the test statistic is

$$\begin{aligned}
\beta_2 &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]^2} \\
&= \frac{\frac{1}{60}(0.104689)}{\left[\frac{1}{60}(1.4572)\right]^2} \\
&= \frac{0.001744816}{[0.024286666]^2} \\
&= \frac{0.001744816}{0.000589842} \\
&\approx 2.9581
\end{aligned}$$

Since $2.198 < 2.9581 < 3.942$, we do not reject H_0 at the 10% level of significance and conclude that the distribution does have the kurtosis of a normal distribution.

(7)

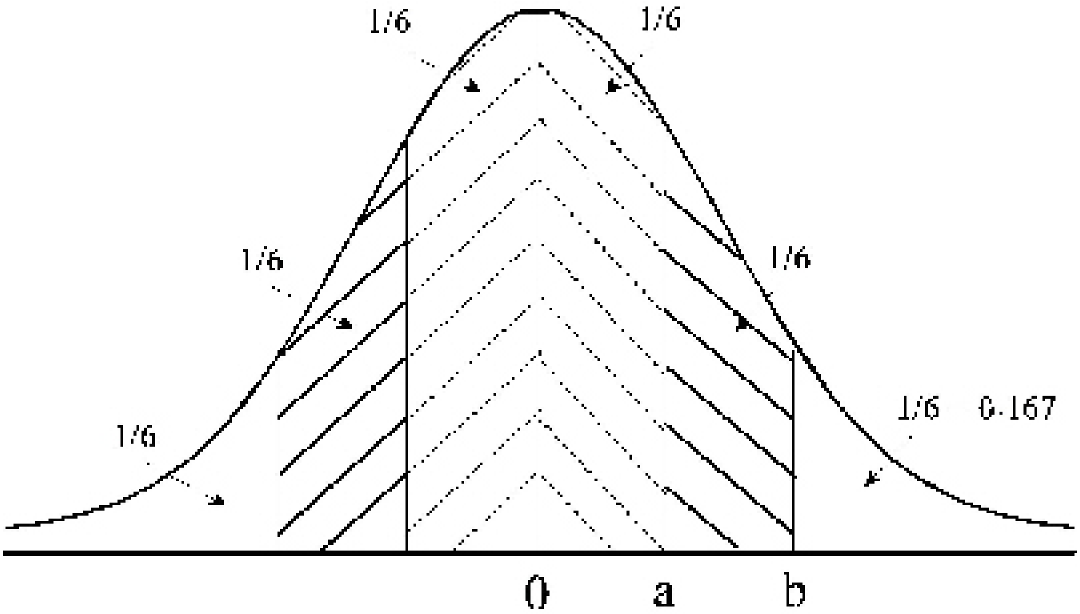
(iii) No, the sample failed one test and we conclude that the distribution does not originate from a normal distribution.

(1)

(c) (i) If we divide the observations into 6 classes with equal expected frequencies, it means that $\pi_i = \frac{1}{6}$ for each interval $\Rightarrow n\pi_i = 10$.
The biggest problem is to *determine the interval limits* in terms of the X-scale such that each interval has a probability of $\frac{1}{6} = 0.167$.

We start with the standardised $n(0;1)$ scale (as always) and transform back to the X-scale by making use of

$Z = \frac{X-\mu}{\sigma} = \frac{X-0.6}{0.16}$.



From the first interval we have $P[Z \leq -b] = 0.167$ and we find b by working with the positive mirror image. i.e., $P(Z \geq b) = 0.167 \Longrightarrow P(Z \leq b) = 0.833$ (this probability is $1 - 0.167$).

From table II we get that $\Phi(0.966) = 0.833 \Rightarrow b = 0.966$.

The first interval is where $Z \leq -0.966$ [Note that P$(Z \leq -0.966) = 0.167$.

$$\begin{aligned}
\Longrightarrow \frac{X-\mu}{\sigma} = \frac{X-0.6}{0.16} &\leq -0.966 \\
X - 0.6 &\leq -0.966 \times 0.16 \\
X - 0.6 &\leq -0.15456 \\
X &\leq 0.6 - 0.15456 \\
X &\leq 0.44544 \\
X &\leq 0.445
\end{aligned}$$

Hence the first interval is where $X \leq 0.445$.

The second interval is where $-b \leq Z \leq -a$. From the sketch above the value "a" is found from Table II as $\Phi(0.432) = P(Z \leq a) = 0.5 + 0.167 = 0.667$. Thus $-a = -0.432$.

Thus

$$\begin{array}{rcccl}
-0.966 & \leq & Z & \leq & -0.432 \\
-0.966 & \leq & \dfrac{X-0.6}{0.16} & \leq & -0.432 \\
-0.966 \times 0.16 & \leq & X-0.6 & \leq & -0.432 \times 0.16 \\
-0.15456 & \leq & X-0.6 & \leq & -0.06912 \\
0.6-0.15456 & \leq & X & \leq & 0.6-0.06912 \\
0.44544 & \leq & X & \leq & 0.53088 \\
0.445 & \leq & X & \leq & 0.531
\end{array}$$

The second interval is $0.445 \leq X \leq 0.531$.

(10)

(ii) Expected frequency is $n\hat{\pi}_i = \frac{1}{6} \times 60 = 10$.

Table of observed and expected frequencies		
Class interval	O_i	$\hat{e}_i = n\hat{\pi}_i$
$X < 0.445$	10	10
$0.445 \leq X < 0.531$	8	10
$0.531 \leq X < 0.600$	14	10
$0.600 \leq X < 0.669$	7	10
$0.669 \leq X < 0.755$	12	10
$X \geq 0.755$	9	10
Totals	**60**	**60**

(1)

(iii) Goodness of fit test:

We have to test

H_0 : The observations come from a $n\left(0.6, 0.16^2\right)$ distribution

H_1 : The observations do not come from a $n\left(0.6, 0.16^2\right)$ distribution.

We use the test statistic

$$\begin{aligned} Y^2 &= \sum_{i=1}^{6} \frac{(N_i - n\pi_i)^2}{n\pi_i} \qquad \text{(\textbf{page 98 study guide})} \\ \therefore Y^2 &= \tfrac{1}{10}\left((10-10)^2 + (8-10)^2 + (14-10)^2 + (7-10)^2 + (12-10)^2 + (9-10)^2\right) \\ &= \tfrac{1}{10}(0+4+16+9+4+1) \\ &= \frac{34}{10} \\ &= 3.4 \end{aligned}$$

The critical value is

$$\begin{aligned} \chi^2_{\alpha;k-r-1} &= \chi^2_{0.05;6-0-1} \\ &= \chi^2_{0.05;5} \\ &= 11.0706 \end{aligned}$$

Reject H_0 if $Y^2 \geq 11.0706$.

Since $3.4 < 11.0706$, we cannot reject H_0 at the 5% level of significance and conclude that the data follow a normal distribution.

FOR YOUR INFORMATION: For the degrees of freedom $k - r - 1$ where k is the number of classes and r is the number of estimated parameters.

Test	Value of r	Parameter unknown
$n\left(20, 5^2\right)$	$r = 0$	
$n\left(20, \sigma^2\right)$	$r = 1$	σ unknown
$n\left(\mu, 5^2\right)$	$r = 1$	μ unknown
$n\left(\mu, \sigma^2\right)$	$r = 2$	both μ and σ unknown

(10)

[45]

QUESTION 2

(a) Start the *JMP* program.

> Enter *Job satisfaction* in the first column and label it *Job satisfaction*.

(make sure to change the scale to nominal)

> Enter *Type of work* in the second column and label it *Type of work*.

(make sure to change the scale to ordinal)

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

Your data should look like this.

Aggressiveness	**Sales**	**Count**
Satisfied	White collar	81
Satisfied	Blue collar	124
Satisfied	Menial	44
Dissatisfied	White collar	49
Dissatisfied	Blue collar	76
Dissatisfied	Menial	56

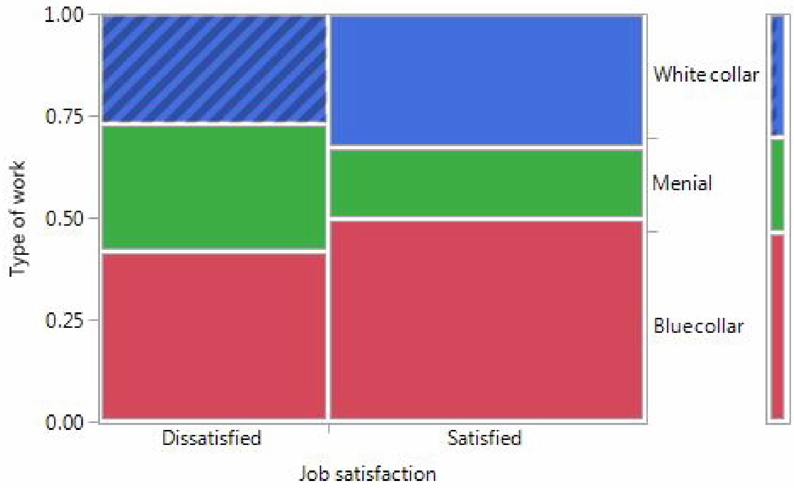
This is a chi-square test of association. To fit the model:

> Choose Analyze>Fit Y by X with *Job satisfaction* as X, Factor and *Type of work* as Y, Response and *Count* as $Freq$.

> Click Ok.

Contingency Analysis of Type of work By Job satisfaction

Mosaic Plot



Freq: Number of workers

Contingency Table

Job satisfaction	Count / Total % / Col % / Row %	Type of work: Blue collar	Menial	White collar	Total
Dissatisfied	Count	76	56	49	181
	Total %	17.67	13.02	11.40	42.09
	Col %	38.00	56.00	37.69	
	Row %	41.99	30.94	27.07	
Satisfied	Count	124	44	81	249
	Total %	28.84	10.23	18.84	57.91
	Col %	62.00	44.00	62.31	
	Row %	49.80	17.67	32.53	
Total	Count	200	100	130	430
	Total %	46.51	23.26	30.23	

Tests

N	DF	-LogLike	RSquare (U)
430	2	5.1181981	0.0113

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.236	0.0060*
Pearson	10.342	0.0057*

The **Mosaic Plot** shows that the sample sizes for satisfied and dissatisfied are almost in the ratio 3:2. The mosaic plot shows that the proportion of blue collar who are satisfied was almost twice those in white collar and three times those in menial. Thus the blue collar comprises almost half of those satisfied. The proportion of those dissatisfied who are blue collar was slightly higher than the other types of works who were almost the same. Thus, the proportions seem to be not the same as evidenced by horizontal lines which are not in alignment. The hypothesis of no association might be rejected.

(11)

(b) H_0: Job satisfaction is not related to type of work.

H_1: Job satisfaction is related to type of work.

(2)

(c) The test statistic is $Y^2 = \sum_{k=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$ and the value is $Y^2 = 10.342$.

(2)

(d) No, the row percentages seems to be different. For dissatisfied are 41.99%, 30.94%, 27.07% and for satisfied are 49.80%, 17.67% and 32.53% for blue collar, menial and white collar respectively. One might expect the null hypothesis to be rejected.

(3)

(e) The critical value is $\chi^2_{0.05;2} = 5.99147$. Since $10.342 > 5.99147$, we reject H_0 at the 5% level of significance and conclude that job satisfaction is related to type of work.

Alternatively the p-value is $= 0.0057$. Since $0.0057 < 0.05$, we reject H_0 at the 5% level of significance and conclude that job satisfaction is related to type of work.

(2)

[20]

QUESTION 3

H_0 : There is no difference in opinion between psychologists and psychiatrists regarding behavioural modification as a useful technique.

H_1 : Psychiatrists feel less favourable towards the usefulness of behavioural modification

For this 2×2 table for the exact test is

Is behavioural modification useful?				
	Yes	**No**	**Row total**	
Psychologists	5	2	7	
Psychiatrists	1	$4 (= x)$	5	$\leftarrow k$
Column total	6	6	12	
		$\uparrow$ n		

Now $k = 5$, $n = 6$ and $x = 4$ [Note: $k = 5$, $n = 6$ and $x = 1$ thus $P(X \leq 1) = 0.121$]

The alternative "Psychiatrists feel less favourable towards the usefulness of behavioural modification" would imply a large value of x to reject H_0 i.e., $P(X \geq x) = \alpha$.

Now $x = 4$ and

$$\begin{aligned} P\left(X \geq x\right) &= 1 - P(X < x - 1) \\ P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - 0.879 \\ &= 0.121 \ \text{(from table D study guide p131)} \end{aligned}$$

Since $0.121 > 0.05$, we do not reject H_0 at the 5% level of significance and conclude that there is no difference in opinion between psychologists and psychiatrists regarding behavioural modification as a useful technique.

[10]

QUESTION 4

(a) If $U = \dfrac{\Sigma\,(X_i - \mu)^2}{\sigma^2}$ then $U \sim \chi^2_n$ (result 1.2).

Then

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

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

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

(7)

(b) (i) We have to test $H_0 : \sigma_1^2 = \sigma_2^2$

against $H_1 : \sigma_1^2 \neq \sigma_2^2$

$$n_1 = 17 \qquad S_1^2 = (1.12)^2 = 1.2544 \qquad n_2 = 12 \qquad S_2^2 = (2.36)^2 = 5.5696$$

The test statistic is

$$\begin{aligned} F &= \frac{\sigma_2^2}{\sigma_1^2} \times \frac{S_1^2}{S_2^2} \\ &= 1 \times \frac{1.2544}{5.5696} \\ &\approx 0.2252 \end{aligned}$$

The critical values is $F_{\alpha/2;n_1-1;n_2-1} = F_{0.025;16;11} \approx 3.33$ (or actual is $F_{0.025;16;11} = 3.30$ (internet F-calculator)) and $F_{1-\alpha/2;n_1-1;n_2-1} = \dfrac{1}{F_{\alpha/2;n_2-1;n_1-1}} = \dfrac{1}{F_{0.025;11;16}} = \dfrac{1}{\frac{1}{2}(2.99+2.89)} = \dfrac{1}{2.94} \approx 0.3401$. Reject H_0 if $F > 3.33$ or $F < 0.3401$.
Since $0.2252 < 0.3401$, we reject H_0 at the 5% level of significance and conclude that the variances are not equal i.e. $\sigma_1^2 \neq \sigma_2^2$.

(10)

(ii) The 95% confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$ is

$$P\left(F_{1-\frac{\alpha}{2};n_2-1;n_1-1} < \frac{\sigma_1^2}{\sigma_2^2}\frac{S_2^2}{S_1^2} < F_{\frac{\alpha}{2};n_2-1;n_1-1}\right) = 1-\alpha$$

$$\left[\frac{F_{1-\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}; \frac{F_{\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}\right]$$

$\alpha = 0.05, \alpha/2 = 0.025$

$F_{1-\frac{\alpha}{2};n_2-1;n_1-1} = F_{0.975;11;16} = \dfrac{1}{F_{0.025;16;11}} \approx \dfrac{1}{3.33} \approx 0.3003$

$F_{\frac{\alpha}{2};n_2-1;n_1-1} = F_{0.025;11;16} = \dfrac{1}{2}(2.99+2.89) = 2.94$

$\therefore$The 95% confidence interval is

$$\begin{gathered} \left[\frac{F_{1-\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}; \frac{F_{\frac{\alpha}{2};n_2-1;n_1-1}}{S_2^2/S_1^2}\right] \\ \left[\frac{0.3003}{5.5696/1.2544}; \frac{2.94}{5.5696/1.2544}\right] \\ \left[\frac{0.3003}{4.44005102}; \frac{2.94}{4.44005102}\right] \\ [0.0676; 0.6622]. \end{gathered}$$

We are 95% confident that the confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$ lies between 0.0676 to 0.6622.

(8)

[25]

[**100**]