Assignment 5 (Week 9 - 10)

STAT 2601 - Business Statistics (2024 Fall)
School of Mathematics and Statistics, Carleton University

Due Date and Time: Wednesday 27 November 2024, before 10:00 am
Total Marks: 30

Q1: **[6]** A random sample of 518 Canadian residents was asked to indicate their preferred news channel for global news from a list of popular channels. The results are summarized in the following table:

Channel	**Frequency**
CBC News	110
CTV News	105
Global News	80
BBC News	50
Al-Jazeera English	100
CNN	63
Sky News	10

Using a 5% significance level, we wish to determine if there is an evidence to suggest that these seven news channel are not equally preferred.

1. State the null and alternative hypotheses. [1]

   Solution: Define

$$
\begin{aligned}
P_1 &= p(\text{prefer watching CBC News}).\\
P_2 &= p(\text{prefer Watching CTV News}).\\
P_3 &= p(\text{prefer Watching Global News}).\\
P_4 &= p(\text{prefer watching BBC News}).\\
P_5 &= p(\text{prefer watching Al-Jazeera English}).\\
P_6 &= p(\text{prefer watching CNN}).\\
P_7 &= p(\text{prefer watching sky News}).
\end{aligned}
$$

   Thus,

$$
\begin{aligned}
&[0.5]\ H_0 : P_1 = P_2 = P_3 = P_4 = P_5 = P_6 = P_7 = \frac{1}{7}\\
&[0.5]\ H_a : \text{ At least two probabilities differ from } \frac{1}{7}
\end{aligned}
$$

2. Compute the test statistic. [2]

   Solution: See the table below

$$
\begin{aligned}
[1]\ \chi^2\text{-test } &= \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = 17.51 + 12.99 + 0.49 + 7.78 + 9.14 + 1.64 + 55.35\\
&\approx 104.9
\end{aligned}
$$

Channel	f_i	[0.5] p_i	[0.5] $e_i = nP_i$	$\frac{(f_i - e_i)^2}{e_i}$
CBC News	110	$\frac{1}{7}$	$518 \times \frac{1}{7} = 74$	$\frac{(110-74)^2}{74} = 17.51$
CTV News	105	$\frac{1}{7}$	$518 \times \frac{1}{7} = 74$	$\frac{(105-74)^2}{74} = 12.99$
Global News	80	$\frac{1}{7}$	$518 \times \frac{1}{7} = 74$	$\frac{(80-74)^2}{74} = 0.49$
BBC News	50	$\frac{1}{7}$	$518 \times \frac{1}{7} = 74$	$\frac{(50-74)^2}{74} = 7.78$
Al-Jazeera English	100	$\frac{1}{7}$	$518 \times \frac{1}{7} = 74$	$\frac{(100-74)^2}{74} = 9.14$
CNN	63	$\frac{1}{7}$	$518 \times \frac{1}{7} = 74$	$\frac{(63-74)^2}{74} = 1.64$
Sky News	10	$\frac{1}{7}$	$518 \times \frac{1}{7} = 74$	$\frac{(10-74)^2}{74} = 55.35$
Total	518	1	518	≈ 104.9

3. Use the critical value method to make a final decision about the hypotheses in part (a). [1]

   Solution: Since we have $df = k-1 = 7-1 = 6$, our critical value at 5% significance level is [0.5] $\chi_{0.05}(6) = 12.592$. Since [0.5] χ^2-test statistic $= 104.89 > \chi_{0.05}(6) = 12.592$, we reject H_0 and accept H_a. Therefore, we conclude that there is sufficient evidence at the 5% significance level to reject the hypothesis that Canadians equally prefer these seven popular news channels.

4. Are the conditions for inference using a chi-square test satisfied? [2 pt]

   Solution: [1] Yes, the result is (approximately) valid because

   - we used a [0.5] random sample, and
   - [0.5] all expected values are at least 5. That is, $e_i = 74 \geq 5$ for all $i = 1, 2, \cdots, 7$.

Q2: **[11 pt]** A survey asked 690 randomly sampled Ottawa residents which shipping carrier they prefer to use for shipping holiday gifts. The results are summarized in the table below.

		Age			Total
		$18-39$	$40-59$	$60+$	
Shipping Method	Canada Post	88	90	102	280
	UPS	52	77	31	160
	FedEx	78	66	56	200
	Something else	8	8	14	30
	Not sure	4	7	9	20
	Total	230	248	212	690

1. Calculate the expected counts for each cell of this Table. [7.5]

Solution:

Shipping Method	**Age** $18-39$	$40-59$	$60+$	Total
Canada Post	88 $(e_{11} = \frac{280 \times 230}{690} = 93.33)$[0.5]	90 $(e_{12} = \frac{280 \times 248}{690} = 100.64)$[0.5]	102 $(e_{13} = \frac{280 \times 212}{690} = 86.03)$[0.5]	280
UPS	52 $(e_{21} = \frac{160 \times 230}{690} = 53.33)$[0.5]	77 $(e_{22} = \frac{160 \times 248}{690} = 57.51)$[0.5]	31 $(e_{23} = \frac{160 \times 212}{690} = 49.16)$[0.5]	160
FedEx	78 $(e_{31} = \frac{200 \times 230}{690} = 66.67)$[0.5]	66 $(e_{32} = \frac{200 \times 248}{690} = 71.88)$[0.5]	56 $(e_{33} = \frac{200 \times 212}{690} = 61.45)$[0.5]	200
Something else	8 $(e_{41} = \frac{30 \times 230}{690} = 10.00)$[0.5]	8 $(e_{42} = \frac{30 \times 248}{690} = 10.78)$[0.5]	14 $(e_{43} = \frac{30 \times 212}{690} = 9.22)$[0.5]	30
Not sure	4 $(e_{51} = \frac{20 \times 230}{690} = 6.67)$[0.5]	7 $(e_{52} = \frac{20 \times 248}{690} = 7.19)$[0.5]	9 $(e_{53} = \frac{20 \times 212}{690} = 6.14)$[0.5]	20
Total	230	248	212	690

2. State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Ottawa residents. [1 pt]

Solution: ([0.5 point] for H_0 and [0.5 point] for H_1).

We can write the hypotheses using the word associated as follows:

H_0 : There is no association between the age group and preferred shipping method for holiday gifts among Ottawa residents

H_a : The age group and preferred shipping method for holiday gifts among Ottawa residents are associated

Another way is to write the hypotheses using the words independent and dependent as follows:

H_0 : The age and preferred shipping method for holiday gifts among Ottawa residents are independent

H_a : The age and preferred shipping method for holiday gifts among Ottawa residents are not independent (dependent)

3. Conduct a test of independence using the critical value method and a 1% significance level. [2.5 pt]

Solution:

$$
\begin{aligned}
[1]\ \chi^2\text{-test } &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = \frac{(88-93.33)^2}{93.33} + \frac{(90-100.64)^2}{100.64} + \frac{(102-86.03)^2}{86.03} \\
&+ \frac{(52-53.33)^2}{53.33} + \frac{(77-57.51)^2}{57.51} + \frac{(31-49.16)^2}{49.16} \\
&+ \frac{(78-66.67)^2}{66.67} + \frac{(66-71.88)^2}{71.88} + \frac{(56-61.45)^2}{61.49} \\
&+ \frac{(8-10.00)^2}{10.00} + \frac{(8-10.78)^2}{10.78} + \frac{(14-9.22)^2}{9.22} \\
&+ \frac{(4-6.67)^2}{6.67} + \frac{(7-7.19)^2}{7.19} + \frac{(9-6.14)^2}{6.14} \\
&= 26.63
\end{aligned}
$$

The critical value at $\alpha = 0.01$ level of significance and degrees of freedom $= (5-1) \times (3-1) = 8$ is [0.5] $\chi^2_{0.01}(8) = 20.09$. [0.5]Since the test statistic $\chi^2 = 26.63$ is grater than the critical value $\chi^2_{0.01}(8) = 20.09$, then we will [0.5]reject the null hypothesis and conclude that there is significant evidence at 1% level of significance that the age and preferred shipping method for holiday gifts among Ottawa residents are associated (not independent = dependent).

Q3: [13 pt] A company wants to study how advertising spending (in thousands of dollars) affects its sales revenue (in thousands of dollars). The company collects data for the past 6 months, as shown below:

Month	Advertising Spending (X)	Sales Revenue (Y)
1	6	71
2	7	82
3	9	90
4	10	105
5	5	55
6	4	50

Consider advertising spending (X) as the independent variable and sales revenue (Y) as the dependent variable. Use the following statistics to answer the questions below:

$$\sum_{i=1}^{6} x_i = 41, \quad \sum_{i=1}^{6} x_i^2 = 307, \quad \sum_{i=1}^{6} y_i = 453, \quad \sum_{i=1}^{6} y_i^2 = 36415, \quad \sum_{i=1}^{6} x_i y_i = 3335, \quad \text{Standard Error of Estimate, } s = 4.36$$

1. Calculate the sample correlation coefficient, r, between the advertising spending and sales revenue and intercept it in context. [2.5]

   Solution:

   $$[1] \quad r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}} = \frac{3335 - \frac{(41)(453)}{6}}{\sqrt{307 - \frac{(41)^2}{6}}\sqrt{36415 - \frac{(453)^2}{6}}} = 0.983$$

   There is a strong [0.5] positive [0.5] linear [0.5] relationship (correlation) between advertising spending and sales avenue.

2. Calculate the coefficient of determination and intercept it in context. [1]

   Solution: [0.5] $R^2 = 0.966$. [0.5] Indicates that approximately 97% of the variability in sales revenue can be explained by advertising spending.

3. Fit a simple linear regression model to predict the sales revenue based on the advertising spending. Write down the fitted regression line. [1.5]

   Solution:

   $$\hat{y} = b_0 + b_1 x,$$

   where

   $$[0.5] \quad b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = 8.93$$

   $$[0.5] \quad b_0 = \bar{y} - b_1 \bar{x} = \left(\frac{453}{6}\right) - (8.93)\left(\frac{41}{6}\right) = 14.51$$

   Thus,

   $$[0.5] \quad \hat{y} = 14.51 + 8.93x$$

   or

   $$\textit{predicted sale avenue} = 14.51 + 8.93\ (\textit{advertising spending})$$

4. Interpret the slope of the regression coefficient. [1]

   Solution: For each additional \$1,000 advertising spending, the average sales revenue will increase \$8,930.

5. Calculate the error in prediction the sales revenue when advertising spending is \$5,000. [1]

   Solution:

   $$\hat{e}_{|x=5} = y_{|x=5} - \hat{y}_{|x=5},$$

   where

   $$y_{|x=5} = 55, \quad \text{and} \quad [0.5]\hat{y}_{|x=5} = 14.51 + 8.93(5) = 59.16.$$

   Thus,

   $$[0.5]\hat{e}_{|x=5} = 55 - 59.16 = -4.16 \equiv -\$4,160$$

6. Test at the 0.05 significance level whether there is a linear relationship between sales revenue and advertising spending. Specify the hypotheses and write the decision using both approaches: the critical value and p-value. [3]

Solution: We can solve this question in two ways: [3 points] for solving this question using any of these two methods. [0.5] for writing the hypotheses, [1] for test statistic, [0.5] for critical value, [0.5] for p-value, and [0.5] for rejecting the null hypothesis.

Method 1: Test the hypotheses about the population slope of the regression line β_1

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

This is a non-directional two-tailed test

The test statistic is

$$t = \frac{b_1 - \beta_1}{s_{b_1}},$$

where

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}} = \frac{s}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}} = \frac{4.36}{\sqrt{307 - \frac{(41)^2}{6}}} \approx 0.842$$

$$\Rightarrow t - \text{test} = \frac{8.93 - 0}{0.842} \approx 10.61$$

The critical value at $\alpha = 0.05$ and degrees of freedom $= n - 2 = 6 - 2 = 4$ is $t_{0.025}(4) = 2.77645$.

Since t-test $= 10.61 > t_{0.025}(4) = 2.77645$, we reject H_0 and conclude that there is sufficient evidence to support a positive linear relationship between sales revenue and advertising spending.

P-value $= 2P(T_{df=4} > 10.61) < 2(0.0005) = 0.001$. This p-value is less than $\alpha = 0.05$, hence we reject H_0 and conclude that there is sufficient evidence to support a positive linear relationship between sales revenue and advertising spending.

Method 2: Test the hypotheses about the population correlation coefficient ρ

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

The test statistic for this non-directional two-tailed test is

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.983 - 0}{\sqrt{\frac{1-0.966}{4}}} = 10.66,$$

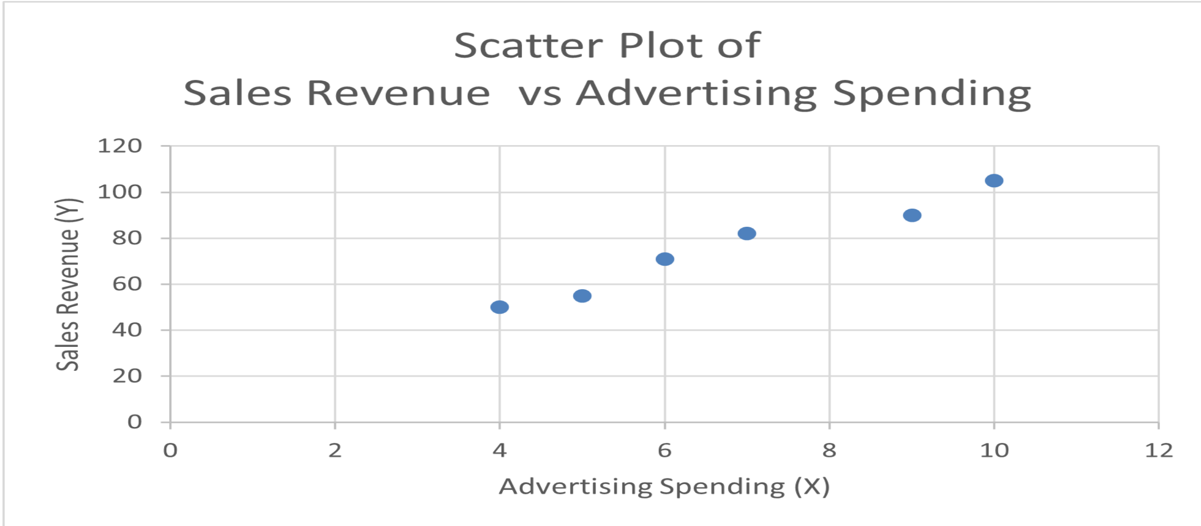
The critical value at $\alpha = 0.05$ and degrees of freedom $= n - 2 = 6 - 2 = 4$ is $t_{0.025}(4) = 2.77645$.

Since t-test $= 10.66 > t_{0.025}(4) = 2.77645$, we reject H_0 and conclude that there is sufficient evidence to support a positive linear relationship between sales revenue and advertising spending.

P-value $= 2P(T_{df=4} > 10.61) < 2(0.0005) = 0.001$. This p-value is less than $\alpha = 0.05$, hence we reject H_0 and conclude that there is sufficient evidence to support a positive linear relationship between sales revenue and advertising spending.

7. Use Excel, as explained in Lab 5, to construct a scatter plot of the sales revenue versus advertising spending for the sample data. [1]

Solution:



8. Use Excel, as explained in Lab 5, to estimate the fitted simple regression model that could be used to predict the sales revenue based on the advertising spending. Show the output results. What Significance F in ANOVA table tells you? [2]

Solution: [1] for the three tables output and [1] for explaining the significance F in ANOVA table

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.982716938							
5	R Square	0.965732581							
6	Adjusted R Square	0.957165726							
7	Standard Error	4.354622018							
8	Observations	6							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	2137.649068	2137.649068	112.7289551	0.000445475			
13	Residual	4	75.85093168	18.96273292					
14	Total	5	2213.5						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	14.50931677	6.013214747	2.41290514	0.073323781	-2.18604388	31.20467742	-2.18604388	31.20467742
18	X Variable 1	8.925465839	0.840646001	10.61738928	0.000445475	6.591458364	11.25947331	6.591458364	11.25947331

The Significance F in the ANOVA table is the p-value for testing the hypotheses

H_0 : The regression model is bad for prediction the sales revenue based on the advertising spending or (there is no linear relationship between sales revenue and advertising spending)

H_1 : The regression model is good for prediction the sales revenue based on the advertising spending or (there is a linear relationship between sales revenue and advertising spending)

This p-value $= 0.000445475062877538$ is very small which is less than the significant level $\alpha = 0.05$ suggests that we have a sufficient evidence to reject H_0 and conclude that the regression model is good for prediction the sales revenue based on the advertising spending.