Name (Last, First):	NEU ID:
Kunjangada Arun Sushma	002473132

Assignment 11

NEU_COE_INFO6105_Fall2024

Instructions:

1. For answering **programming questions**, please use Adobe Acrobat to edit the pdf file in two steps **[See Appendix: Example Question and Answer]**:
    a. Copy and paste your R or Python code as text in the box provided (so that your teaching team can run your code);
    b. Screenshot your R or Python console outputs, save them as a .PNG image file, and paste/insert them in the box provided.
    c. Show all work—credit will not be given for code without showing it in action, including a screenshot of R or Python console outputs.
2. To answer **non-programming questions**, please type or handwrite your final answers clearly in the boxes. Show all work - credit will not be given for numerical solutions that appear without explanation in the space above the boxes. **You're encouraged to use R or Python to graph/plot the data and produce numerical summaries; please append your code and screenshot of the outputs at the end of your PDF submission.**
3. **[Total 96 pts = 93 pts + 3 Extra Credit pts]**

Grading Rubric

Each question is worth 3 points and will be graded as follows:

3 points: Correct answer with work shown

2 points: Incorrect answer but attempt shows some understanding (work shown)

1 point: Incorrect answer but an attempt was made (work shown), or **correct answer without explanation (work not shown)**

0 points: Left blank or made little to no effort/work not shown

Reflective Journal [3 pts]

(Copy and paste the link to your live Google doc in the box below)

https://docs.google.com/document/d/1ptEhnYHniNtT1yxDPcvXK7LpJaPGzi80BCSGZZhom7Y/edit?usp=sharing

I. Goodness of Fit Test (12 pts)

Are more babies born on a specific day of the week than others? To determine if the distribution of births across the week happened in different proportions than expected, a researcher took a random sample of 84 births in the year from a local hospital and recorded what day they were born on. The data is given in the following table:

Day of the week	Sun	Mon	Tue	Wed	Thu	Fri	Sat
# of births	6	7	9	14	19	17	12

Based on these data, is it reasonable to conclude that the proportion of births is not the same for all days of the week? Use $\alpha = 0.05$.

Answer:

H_0: The proportion of birth is the same for all days of the week.

H_a: The proportion of birth is not same for all days of the week.

$$\text{Expected frequency} = \frac{\text{Total number of birth}}{\text{Number of days}} = \frac{84}{7} = 12$$

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(6-12)^2}{12} + \frac{(7-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12} + \frac{(19-12)^2}{12} + \frac{(17-12)^2}{12} + \frac{(12-12)^2}{12}$$

$$X^2 = 3 + 2.08 + 0.75 + 0.33 + 4.08 + 2.08 + 0 = \underline{12.32}$$

Degree of freedom (df) = Number of categories $- 1 = 7 - 1 = \underline{6}$.

At $\alpha = 0.05$ with $df = 6$, the critical value is $\underline{12.592}$ from the chi square distribution table

$$X^2_{critical} = 12.592$$

The calculated test statistics $(X^2 = 12.32)$ < the critical value (12.592)

$\therefore$ We fail to reject the null hypothesis.

At $\alpha = 0.05$, we don't have sufficient evidence to conclude that the proportion of births differs across days of the week. While there appears to be some variation in the observed frequencies (with Thursday & Friday having notably more births), This difference is not statistically significant at the 5% level.

II. Chi-Square Test for Homogeneity (12 pts)

A study at a university wanted to see if there was a gender difference with respect to drinking behavior. Two independent random samples of male and female college students at the university asked them to rate their drinking behavior as none, low, moderate, or high. The results are shown in the table below.

	Drinking Level			
	None	Low	Moderate	High
Male	140	478	300	63
Female	180	580	285	40

a) What would be the null and alternative hypotheses to test to see if there was a gender difference with respect to drinking behavior?

Null Hypothesis (H_0): There is no association between gender & drinking behaviour; the distribution of drinking behaviour is the same for males & females. (they are independent)

Alternative Hypothesis (H_a): There is an association between gender & drinking behaviour; the distribution of drinking behaviour differs between males and females. (they are dependent).

b) What would the expected counts be for a male with a moderate drinking level? Show your work!

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}} = \frac{(140+478+300+63)\times(300+285)}{(140+478+300+63+180+580+285+40)}$$

$$= \frac{981 \times 585}{2066} \simeq \boxed{277.78}$$

c) You run the test and get a chi-square statistic of 15.157. What is the p-value?

$X^2 = 15.157$

df = (Number of rows − 1) × (Number of columns − 1) = (2−1) × (4−1) = 3.

For $X^2 = 15.157$ and df = 3, the p-value is approximately 0.0017 using chi-square distribution.

d) What can you conclude at the 5% significance level?

p value (0.0017) < α (0.05), we reject the null hypothesis

We have sufficient evidence to conclude that there is a significant association between gender and drinking behaviour at the 5% significance level.

III. Chi-Square Test and the Follow-Up Analysis (15 pts)

Does the treatment of a stress fracture in a foot affect the success or failure of healing the bone? A recent experiment in a medical journal took four separate random samples of various treatment methods used to treat a fractured foot. In each of these random samples, they recorded whether the patient saw success in the healing of the fracture. A Chi-Square test for Homogeneity was performed and the follow-up analysis is given below.

	Success	Failure
Surgery	54 50.471 0.247	12 15.529 0.802
Weight-Bearing Cast	41 51.235 2.045	26 15.765 6.645
Non-Weight Bearing Cast for Less Than 6 Weeks	17 19.118 0.235	8 5.882 0.762
Non-Weight Bearing Case for 6 Weeks	70 61.176 1.273	10 18.824 4.136

Key:
Observed
Expected
Contribution

a) What would be the null and alternative hypotheses in this situation?

Null Hypothesis (H_0): The success or failure of healing a fractured foot is independent of the treatment method used.
Alternative Hypothesis (H_a): The success or failure of healing a fractured foot depends on the treatment method used.

b) Show how the value of "15.529" under "surgery, failure" was obtained.

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}} = \frac{66 \times 56}{238} \approx 15.529$$

Row total for surgery $= 54 + 12 = 66$
Column total for failure $= 12 + 26 + 8 + 10 = 56$
Grand total $= 54 + 12 + 41 + 26 + 17 + 8 + 70 + 10 = 238$

c) What would be the chi-square statistic and p-value for this test?

(Chi-square statistic)

$$x^2 = \sum \frac{(O-E)^2}{E} = 0.247 + 0.802 + 2.045 + 6.645 + 0.235 + 0.762 + 1.273 + 4.136$$
$$= 16.145$$

Degree of freedom (df): (Number of rows -1) $\times$ (Number of columns -1) $= (4-1) \times (2-1) = 3$

From chi-square table or statistical table, the p-value for $x^2 = 16.145$ with $df = 3$ is approximately 0.001.

d) What is your conclusion in the context of the problem, at the 1% significance level?

The p-value (0.001) < α (0.01)

$\therefore$ We reject the null hypothesis

At the 1% significance level, there is sufficient evidence to conclude that the success or failure of healing a fractured foot depends on the treatment method used.

e) Identify the two largest contributions to the chi-square statistic. What do these contributions imply in the context of the problem?

The two largest contributors are:

1. Weight - Bearing Cast, Failure: 6.645
2. Non-Weight Bearing Cast for 6 weeks, Failure: 4.136

Implications:

- The "weight-bearing cast; failure" indicates a much higher failure rate than expected for this.
- The "Non-weight Bearing Cast for 6 weeks, failure" shows fewer failures than expected for this treatment.

IV. Chi-Square Test for Association/Independence (21 pts)

Are the smoking habits of college students related to their parents' smoking habits? Below are data from a survey of students gathered across five different colleges.

	Both Parents Smoke	One Parent Smokes	Neither Parent Smokes
Student Smokes	300	316	88
Student Does Not Smoke	1280	1723	1068

a) How can the data be gathered so that we would perform a chi-square test for homogeneity?

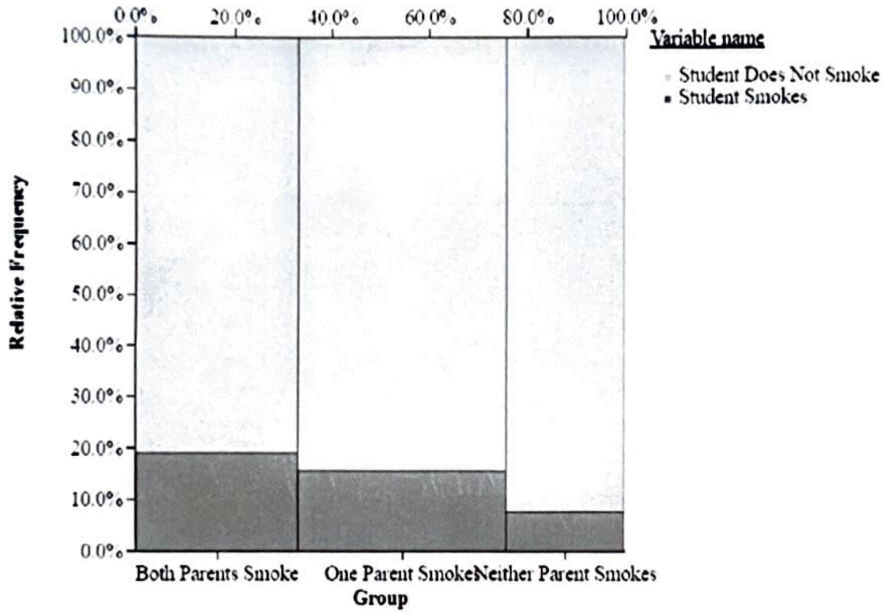
Data Collection: Seperate random sample should be taken from each parent smoking category: "Both Parents Smoke", "One Parent-Smokes", "Neither Parent Smokes".
Observation: Within each sample, the smoking status of the Students (smokes/doesn't smoke) is recorded.
Independence: The samples must be independent of one another.
Objective: This approch tests whether the distribution of students smoking habits (smokes or does not smoke) is the same across the categories of parents smoking habits.

b) How can the data be gathered so that we would perform a chi-square test for association/independence?

Data Collection: Data should be gathered from a single random sample of a population
Variables Measured: For each individual in the sample, two categorical variables should be recorded:
- Parents smoking habits (Both Parents Smoke, One Parents Smokes, Neither Parent Smokes)
- Students smoking status (Student Smokes, Student Does not Smoke).

Objective: The ~~global~~ goal is to test whether there is an association or dependence between the two categorical variables (parents smoking habits and student smoking habits.

c) Below is a mosaic plot of the data. In a mosaic plot, the width of each vertical bar represents the relative size of each parent smoking habit category. Are the categories (Both Parents Smoke, One Parent Smokes, Neither Parent Smokes) equally represented in the survey? Explain.



No, the categories are not equally represent in the survey. Looking at the mosaic plot, the width of the bars is different for each category. "One Parent Smokes" has the widest bar, followed by "Both Parent Smoke", followed with "Neither Parent Smokes" has the narrowest bar.

d) What would be the null and alternative hypotheses from a chi-square test for association/independence?

Null Hypothesis (H_0): There is no association between parents smoking habits and students smoking habits (they are independent)

Alternative Hypothesis (H_1): There is an association between parents smoking habits and students smoking habits (they are dependent).

e) Using you calculator, find the expected counts, chi-square statistic, degrees of freedom, and the p-value.

$X^2 = \sum \frac{(O-E)^2}{E}$ $E = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$ df = (Number of Rows − 1) × (Number of Columns − 1)

Expected Counts	Both Parents Smoke	One Parent Smokes	Neither Parent Smokes
Student Smokes	232.947	300.619	170.434
Student Does Not Smoke	1347.053	1738.381	985.566

χ^2 = 70.32

df = 2

P-value = 0.0000000000000000l

f) Based on your p-value, what can you conclude?

Since p value < 0.0001 ie, p-value is much smaller than significance level, we reject the null hypothesis. There is a strong evidence of an association between parents smoking habits and students smoking habits.

g) Which cell in the table contributed the most to the chi-square statistic? How does this information expand on your solution in part (f)?

- The cell with largest contribution to chi-square statistic is the cell for "Student Smokes" when "Neither Parent Smokes" because the observed count (88) significantly deviates from expected count (170.434)

- This suggest that having non smoking parents is associated a lower likelihood of the student smoking, which expands on our conclusion by showing specifically where the evidence of association lies.

V. Confidence Intervals for Slopes (12 pts)

Your local Barnes and Nobel store has hired a new social media manager, and they are tracking data on how many books were sold each week from a random sample of new hires at the store.

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number	12	21	18	13	28	20	29	35	26	20	36	40	28	52	32	32	55	56	49	60

a) Use R or Python to determine the LSRL for the data above:

```
[2]: import numpy as np
import scipy.stats as stats

# Data
weeks = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20])
books_sold = np.array([12, 21, 18, 13, 28, 20, 29, 35, 26, 20, 36, 40, 28, 52, 32, 32, 55, 56, 49, 60])

# Calculate the slope and intercept of the regression line
slope, intercept, r_value, p_value, std_err = stats.linregress(weeks, books_sold)

print("Least Squares Regression Line (LSRL):")
print(f"y = {intercept:.2f} + {slope:.2f}x")

Least Squares Regression Line (LSRL):
y = 10.68 + 2.14x
```

b) Interpret the slope of the regression in the context of the problem.

The slope of ($b_1 = 2.14$) means that each additional week (x) the number of books sold increases on average by 2.[redacted]4 books.

c) Assuming the conditions needed for inference were checked and met, use the data below to construct and interpret a 90% T-Interval for the true regression equation slope. What does this interval suggest about how many books the new hires are selling?

n	s	s_x	$\bar{x}$	r^2
20	7.651	5.916	10.5	0.742

$$B_1 = r\left(\frac{S_y}{S_x}\right) = 0.742 \times \left(\frac{7.651}{5.916}\right) = 0.960$$

$$SE_{\beta_1} = \frac{S_y}{S_x} \times \sqrt{\frac{1-r^2}{n-2}}$$

$r^2 = 0.74^2 = 0.5506$ $\quad n-2 = 20-2 = 18$

$1-r^2 = 1-0.5506 = 0.4494$ $\quad \frac{1-r^2}{n-2} = \frac{0.4494}{18} = 0.02497$

$\sqrt{0.02497} = 0.158$

$$SE_{B_1} = \frac{7.651}{5.916} \times 0.158 = 0.204$$

For 90% Confidence interval, $t^* = 1.734$

CI = $B_1 \pm t^* \times SE_{\beta_1} = 0.960 \pm 1.734 \times 0.204$

$= 0.606, 1.314$

This suggest on an average the number of books sold by new hires increases between 0.606 and 1.314 books per week as weeks go by.

VI. Significance Testing for Slope (21 pts)

Data was gathered on a random sample of 25 high school sophomores at your school. Each sophomore's score on a standardized chemistry exam and their score on a standardized reading exam was taken. School officials wanted to see if a sophomore's reading score (in points) could help predict their chemistry score (in points).

a) Identify the value of the standard deviation of the residuals and interpret this value in the context of the problem.

The standard deviation of residuals is $S = 25.83$ points. This means that on an average the actual chemistry score deviate from predicted score by about 25.83 points.

b) Identify the value of the estimated standard deviation of the slope and interpret this value in the context of the problem.

The standard error of the slope (SE Coef for Reading) is 0.0351. This represent standard deviation of sampling distribution of the slope estimate. It tells us how much we did expect the sample slope of 0.731 to vary from sample to sample of 25 sophomores.

c) What would be a null and alternative hypothesis for a hypothesis test based on the computer output above?

$H_0: \beta = 0$ (reading score do not help predict chemistry score)
$H_1: \beta \neq 0$ (reading score do help predict chemistry score).

d) What would the value of the parameter, β, measure in this test?

β represents the true slope of the population regression line.
It measures the average change in chemistry score for each one-point increase in the reading score.

e) Explain how a t-test statistic of 20.83, and a p-value of <0.0001 were obtained.

$$t = \frac{\text{Coef (Slope)}}{\text{SE Coef (Slope)}} = \frac{0.731}{0.0351} \approx 20.83$$

P(value) (< 0.0001) is computed t value 20.83 with $(n-2)$ $\overset{25-2}{}$ 23 degrees of freedom, indicates strong evidence against H_0.

f) Create a 99% confidence interval using the data from the computer output above.

$CI = \text{Slope} \pm t^* \cdot \text{SE Coef (slope)}$

99% confidence interval $df = 23$ $t^* = 2.807$

$CI = 0.731 \pm 2.807 \times 0.0351$

$CI = 0.731 \pm 0.0985$

$CI = (0.6325, 0.8295)$

g) What would be the correct conclusion of your significance test in the context of the problem? Explain how your results in (e) and (f) agree on this conclusion.

- Since p-value is extremly small (< 0.0001), we reject the null hypothesis. This suggest that there is strong evidence the reading score are significantly associated with chemistry score.
- The confidence interval for the slope does not contain 0, further confirming that the slope is significantly different from zero. Both p-value & confidence interval lead to the same conclusion.

THE END