Introductory Statistics

Midterm # 3

Learning Objectives covered in this midterm:

Question # 1 LOs:

- Complete a one-way ANOVA hypothesis test
- Complete pair-wise comparisons for ANOVA
- Calculate a confidence interval and p-value for pair-wise comparisons and explain what it means

Question # 2 LOs:

- Complete a chi-square test for goodness of fit and write its conclusion in context of the problem

Question # 3 LOs: Choose one out of two!

- Complete a chi-square test of homogeneity
- Write the conclusion of a chi-square test of homogeneity in context of the problem
- Complete a chi-square test of independence
- Write the conclusion of a chi-square test of independence in context of the problem

Question # 4 LOs:

- Perform a test for significance of slope and interpret the results
- Check the conditions that are necessary to perform a test for significance of slope
- Understand what is measured by SSRegression, SSResiduals, and SSTotal in a regression context
- Discuss the factors that affect the value of F-statistics in a regression context
- Find and interpret the confidence interval for the mean response
- Find and interpret the prediction interval for an individual response

Question # 5 LOs:

- Write and describe a multiple linear regression model equation
- Calculate and describe the unadjusted coefficient of determination
- Know what an indicator variable is
- Know what an interaction term is and use a scatterplot to understand the interaction effects

Question # 6 LOs:

- Complete a randomization test involving a difference in proportions

Introductory Statistics - Midterm # 3

Name: Date:

1. We would like to investigate the effectiveness of antidepressant medication and how it is related to the severity of the depression. Based on pretreatment depression scores, patients were divided into four groups based on their level of depression. After receiving the antidepressant medication, depression scores were measured again and the amount of improvement was recorded for each patient in the table below.

Low Moderate	High Moderate	Moderately Severe	Severe
2.3	0.9	1.1	2.9
0.9	0.6	1.5	3.4
1.5	1.4	1.9	3.9
1.3	2.2	0.6	2.7
0	2.6	0.8	3.3
3	2.3	2.6	4

   a. Explain why a one-way ANOVA should be considered for this situation.

   b. State the null and alternative hypotheses for this scenario.

   c. Based on the information of the study and the data set provided above, do you have any concerns about testing with a one-way ANOVA? Explain your answer.

d. Use the statistical tool (https://lumen-learning.shinyapps.io/anova/) and fill in the summary table for the ANOVA test and report its P-value.

	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F-statistics
Group				
Error				
Total				

P-value = _____

e. What is your conclusion in context of the problem? Use a significance level of 5%.

f. Use the Tukey method adjusted confidence interval and P-value to conduct multiple comparisons to identify the differences between the four groups. Which mean(s) are different from one another? Explain.

2. A six-sided die has been weighted so that they're more likely to roll a six. The claimed proportions are the following:

Appeared	1	2	3	4	5	6
Proportion	0.06	0.11	0.11	0.11	0.11	0.50

You decided to test the die and roll the die 100 times into a rectangular, felt-lined dice box and got the following results:

Appeared	1	2	3	4	5	6
Frequency	2	7	5	6	5	75

Let's conduct the goodness of fit test to see if the distribution of the die is different than in the claimed proportions.

Statistical tool: https://lumen-learning.shinyapps.io/chisquaredtest/

a. How could we formulate this question as two testable hypotheses? State the hypotheses.

b. Does the sample satisfy the conditions for a chi-square goodness of fit test? Explain.

c. Use the statistical tool and report the test statistic χ^2, degree of freedom, and the P-value. Then, state your conclusion in context of the problem.

3. Pick one of the following scenarios to complete. Circle the scenario you want me to grade.

Scenario A: Education level

Americans' concern about race relations has increased among all major racial/ethnic groups since 2014 amid national attention to the killing of unarmed Black men and women in encounters with police. Gallup has monitored Americans' concern about race relations annually as part of a March survey that asks respondents to rate their concerns about numerous issues facing the country.[1]
When thinking of race relations, Americans may be focused generally on how different racial groups in the country relate to each other, such as Black and White adults or Asian and White adults. But, they could also be factoring in other demographics, such as education level and age.

Statistical tool: https://lumen-learning.shinyapps.io/chisquaredtest/

Below is the contingency table reported by Gallup for their March 2022 survey regarding their level of race relations worry and their level of education.

Worry	College Grad	Some College	HS Grad or Less
A great deal	114	79	103
A fair amount	86	56	65
Only a little	41	41	42
Not at all	24	26	44

Let's investigate if there is an association between race relations and the education level.

a. Which χ^2-test are we performing and why?

b. State its null and alternative hypotheses.

[1] https://news.gallup.com/poll/392705/concern-race-relations-persists-floyd-death.aspx

c. Are the conditions met to perform a chi-square test? Explain.

d. Use the statistical tool and report the test statistic χ^2, degree of freedom, and the P-value. Then, state your conclusion in context of the problem. Use a significance level of 5%.

e. Interpret and compare the standardized residuals for individuals with college degrees and individuals with high school degrees or less.

f. Based on the results of this test alone, can you assure someone that if they have more education, they will worry more about race relations? Explain.

g. Give an example of a lurking variable that could arise when considering the association of these two variables.

Scenario B: Age

Americans' concern about race relations has increased among all major racial/ethnic groups since 2014 amid national attention to the killing of unarmed Black men and women in encounters with police. Gallup has monitored Americans' concern about race relations annually as part of a March survey that asks respondents to rate their concerns about numerous issues facing the country.[2]
When thinking of race relations, Americans may be focused generally on how different racial groups in the country relate to each other, such as Black and White adults or Asian and White adults. But, they could also be factoring in other demographics, such as education level and age.

Statistical tool: https://lumen-learning.shinyapps.io/chisquaredtest/

Below is the contingency table reported by Gallup for their March 2022 survey regarding their level of race relations worry and their age.

Worry	18-34 years old	35-54 years old	55+ years old
A great deal	76	99	113
A fair amount	44	66	96
Only a little	29	32	62
Not at all	33	35	20

Let's investigate if the different worries have the same distribution when it comes to their age.

a. Which χ^2-test are we performing and why?

b. State its null and alternative hypotheses.

[2] https://news.gallup.com/poll/392705/concern-race-relations-persists-floyd-death.aspx

c. Are the conditions met to perform a chi-square test? Explain.

d. Use the statistical tool and report the test statistic χ^2, degree of freedom, and the P-value. Then, state your conclusion in context of the problem.

e. Discuss the result of your test in terms of practical significance and statistical significance.

f. Interpret and compare the standardized residuals for individuals who said that they worry a great deal when it comes to race relations.

4. Let's conduct a hypothesis test to find out whether or not the Internet and Facebook usage from 32 countries in 2012 have a significant linear relationship, in addition to measuring the evidence of an association between Internet and Facebook usage.
   Select the **Internet & Facebook** data set within the statistical tool:
   https://lumen-learning.shinyapps.io/linear_regression/

   a. First, let's conduct a hypothesis test for significance of slope. State its null and alternative hypotheses.

   b. Is a line a reasonable model for the relationship between the two variables? Explain.

   c. Using the residual plot, would it be appropriate to do a hypothesis test for significance of slope? Explain.

   d. Use the output of the statistical tool to obtain the test statistic and P-value. State the conclusion for the hypothesis test for significance of slope in context of the problem.

   e. Obtain the 95% confidence interval for slope and use it to determine if the slope of a regression line is statistically significant. Justify your answer accordingly.

f. Use ANOVA F-test to measure the evidence of an association between Internet and Facebook usage. Fill out the table below. Then, use the statistical output to make your decision and state your conclusion for the hypothesis test in context of the problem.

Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F value	P-value
Regression					
Residuals					
Total					

Conclusion:

g. Is there a connection between the hypothesis test for significance of slope and the ANOVA F-test, which measure the association between the two variables? Explain.

h. Fill-in-the-blanks.

i. We are 95% confident that the mean percentage of Facebook Use when the average Internet Use is 48% is between _____ and _____.

ii. We are 95% confident that the percentage of Facebook Use when the Internet Use of a country is 48% is between _____ and _____.

5. A dataset provides county level data on education (measured as the percentage of residents aged at least 25 in the county who had at least a high school degree) and crime rate (measured as the number of crimes in that county in the past year per 1000 residents). Another variable measured for these counties is urbanization, measured as the percentage of county residents living in metropolitan areas (variable name: Urbanization (Percent)). The variable Urbanization (Categorical) groups this percentage into categories "rural" (counties with no urbanization), "mixed" (between 1% and 50% urbanization) and "urban" (more than 50% urbanization).
We would like to predict the crime rate using the county level data on education and the percentage of county residents living in metropolitan areas.

   a. Identify the response variable and explanatory variables.

   b. Using the following results, write the multiple linear regression equation.

	Estimate
Intercept	59.1
Education	-0.583
Urbanization	0.683

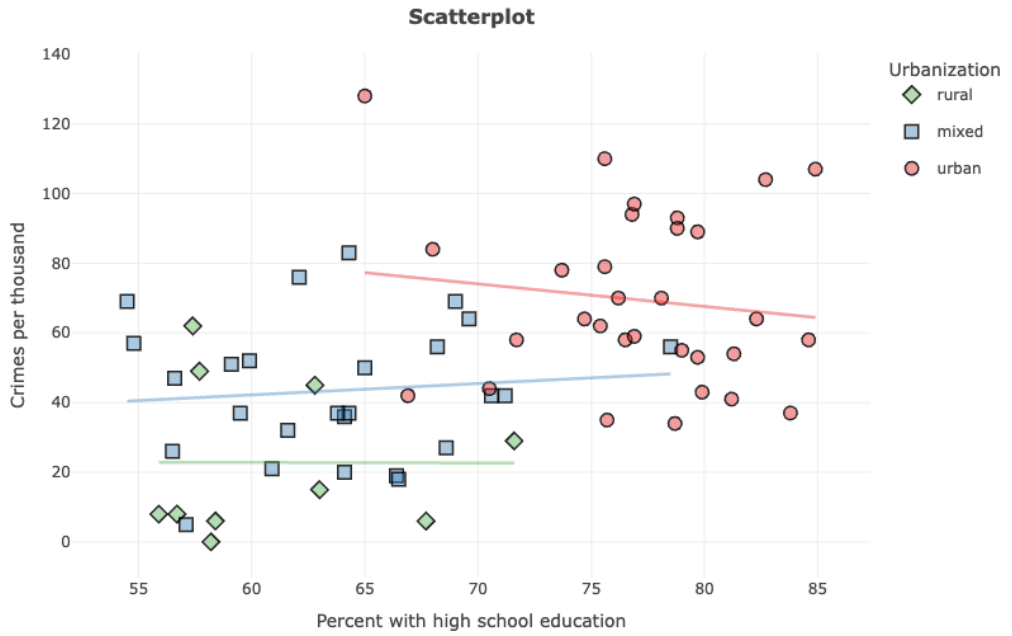
   c. What is the interpretation of the coefficient for the variable education in the context of the dataset?

   d. What is the interpretation of the coefficient for the variable urbanization in the context of the dataset?

   e. The unadjusted coefficient for determination value for this model is 47.1%. Interpret this value.

f. We are interested in adding the explanatory variable of Urbanization (categorical) to our prediction model. How many indicator variables do we need to create in order to add the variable to the model? Give an example on how you would label the indicator variable in this context.

g. Do you think an interaction term that looks at the relationship between Education and Urbanization would be appropriate to investigate in a multiple linear regression model? From the regression lines, what can you say regarding the effect of education and urbanization on the crime rate? Explain.



6. We would like to investigate whether or not the COVID-19 Pfizer vaccine helps prevent COVID-19 infection. The contingency table is provided below and can also be found under "COVID-19 Vaccine (Pfizer)" in the statistical tool.

	Yes	**No**
Pfizer Vaccine	8	21712
Placebo	162	21566

Statistical tool: https://lumen-learning.shinyapps.io/association_categorical/

a. What is the difference in proportions for the "Yes" category?

b. State its null and alternative hypotheses.

c. Using the statistical software to generate 1000 simulations. Find and interpret the 95% confidence interval.

d. Is there enough evidence whether or not COVID-19 Pfizer vaccine helps prevent COVID-19 infection? Explain your reasoning based on the simulation results, including a discussion of the purpose of the simulation process and what information it revealed to help you answer this research question.