Introductory Statistics - Midterm # 3
Answer Key

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.
   Because there are more than two groups (four groups based on the level of depression) and we want to compare the means of these multiple groups to determine if there are any significant differences among them regarding the improvement in depression scores after taking antidepressant medication.

   b. State the null and alternative hypotheses for this scenario.
   Null Hypothesis: There is no significant difference in the mean improvement scores among the groups representing different levels of depression after taking the antidepressant medication. Mathematically, this can be expressed as
   $H_0: \mu_{LM} = \mu_{HM} = \mu_{MS} = \mu_S$

   Alternative Hypothesis: At least one group has a different mean improvement score compared to the others.

   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.
      - Data: Factor of interest is categorical data, with a numeric response variable, and we are interested in the mean of the response variable, so, this condition is met.

- Independent: The groups are independent, randomly assigned experimental groups, so an ANOVA is appropriate.
- Variability: The sample sizes are equal and the largest standard deviation is no more than two times the smallest standard deviation. This condition is met.

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	3	15.41	5.137	7.882
Error	20	13.04	0.6518	
Total	23	28.45		

P-value = _____0.0012_____

e. What is your conclusion in context of the problem? Use a significance level of 5%.
With a p-value of 0.0012, which is less than the significance level (α) of 0.05, there is sufficient evidence to reject the null hypothesis.
In the context of the study, this means there are significant differences in the effectiveness of the antidepressant medication across the different levels of depression severity. In other words, the improvement in depression scores after taking the medication varies significantly among patients with different initial levels of depression.

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.
- The means of Low Moderate and Severe is significantly different from one another because the CI (-3.17, -0.56) doesn't include 0. Additionally, the P-value is 0. Since all of the values are negative the mean of the Severe group is larger than the mean in the Low Moderate group, which indicates that the improvement of the Severe group is much larger after receiving antidepressant medication in comparison to the Low Moderate group.
- The means of High Moderate and Severe is also significantly different from one another because the CI (-3, -0.40) doesn't include 0. Additionally, the P-value is 0.01. Since all of the values are negative the mean of the Severe group is larger than the mean in the High Moderate group as well, which indicates that the improvement of the Severe group is much larger after receiving antidepressant medication in comparison to the High Moderate group.

- Lastly, the means of Moderately Severe and Severe is also significantly different from one another because the CI (-3.25, -0.65) doesn't include 0. Additionally, the P-value is 0. Since all of the values are also negative the mean of the Severe group is larger than the mean in the Moderately Severe group, which indicates that the improvement of the Severe group is much larger after receiving antidepressant medication in comparison to the Moderately Severe group.

From these comparisons, overall, it seems that the improvement of the Severe group is much larger after receiving antidepressant medication in comparison to the other group within this experiment.

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.
Null Hypothesis: The observed distribution of the die rolls follows the claimed proportions, indicating no significant difference between the observed and expected outcomes.

Alternative Hypothesis: The observed distribution of the die rolls does not follow the claimed proportions, suggesting a significant difference between the observed and expected outcomes.

b. Does the sample satisfy the conditions for a chi-square goodness of fit test? Explain.
   - Random: Observed counts come from an experiment, this condition is met.
   - 10%: We can assume that the sample size is less than a tenth of the population size.

- Large Sample: The expected counts (6, 11, 11, 11, 11, 50) are greater than five, therefore this condition is met.

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.
Test statistic = 25.44
df = 5
P-value < 0.0001
Conclusion: With a p-value of 0.0001, which is less than the significance level (α) of 0.05, there is strong evidence to reject the null hypothesis. This suggests that the observed outcomes of the die rolls significantly deviate from the expected frequencies. In other words, the weighted die is likely affecting the outcomes, indicating a departure from the claimed proportions.

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.

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

a. Which χ^2-test are we performing and why?
To investigate the association between race relations worry and education level, we can perform a Chi-square test for Independence. This test will help determine whether there is a statistically significant relationship between the two categorical variables: worry level (in categories "A great deal," "A fair amount," "Only a little," and "Not at all") and education level (in categories "College Grad," "Some College," and "HS Grad or Less").

b. State its null and alternative hypotheses.
Null Hypothesis: There is no association between race relations worry and education level.
Alternative Hypothesis: There is an association between race relations worry and education level.

c. Are the conditions met to perform a chi-square test? Explain.
- The data represent the counts for two categorical variables measured for individuals in one sample from one population.
- The data was obtained through a random survey, this condition is met.
- The expected cell count is much larger than five. This condition is met.

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%.
Test statistic = 11.13
df = 6
P-value = 0.0845
Conclusion: With a p-value of 0.0845 at a significance level (α) of 0.05, we fail to reject the null hypothesis. In the context of the study, this suggests that there is not enough evidence to conclude that there is a significant association between individuals' level of worry about race relations and their education level. The worry about race relations does not appear to be significantly different across different education levels according to the data collected in the Gallup survey.

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

	CG	SC	HS
great	0.82	-0.66	-0.2
fair	1.7	-0.37	-1.4
little	-0.94	1.4	-0.35
not	-2.4	-0.083	2.5

- For those who worry a great deal about race relations, the standardized residual indicates that college graduates are more concerned about race relations than expected based on their education level.
- For those who worry a fair amount, the standardized residual also indicates that college graduates are more concerned about race relations than expected based on their education level.
- Lastly, for those who are not at all worried, the observed value of High School Graduate or less is higher than expected, suggesting more engagement with the issue of race relations than expected.

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.
No, the results of this chi-square test do not provide assurance that having more education directly leads to worrying more about race relations. While the chi-square test did not find a statistically significant association between education level and concern about race relations in this specific survey, this does not imply a causal relationship or a guarantee that education level determines one's worry about race relations.

g. Give an example of a lurking variable that could arise when considering the association of these two variables.
There are many lurking variables such as the respondents' exposure to diverse environments or communities, socioeconomic status, geographic location, etc.

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

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

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?
Chi-square homogeneity test is used to determine whether the distribution of a categorical variable is the same across different groups or categories of another variable. In this scenario, we are comparing the distribution of worries about race relations across different age groups, which fits the chi-square homogeneity test.

b. State its null and alternative hypotheses.
Null Hypothesis: The distribution of worries about race relations is the same across different age groups.
Alternative Hypothesis: The distribution of worries about race relations is different across different age groups.

c. Are the conditions met to perform a chi-square test? Explain.
- Independence: The responses of individuals in each age group (18-34, 35-54, and 55+) must be independent of each other. This means that individuals within each age group should be randomly selected or assigned, and their responses should not be influenced by the responses of others. Therefore, this condition is met.
- Expected Frequency: The expected frequency in each cell of the contingency table is at least 5. This condition is met.

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.
Test statistic = 21.31
df = 6
P-value = 0.0016
Conclusion: With a P-value of 0.0016 at the 0.05 significance level, we reject the null hypothesis. This suggests that there is a significant difference in the distribution of worries about race relations among different age groups according to the data collected in the March 2022 survey.

e. Discuss the result of your test in terms of practical significance and statistical significance.
In terms of statistical significance, the chi-square test indicates that there is a significant difference between worries about race relations and age groups, as evidenced by the low p-value (0.0016) which falls below the commonly used alpha level of 0.05. This means that the differences in worry levels across age groups are unlikely to have occurred by random chance alone, suggesting a genuine relationship in the population sampled.

While the results are statistically significant, the practical importance of this difference depends on the context and the specific goals of the study. Understanding that different age groups have varying levels of concern about race relations is valuable for social and demographic analysis. It might be essential to delve deeper into the context, such as understanding the reasons behind these worries and their implications on social harmony and policies.

f. Interpret and compare the standardized residuals for individuals who said that they worry a great deal when it comes to race relations.
- For the age group 18-34: A standardized residual of 0.29 suggests a slightly higher observed frequency for those who worry a great deal, but it's not significantly deviating from the expected values.
- For the age group 35-54: With a standardized residual of 0.69, there is a higher observed frequency of individuals worrying a great deal compared to what was expected for this age group.
- For the age group 55+: A standardized residual of -0.91 indicates a lower observed frequency of individuals worrying a great deal. While negative, it's not significantly deviating from the expected values.

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.
Null Hypothesis: The slope of the regression line, representing the relationship between Internet and Facebook usage, is equal to zero.
$H_0: \beta_1 = 0$
Alternative Hypothesis: The slope of the regression line, representing the relationship between Internet and Facebook usage, is not equal to zero.
$H_A: \beta_1 \neq 0$

b. Is a line a reasonable model for the relationship between the two variables? Explain.
The coefficient of determination (R^2) tells us the proportion of the variance in the dependent variable (Facebook usage) that is predictable from the independent variable (Internet usage). In this case, R^2 = 37.7%, indicating that approximately 37.7% of the variation in Facebook usage can be explained by Internet usage. We might need u=to use residual plot to see if there is a pattern to understand if a line is a reasonable model.

c. Using the residual plot, would it be appropriate to do a hypothesis test for significance of slope? Explain.
The residuals show NO pattern (e.g., a curve), it suggests that a linear model is the best fit. Therefore, it is appropriate to do a hypothesis test for significance of slope.

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.
Test statistic = 4.26
P-value = 0.0002
Conclusion: Reject the null hypothesis. This means that the relationship between Internet and Facebook usage in the 32 countries in 2012 is statistically significant. In other words, the data provides convincing evidence that there is a significant linear relationship between the two variables.

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.
95% confidence interval: (0.229, 0.649)
Because 0 is not included in this interval, this means that there is a relationship between Internet and Facebook usage in the 32 countries in 2012. In other words, the data provides convincing evidence that there is a significant linear relationship between the two variables.

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	1	2991.7	2991.7	18.2	0.00
Residuals	30	4942.6	164.8		
Total	31	7934.3			

Conclusion: With a p-value of 0, the ANOVA F-test provides overwhelming evidence to reject the null hypothesis. This means that there is a significant association between Internet and Facebook usage in the 32 countries studied in 2012.

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.
Yes, there is a connection between the hypothesis test for the significance of the slope in linear regression and the ANOVA F-test. When we are analyzing the relationship between Internet usage and Facebook usage across 32 countries in 2012, the hypothesis test for the significance of the slope in linear regression evaluates whether there is a significant linear relationship between the amount of Internet usage and Facebook usage in these countries. On the other hand, the ANOVA F-test, would also assess the association between Internet and Facebook usage but in a broader sense. It checks if there are any significant differences in Facebook usage across different levels of Internet usage.

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 ___130___ and _____308_____.

   ii. We are 95% confident that the percentage of Facebook Use when the Internet Use of a country is 48% is between _____126____ and _____312_____.

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.
   The response variable is the crime rate, measured as the number of crimes in the county in the past year per 1000 residents.
   The two variables, education rate and urbanization percentage, are the explanatory variables used to predict the crime rate.

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

	Estimate
Intercept	59.1
Education	-0.583
Urbanization	0.683

Crime Rate = 59.1 -0.583(Education) + 0.683(Urbanization)

c. What is the interpretation of the coefficient for the variable education in the context of the dataset?
For every one percentage point increase in the education rate (meaning more residents with at least a high school degree), the predicted crime rate decreases by 0.583 crimes per 1000 residents on average, assuming all other factors such as urbanization percentage are constant.

d. What is the interpretation of the coefficient for the variable urbanization in the context of the dataset?
For every one percentage point increase in the urbanization rate (meaning more residents living in metropolitan areas), the predicted crime rate increases by 0.683 crimes per 1000 residents on average, assuming all other factors such as education rate are constant.

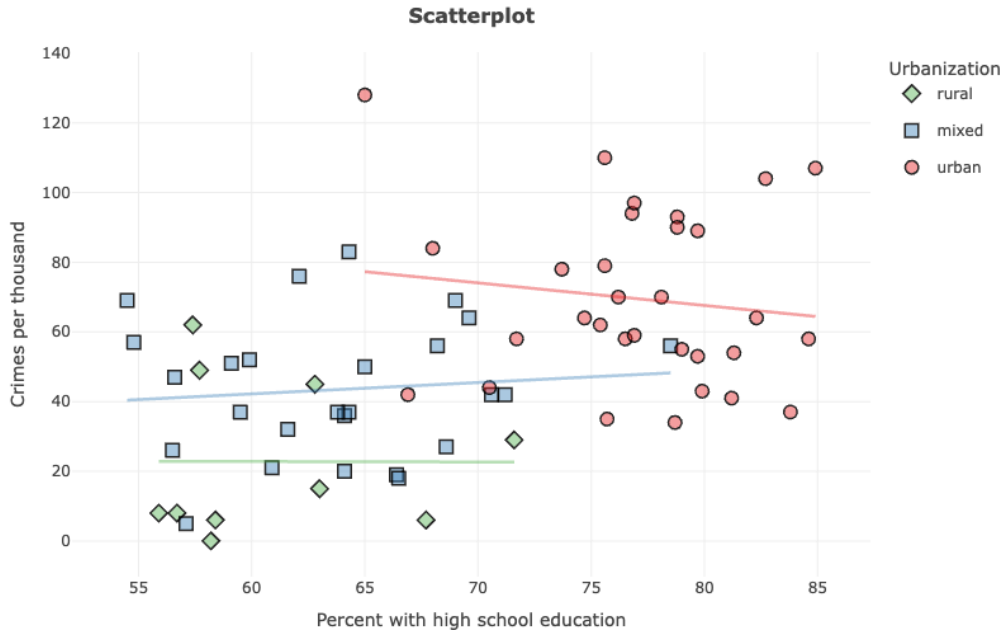
e. The unadjusted coefficient for determination value for this model is 47.1%. Interpret this value.
Approximately 47.1% of the variability in the crime rate across counties can be accounted for by differences in education rates and urbanization rates.

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.
To add the categorical variable "Urbanization (categorical)" to the prediction model, you would need to create two indicator variables because there are three categories: "rural," "mixed," and "urban."

For example, if you create two indicator variables, you could label them as follows:

- Indicator variable for "mixed" urbanization: This variable would take the value 1 if the county is in the "mixed" urbanization category and 0 otherwise.
- Indicator variable for "urban" urbanization: This variable would take the value 1 if the county is in the "urban" urbanization category and 0 otherwise.

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.



Yes, investigating an interaction term between Education and Urbanization can be very appropriate in a multiple linear regression model, especially if you suspect that the effect of Education on the crime rate might differ based on the level of Urbanization.
Because the regression lines are not parallel, it indicates that the effect of Education on the crime rate depends on the level of Urbanization. In other words, the relationship between Education and crime rate differs for different levels of Urbanization. This is precisely what an interaction term in a regression model captures.

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?
p_Pfizer - p_Placebo = 0.000368 - 0.00746 = -0.00709

b. State its null and alternative hypotheses.
Null Hypothesis: There is no significant difference in the proportions of COVID-19 infection between individuals who received the Pfizer vaccine and those who received the placebo.
Alternative Hypothesis: There is a significant difference in the proportions of COVID-19 infection between individuals who received the Pfizer vaccine and those who received the placebo.

c. Using the statistical software to generate 1000 simulations. Find and interpret the 95% confidence interval.
Answers may vary depending on the generated 1000 simulations.
Sample answer:
95% CI: (-0.00823, -0.00596)
Interpretation: We are 95% confident that the true difference in proportions between individuals who received the Pfizer vaccine and those who received the placebo, in terms of getting infected with COVID-19, falls between approximately -0.00823 and -0.00596.

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.
The simulation process can be used to estimate confidence intervals for various parameters, such as the difference in proportions between vaccinated and placebo groups. These intervals provide a range of values within which the true population parameter is likely to lie. Simulations allow researchers to study the variability in outcomes, understand the potential variability in the real-world, and make informed decisions.

In the context of the given 95% confidence interval (-0.00823, -0.00596) obtained from 1000 simulations, the negative values in the interval indicate that, on average, there is a lower proportion of COVID-19 infections among individuals who received the Pfizer vaccine compared to those who received the placebo. The interval does not include zero, suggesting a statistically significant difference in infection rates between the two groups.

This result suggests that the Pfizer vaccine is associated with a statistically significant reduction in COVID-19 infections compared to the placebo, based on the given data and simulation results.