Final Exam
Solutions to the most important exercises done in class
MATH 338 - Prof. Guerrero

Question 1: [*6.37 - Offshore drilling, Part III.*] The table below summarizes a data set we first encountered in Exercise 6.23 that examines the responses of a random sample of college graduates and non-graduates on the topic of oil drilling. Complete a chi-square test for these data to check whether there is a statistically significant difference in responses from college graduates and non-graduates.

	College Grad	**Non-Grad**
Support	154	132
Oppose	180	126
Do Not Know	104	131
Total	438	389

Solution: To test whether there is a statistically significant difference in responses from college graduates and non-graduates, we perform a chi-square test:

Let $O_{i,j}$ be the observed frequency and $E_{i,j}$ be the expected frequency for cell i, j (where i represents the i^{th} row and j the j^{th} column). The chi-square test statistic is given by:

$$\chi^2 = \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

where the sum is taken over all cells of the table. The degrees of freedom for the test are given by $df = (R-1) \times (C-1)$, where R is the number of rows and C is the number of columns in the table.

The hypotheses are as follows:

- H_0: The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California.
- H_A: Opinions regarding the drilling for oil and natural gas off the coast of California have an association with earning a college degree.

Before calculating the test statistic we should check that the conditions are satisfied.

1. **Independence:** The samples are both random, unrelated, and from less than 10% of the population, so independence is reasonable.
2. **Sample size:** Under H_0 the expected counts can be calculated as follows:

$$E_{1,1} = \frac{438 \times (154+132)}{827} = 151.5 \qquad E_{1,2} = \frac{389 \times (154+132)}{827} = 134.5$$
$$E_{2,1} = \frac{438 \times (180+126)}{827} = 162.1 \qquad E_{2,2} = \frac{389 \times (180+126)}{827} = 143.9$$
$$E_{3,1} = \frac{438 \times (104+131)}{827} = 124.5 \qquad E_{3,2} = \frac{389 \times (104+131)}{827} = 110.5$$

All expected counts are at least 5.

The chi-squared statistic, the degrees of freedom associated with it, and the p-value can be calculated as follows:

$$\begin{aligned}\chi^2 = \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} &= \frac{(154 - 151.5)^2}{151.5} + \frac{(132 - 134.5)^2}{134.5} + \frac{(180 - 162.1)^2}{162.1} \\ &\quad + \frac{(126 - 143.9)^2}{143.9} + \frac{(104 - 124.5)^2}{124.5} + \frac{(131 - 110.5)^2}{110.5} \\ &= 11.47\end{aligned}$$

$$df = (R - 1) \times (C - 1) = (3 - 1) \times (2 - 1) = 2$$

$$\text{p-value} = P(\chi_2^2 > 11.47) \Rightarrow \text{p-value} = 0.003$$

Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is some difference in the rate of support for drilling for oil and natural gas off the Coast of California based on whether or not the respondent graduated from college. Support for off-shore drilling and having graduated from college do not appear to be independent.

Question 2: [*7.39 - Coffee, depression, and physical activity.*] Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. Participants in a study investigating the relationship between coffee consumption and exercise were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed.

	Caffeinated coffee consumption					
	≤ 1 cup/week	2–6 cups/week	1 cup/day	2–3 cups/day	≥ 4 cups/day	Total
Mean	18,7	19,6	19,3	18,9	17,5	
SD	21,1	25,5	22,5	22,0	22,0	
n	12,215	6,617	17,234	12,290	2,383	50,739

(a) Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

(c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
coffee					0.0003
Residuals		25,564,819			
Total		25,575,327			

(d) What is the conclusion of the test?

Solution:

(a) **Hypotheses:**

- H_0: The mean MET for each group is equal to each other.

$$\mu_{\leq 1 \text{ cup/week}} = \mu_{2-6 \text{ cups/week}} = \mu_{1 \text{ cup/day}} = \mu_{2-3 \text{ cups/day}} = \mu_{\geq 4 \text{ cups/day}}$$

- H_A: At least one pair of means is different.

(b) **Conditions for ANOVA:**

1. **Independence:** We don't have any information on how the data were collected, so we cannot assess independence. To proceed, we must assume the subjects in each group are independent.
2. **Approximately normal:** The data are bound below by zero and the standard deviations are larger than the means, indicating very strong skew. However, since the sample sizes are extremely large, even extreme skew is acceptable.
3. **Constant variance:** This condition appears to be met, as the standard deviations are pretty consistent across groups.

In order to proceed with the test, we will need to assume independence.

(c) **Filling in the table:**

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
coffee	5-1=4	25575327-25564819=10508	10508/4=2627	2627/505=5.2	0.0003
Residuals	50739-5=50734	25,564,819	25564819/50624=505		
Total	50739-1=50738	25,575,327			

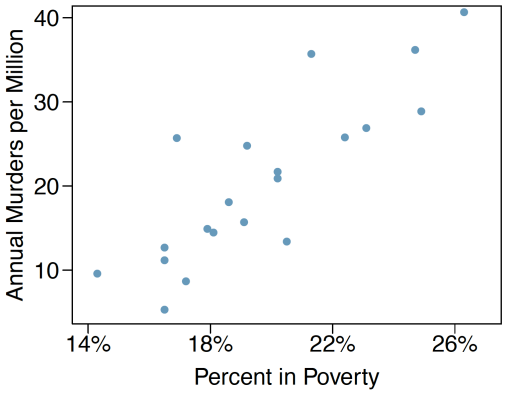
(d) **Conclusion:**

Since p-value is low, reject H_0. The data provide convincing evidence that the average MET differs between at least one pair of groups.

Question 3: [*8.25 - Murders and poverty, Part I.*] The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	<0.001

$s = 5.512$ $R^2 = 70.52\%$ $R^2_{\text{adj}} = 68.89\%$



(a) Write out the linear model.
(b) Interpret the intercept.
(c) Interpret the slope.
(d) Interpret R2.
(e) Calculate the correlation coefficient.

Solution:

(a) **Model:** $\widehat{murder} = -29.901 + 2.559 \times poverty\%$.

(b) **Intercept Interpretation:** The expected murder rate in metropolitan areas with no poverty is -29.901 per million. This is obviously not a meaningful value; it just serves to adjust the height of the regression line.

(c) **Slope Interpretation:** For each additional percentage increase in poverty, we expect murders per million to be higher on average by 2.559.

(d) **Variability Explanation:** Poverty level explains 70.52% of the variability in murder rates in metropolitan areas.

(e) **Square Root of R^2:** $\sqrt{0.7052} = 0.8398$.

Question 4: [*8.35 Murders and poverty, Part II.*] Exercise 3 [8.25] presents regression output from a model for predicting annual murders per million from percentage living in poverty based on a random sample of 20 metropolitan areas.

(a) What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

Solution:

(a) **Hypotheses:**

$$H_0 : \beta_1 = 0; \quad H_A : \beta_1 \neq 0$$

(b) **P-value and Conclusion:** The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate.

(c) **Confidence Interval:** Given $n = 20$, $df = 18$, $T_{18} = 2.10$;

$$2.559 \pm 2.10 \times 0.390 = (1.74; 3.378)$$

For each percentage point poverty is higher, the murder rate is expected to be higher on average by 1.74 to 3.378 per million.

(d) **Conclusion:** Yes, we rejected H_0 and the confidence interval does not include 0.