Practice Exam #1 - Sketch Solution (S24)

Ryan Miller

The following information will appear verbatim on the first page of Exam 1. You do not need to memorize this information, but you should be familiar with it.

Directions

- Answer each question using *no more than specified number of sentences* and not attempt to avoid these guidelines by using run-on sentences. Answers that are unnecessarily verbose may result in point loss.
- Do not include superfluous information in your answers, you may be penalized if you make an inaccurate statement even if you go on to provide a correct answer. Your answers should be clear, concise, and include only what is needed to answer the question that was asked.

Formula Sheet

Definitions:

- **Risk**: relative frequency of an event/outcome
- **Relative Risk**: ratio of the risks across two groups
- **Odds**: ratio of how often an event/outcome is observed relative to how often it is not observed
- **Odds ratio**: ratio of odds across two groups

Formulas:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

Simple linear regression (model):

$$Y = b_0 + b_1 X + \epsilon$$

Fitted regression line:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X$$

Question #1

In a scientific study, 50 people suffering from insomnia were divided into two groups. A group of 20 subjects participated in a one-hour therapy session, while the other group consisting of the remaining 30 subjects did not receive any treatment. Three months later, 13 people in the therapy group reported improved sleep, while 12 people in the group not receiving therapy reported an improvement.

Part A: If these data were stored in "tidy" format with each case as a row and each variable as a column. How many rows and columns would the data frame contain? Do not consider any subject identifiers or variables not listed in the prompt. You do not need to explain your answer.

- 50 rows (1 per subject)
- 2 variables (group = trt/control, outcome = improvement/not)

Part B: Of the variables present in this data set, identify which is the explanatory variable and which is the response variable. Briefly explain your answer using at most 2-sentences.

- The explanatory variable is "group" and the response is "outcome", because the group is assigned prior to the outcome being observed and can be viewed as a potential cause of the outcome.

Part C: Describe or sketch an appropriate data visualization that could be used to explore whether the explanatory and response variables you identified in Part B are associated. If providing a written description, limit your answer to no more than 2-sentences. If providing a sketch, you do not need to be overly precise so long as I can judge that it is the right type of graph.

- Any *conditional* stacked bar chart that conditions on "group" is acceptable. Because the data have unequal group sizes, the bar chart should be conditional.

Part D: Create a contingency table summarizing the results of this study. Make sure to use the explanatory variable to define the table's rows and the response variable to define the table's columns.

	improved	not
therapy	13	7
control	12	18

Part E: Find the *odds ratio* that compares the odds of improved sleep in the group receiving therapy with the odds of improved sleep in the group not receiving therapy. Show your work for any calculations.

- Odds of improvement for therapy group = 13/7 = 1.857
- Odds of improvement for control group = 12/18 = 0.667
- Odds ratio = 1.857/0.667 = 2.8

Part F: Provide a 1-sentence interpretation of the odds ratio you found in Part E. Then, briefly indicate whether this odds ratio suggests an association between these variables. In total your entire response should be exactly 2-sentences.

- The odds of a person in the therapy group experiencing improved sleep were 2.8 times the odds of a person in the control group experiencing improved sleep. This is a large increase in the likelihood of improvement and thus the therapy is associated with the outcome.

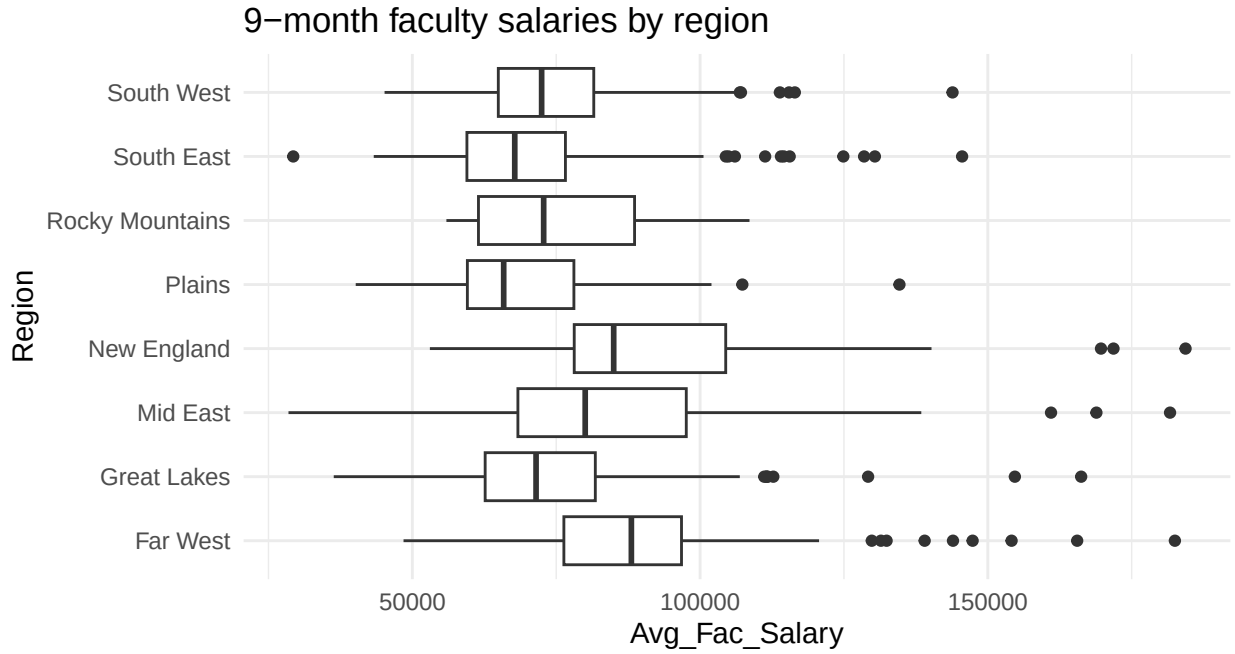
Question #2

The data for this question are from the 2019 College Scorecard. We've previously used these data in class, though the data visualization and descriptive statistics below include only colleges with at least 1000 enrolled students.

1. "Region" - the census-designated geographic region where each college is located.
2. "Avg_Fac_Salary" - the average 9-month salary of the faculty members at each college.

Table 2: Comparative summary of the median salaries of students from 1095 different colleges and universities according to geographic region

Region	N	Mean	StDev	Median	IQR
Far West	92	91289.15	23802.86	88060.5	20443.50
Great Lakes	156	74451.75	18011.65	71496.0	19140.75
Mid East	179	84800.92	22682.08	80037.0	29259.00
New England	65	92525.12	27103.22	84987.0	26325.00
Plains	97	69770.41	15275.53	65871.0	18513.00
Rocky Mountains	29	75944.48	15920.27	72819.0	27126.00
South East	238	70039.78	16814.71	67797.0	17118.00
South West	79	75846.30	17185.57	72468.0	16591.50



Part A: Is there an association between the variables "Region" and "Avg_Fac_Salary"? Explain your answer in at most 2-sentences.

- Yes, some regions such as New England or the Far West have much higher median salaries than regions like Plains or Great Lakes. Thus, knowing the region of a college could help you come up with a better estimate of the avg faculty salary of that college.

Part B: Describe the distribution of "Avg_Fac_Salary" *within* the New England region. Limit your description to at most 2-sentences.

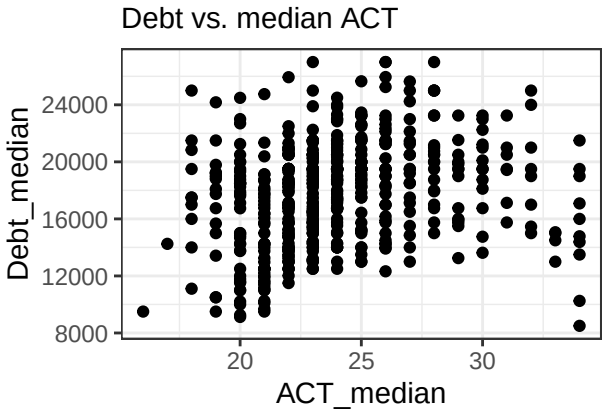
- The distribution is *centered around 80,000 dollars*, it is *skewed right*, and the IQR describing the spread of the middle 50% of colleges in this region is 26,000 dollars, and there are a few outliers with very high average salaries.

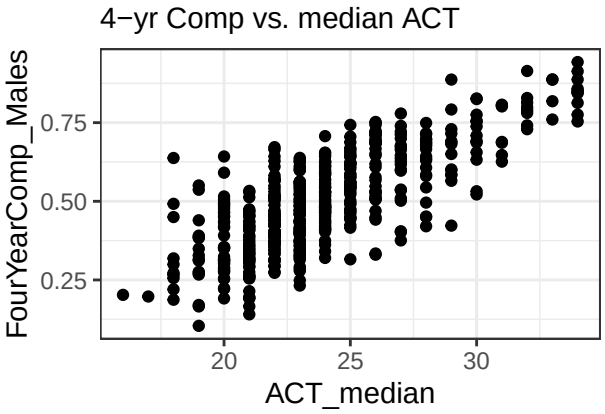
Part C: Using a *robust* descriptive statistic, which region exhibits the largest amount of variability in "Avg_Fac_Salary"? You do not need to explain your answer.

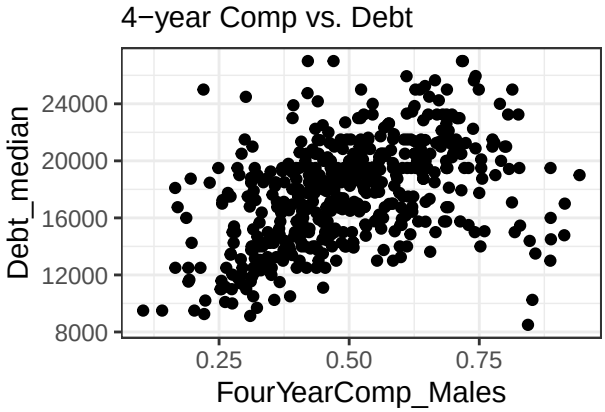
- Mid East (IQR of 29259.00)

Question #3

This question also uses the 2019 College Scorecard data used in Question #2. Below are scatter plot of the variables "ACT_median" (in points) and "Debt_median" (in US dollars) for colleges with at least 1,000 enrolled students, as well as some additional `R` output:








```
## Pearson correlation
cor(col$ACT_median, col$Debt_median, method = "pearson")
```

```
## [1] 0.2281373
```

```
## Spearman correlation
cor(col$ACT_median, col$Debt_median, method = "spearman")
```

```
## [1] 0.2982612
```

```
## Simple linear regression
model1 = lm(Debt_median ~ ACT_median, data = col)
coef(model1)
```

```
## (Intercept)  ACT_median
##  11990.2344    240.9512
```

```
## Multivariable linear regression
model2 = lm(Debt_median ~ ACT_median + FourYearComp_Males, data = col)
coef(model2)
```

```
##         (Intercept)          ACT_median FourYearComp_Males
##          17308.9789           -300.6014          15341.7652
```

Part A: Using only the scatter plot of "ACT_median" vs. "Debt_median", qualitatively describe the relationship between these variables. Limit your response to at most 2-sentences.

- weak, non-linear, somewhat positive for most of the graph w/ a negative relationship for higher median ACTs.

Part B: Is it appropriate to rely upon Pearson's correlation coefficient to describe the relationship between "ACT_median" vs. "Debt_median"? Briefly explain, limiting your response to at most 2-sentences.

- Probably not, these variables have a non-linear relationship that is positive until a median ACT of around 26/27, then negative after that.

Part C: Interpret the effect of "ACT_median" in the the *simple linear regression* model where "ACT_median" is used to predict "Debt_median". Limit your response to a single sentence.

- For every 1 pt higher the median ACT of a college is, you'd expect the median debt of its graduates to increase by about 240 dollars.

Part D: Interpret the effect of "ACT_median" in the the *multivariable linear regression* model where "ACT_median" and "FourYearComp_Males" are used to predict "Debt_median".

- Given two colleges have the same 4-year completion rate, you'd expect each additional 1 pt higher the median ACT is to lead to an expected 300 *decrease* in the median debt of its graduates.

Part E: Briefly explain why the estimated coefficient for "ACT_median" is positive in one model but negative in the other. How is this possible? And why does it happen? Limit your response to at most 4-sentences.

- The first model doesn't control for the impact of four-year completion rate, which has a moderately strong positive relationship with median debt and a strong positive relationship with median ACT. This means that four-year completion rate meets the definition of confounding, so it shouldn't be surprising that it influences the relationship between ACT median and median debt. Considering these relationships, it's likely that ACT median is reflecting/exhibiting much of the positive relationship between 4-year completion rate and median debt through omitted variable bias in the simple linear regression model.

Part F: In Parts A and B you described the relationship between "ACT_median" and "Debt_median". Could your description be an example of the *ecological fallacy*? Briefly explain one way by which the ecological fallacy could occur in this situation. Limit your response to at most 3-sentences.

- Yes, both of these variables are aggregations of individual students who have ACT scores and leave with debt. Just because a positive relationship exists when students are aggregated within a college, does not necessarily mean that individual students are more likely to higher personal levels of student loan debt if they scored highly on the ACT.