Practice Exam #1 (Sta-209, S25)

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:

Mean:

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

Standard deviation

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

Pearson's Correlation Coefficient:

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

Simple linear regression (theoretical model):

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

Simple linear regression (fitted model):

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

Coefficient of Determination (R^2):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

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.

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.

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.

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.

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.

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.

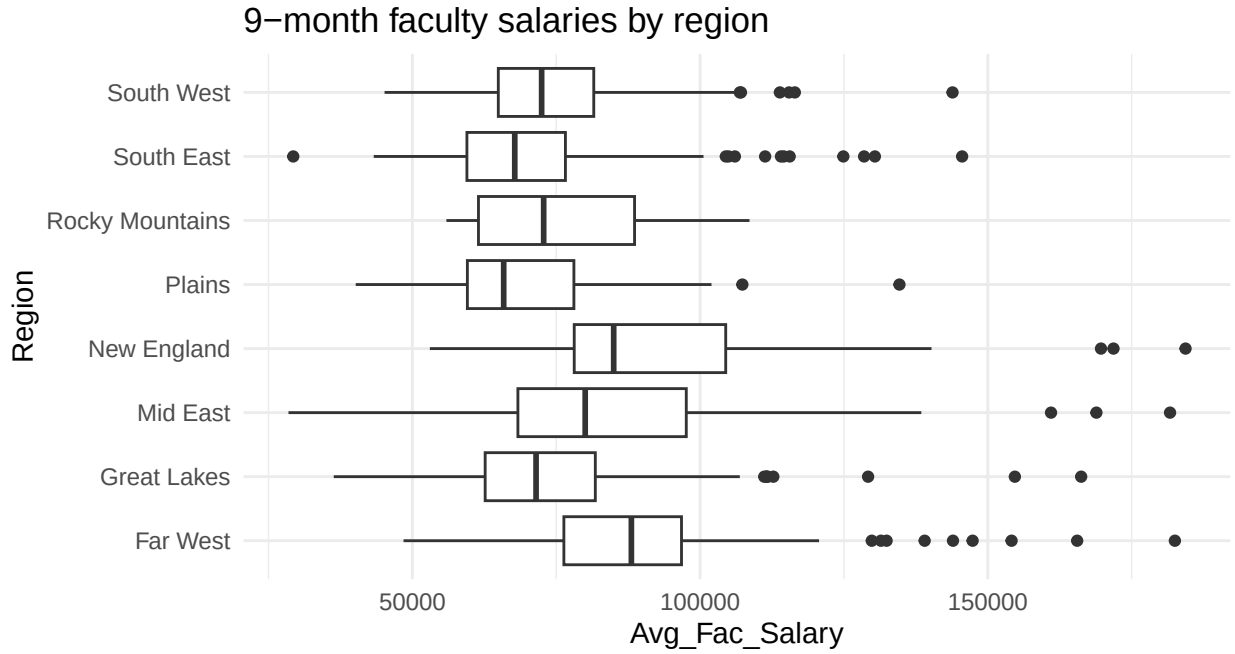
Question #2

This question uses data from the 2019 College Scorecard. The visualizations and descriptive statistics below include all colleges that primarily grant undergraduate degrees and have at least 1000 full-time enrolled students. The information below summarizes two variables:

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 1: 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



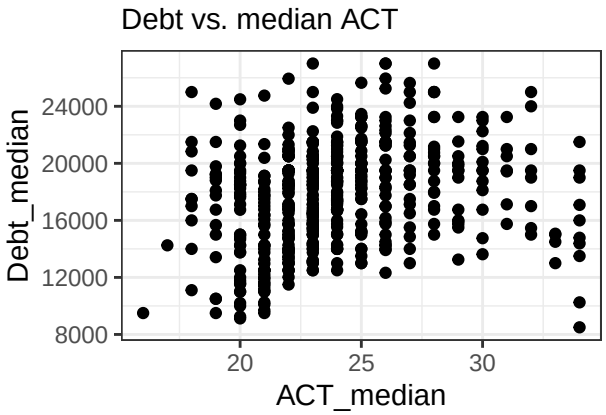
Part A: Is there an association between the variables "Region" and "Avg_Fac_Salary"? Explain your answer in at most 2-sentences.

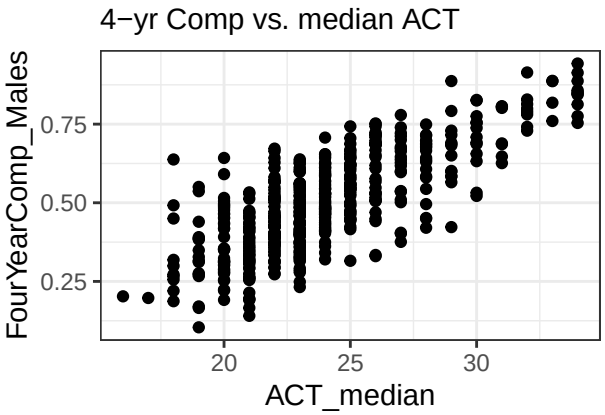
Part B: Describe the distribution of "Avg_Fac_Salary" *within* the New England region. Limit your description to at most 2-sentences.

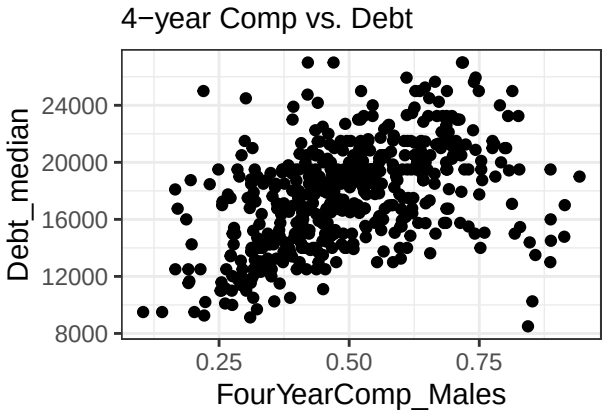
Part C: Using a *robust* measure of spread, which region exhibits the largest amount of variability in "Avg_Fac_Salary"? You do not need to explain your answer.

Question #3

This question also uses the 2019 College Scorecard data used in Question #2. Below are scatter plots of the variables "ACT_median" (in points), "Debt_median" (in US dollars), and "FourYearComp_Males" (proportion of male students who complete a degree in 4 years) 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.

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.

Part C: Interpret the coefficient 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.

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

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.

Part F: Consider Parts A and B where you described the relationship between "ACT_median" vs. "Debt_median". How would an interpretation of this relationship that *commits the ecological fallacy* differ from a proper interpretation? Briefly explain using no more than 3 sentences.

Part G: Now consider one additional regression model that also includes the college's type (either "Private" or "Public") as a third explanatory variable. Provide a 1-2 sentence interpretation of the coefficient for the variable "TypePublic" in the fitted model shown below:

```
## Another Multivariable linear regression
model3 = lm(Debt_median ~ ACT_median + FourYearComp_Males + Type, data = col)
coef(model3)
```

```
##        (Intercept)         ACT_median FourYearComp_Males         TypePublic
##         19450.4185          -278.7298         12334.6165         -2802.7207
```