Exam 1

STA209-04: Applied Statistics

February 22, 2019

Please carefully read each question. You will have 80 minutes to complete this exam. Show all of your work. All short answers should be no more than three sentences in length. Write your name on the upper right hand corner of your answer sheet.

1) [30 pts] In order to determine the relationship between smoking status and mortality among women, a survey was administered to female residents of Whickham, a small town in northeast England, during the years 1972 - 1974. Twenty years later, a follow-up was conducted to assess whether participants were deceased or still living.

i) How would you describe the design of this study? What concerns, if any, do you have about this design? Explain.

ii) Based on the provided information, describe the cases and variables of this study. When describing the variables, be as specific as possible about their type.

iii) Suppose the results of this study showed that more smokers were deceased at follow-up as compared to non-smokers. **True or False**: Based on these results, we can conclude that smoking causes increased mortality among women. Explain.

iv) Suppose the results of this study showed that more smokers were deceased at follow-up as compared to non-smokers. **True or False**: These study results may be generalized to women living in Europe. Explain.

The following table summarizes the results of the previously described study:

	Smoker		**Non-smoker**	
Age Group	**Alive**	**Deceased**	**Alive**	**Deceased**
18-54	372	46	393	25
55-64	64	41	81	40
65+	7	42	28	165

v) What type of variable is 'Age Group'? Be as specific as possible.

vi) Which type of graph would be most appropriate to use if you were interested in comparing the amounts of smokers and non-smokers contained in your sample? Explain.

vii) Which type of graph would be most appropriate to use if you were interested in comparing the proportions deceased between smokers and non-smokers across each age group? Explain.

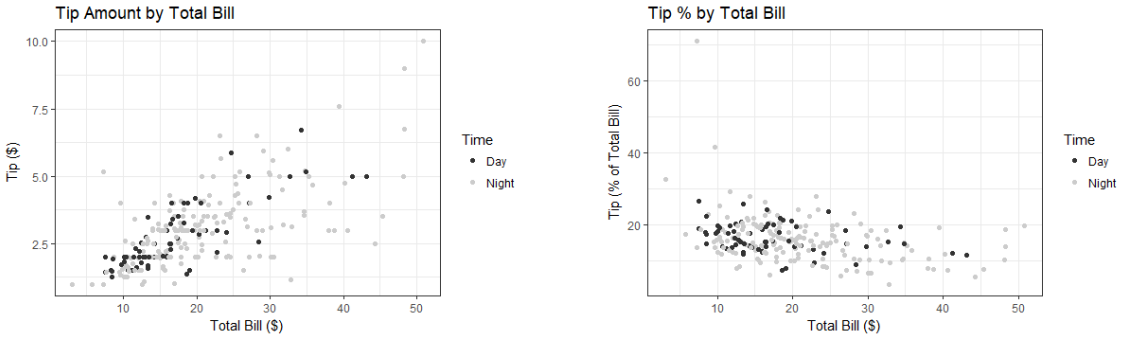
viii) Compute the marginal death rates for smokers and non-smokers. Based on these results, what can you conclude?

ix) Compute the age group specific death rates among smokers and non-smokers. Based on these results, what can you conclude?

x) Between the conclusions drawn in **(viii)** and **(ix)**, which should be reported to the general public? Why?

2) [40 pts] A friend of yours is down on their luck after having their car, "Old Reliable", break down on them. After receiving an estimate from several mechanics, it was determined that the repair bill would be more than what it would cost to purchase another car. Being a broke college student, your friend was forced to find a job as a server hoping to save enough for a car purchase.

Wanting to help your friend in this bad situation, you offered to put your statistics and data analysis skills to use in order to help them find ways to maximize their tip. Using a dataset containing tipping information (that you happened to find), you produced the following two scatterplots to support a recommendation.



i) Focusing on "Tip Amount by Total Bill", would it be appropriate to use the correlation coefficient to quantify the strength of the association? Explain.

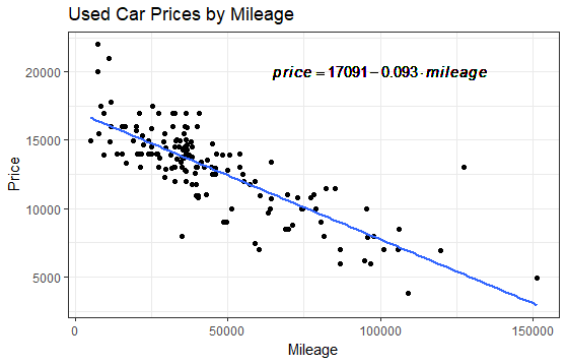
ii) Assume you computed a correlation coefficient for the association between Total Bill and Tip ($), considering only the "Day" datapoints. **True or False**: The correlation coefficient would be positive, further from 0, and closer to 1. Explain.

iii) Assume you found the correlation coefficient for the association between Total Bill and Tip (% of Total Bill) to be -0.33. **True or False**: Total Bill and Tip (% of Total Bill) have an inverse linear relationship. Explain.

iv) Which of these two figures would be most useful in answering the question: "Which time of day are people more generous tippers?"

v) Suppose your friend only had the option to work night shifts. After talking to some of his night-shift coworkers, he determined that the most common total bill amount was 35$. Based on the data which generated the above figures, what is the tip amount (in dollars) that your friend can expect for a total bill amount of 35$? Note that the average total bill in the data is 19.79 and the standard deviation is 8.90. The average tip amount in dollars is 3.00 and the standard deviation is 1.38. The correlation between these two is 0.68.

After a few months of working, your friend is finally looking to purchase a car. Extremely grateful for your initial statistical consultation, your friend called on you for yet another analysis. He managed to collect data on used car mileage and price and is hoping you could help him find a good deal. The figure below displays the data and fitted regression line.



vi) Your friend doesn't understand regression. Interpret the slope coefficient in terms he would understand.

vii) What amount of mileage should your friend look for if he expects to pay for at most a $13,000 car?

viii) **True or False**: The amount of mileage determined in **(vii)** would be the same as what would be predicted by the regression of mileage on price (i.e. price is the explanatory variable and mileage is the response). Explain.

The table below displays data on a few cars that your friend is really interested in:

Car	Price	Mileage	Residual
A	11749	57341	7.35
B	10000	63926	-1127.33
C	9995	95364	1800.55
D	12995	32743	-2106.62

ix) Among the cars listed, which would you suggest your friend buy? Explain.

x) Suppose your friend came upon a time-sensitive deal he felt he had to accept without your consultation. In this deal, he only paid $1,000 for a car with 160,000 miles on it. He wants to know (based on your model) whether he was smart to accept the offer. What would you tell him?

3) [30 pts] In 1867, the results of a study investigating the benefits of sterile technique in surgery were published. This study tested a sterile operating procedure developed by Joseph Lister in which surgeons were required to wash their hands, wear clean gloves, and disinfect surgical instruments with carbolic acid. The table below provides a summary of the experimental results.

Group	Died	Survived
Sterile	6	34
Non-Sterile	16	19

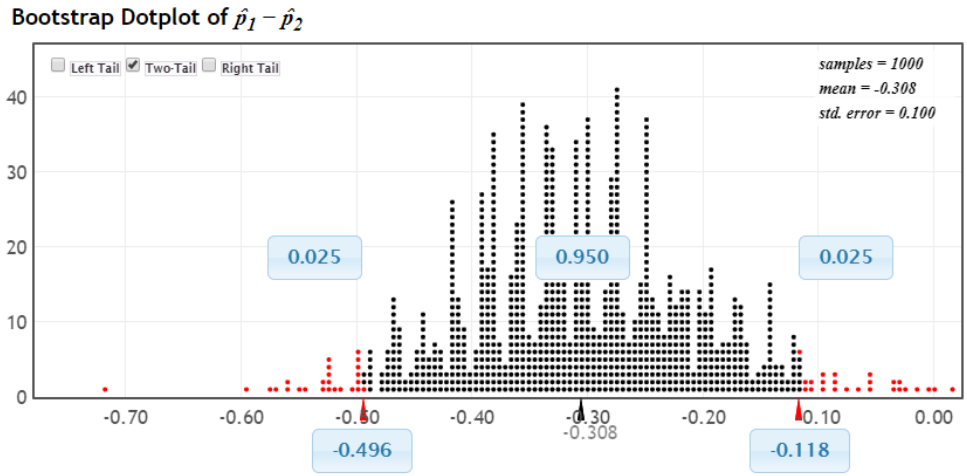
i) Suppose the results from these data can be used to draw inference for all surgical patients in 1867. Suppose also that we found the difference in death rates (Sterile - Not) to be -0.31. **True or False**: Based on these results, we can conclude that surgical patients whose surgeons followed sterile protocols are exactly 31% less likely to die from surgery than non-sterile surgical patients. Explain.

ii) Assume that this was a randomized experiment. **True or False**: Randomization eliminates confounding. Explain.

iii) Suppose we repeated this experiment 100 times in order to determine the sampling distribution for the difference in death rates. **True or False**: The sampling distribution will be centered at the first experiment's sample mean. Explain.

iv) Suppose we found the difference in death rates (Sterile - Not) to be -0.31. **True or False**: Sterilized surgery is correlated with a lower death rate. Explain.

Shown on the following page is the bootstrap distribution of the difference in death rate between sterile and non-sterile surgical patients.



v) Suppose that your study sample is biased. **True or False**: Increasing the number of bootstrap samples will reduce this bias. Explain.

vi) Suppose that your study sample is representative of the population. **True or False**: Increasing the size of each bootstrap sample will increase the standard error of the bootstrap distribution. Explain.

vii) Using either the standard error or percentile method, construct a bootstrap confidence interval of the difference in proportion. The difference in proportion for the original sample was -0.31. Justify your choice of confidence interval construction method.

viii) Explain what is meant be the phrase, "We are 95% confident that...".

ix) Would an 80% confidence interval be wider or more narrow than the confidence interval computed in **(vii)**? Explain.

x) Interpret the confidence interval computed in **(vii)**. What can we conclude based on this interval?