Homework 1

1. An article under the headline “26 Million Homeless, Survey Says” in the December 28, 1994, issue of the Columbus Dispatch reported that 13.5 million Americans have been homeless for at least a few days sometime during their lives, based on a study by Professor Bruce Link at Columbia University. Another 12.5 million have stayed off the street only by moving in with friends or family for periods that ranged from a few days to a year. The estimates were based on a telephone survey of 1,507 adults that asked whether they recalled ever being homeless, for how long and where they slept — in the street, a shelter, abandoned buildings or someone else’s home. Comment on this article and the results of this survey, paying particular attention to possible weaknesses in the sampling strategy. Be as specific as possible.

2. The file mileage2017.mpj gives data for 1203 auto models from model year 2017, as given in the Model Year 2017 Fuel Economy Guide (www.fueleconomy.gov). The file gives the engine displacement (in liters), the number of cylinders in the engine, whether the transmission is automatic (1) or manual (0), and the official EPA city, highway, and combined mileages (in miles per gallon). You should view these data as a snapshot of an ongoing stable random process related to the characteristics of autos sold in the U.S., and hence inference is meaningful considering that as the population. (a) Calculate the mean and standard deviation of the number of cylinders, and of highway mileage. Treating the given data as being a random sample from the general population of autos, for each variable, give an interval within which we would expect roughly 95% of all autos to fall. For which of these variables do you think the interval is likely to be more useful? Why?
(b) Compute the correlation coefficients between (i) highway mileage and number of cylinders, and (ii) highway mileage and engine displacement. Which variable, number of cylinders or engine displacement, seems to have the stronger relationship with highway mileage? Explain how you got your answer.
(c) Construct and report the regression line that allows you to predict highway mileage from engine displacement. What is the interpretation of the slope of the line?
(d) Construct and report the regression line that allows you to predict highway mileage from city mileage. What is the interpretation of the slope of the line?
(e) Take a look at the entries in the first three columns in the data set. Is there anything about them that seems noteworthy? Is there anything that might make you wonder about the validity of any of the implications of parts (a) through (d)?

3. The file yogurt.mpj gives data from pages 21-22 of the August, 2017 issue of Consumer Reports about whole-milk yogurt products. The file gives an indicator if the yogurt is regular(0) or Greek (1), whether the flavor is Plain (0) or Berry (1), calories, total fat (g), total saturated fat (g), protein (g), carbohydrates (g), sugars (g), sodium (mg), and calcium as a percentage of the recommended daily value, all per serving size, taste rating (3=Fair, 4=Good, 5=Excellent), serving size (oz), package size (oz), and average price per package in dollars. Using any statistics or graphics that you wish, explore the relationships in the data between price and the characteristics of the

© 2019, Jeffrey S. Simonoff
yogurt. Do you “get what you pay for” in any reasonable sense? Are there particular yogurts that seem particularly under- or overpriced? How might you use your results in deciding what yogurt to buy? Be sure to examine your data in way(s) that make sense in terms of answering these kinds of questions.

Reminder: The files mileage2017.mpj and yogurt.mpj, and other data files that you will need (or I will use) during the semester, can be obtained from the class web site. They also can be obtained locally on the x drive (x:\sor\jsimonof\1305). I recommend that you word process this and future homeworks, but this is not required; you should make sure that the homework is neat, however (you will lose credit for sloppiness).

Note: You might be uncomfortable doing this homework for a few reasons. First, it will probably take you longer to complete than (most of) the other homeworks, so you should leave yourself enough time to complete it. You might also find it open-ended, leaving you with a sense of not knowing exactly how to proceed. This is perfectly normal, and, in fact, desirable. Data analysis is by its nature amorphous and open-ended, and one of the purposes of this homework is to give you some experience with that. In any data analysis problem you should try to focus on the questions that might be of interest to you, and how you might use statistical methods to help you to answer those questions. You’ll also notice when doing this homework that we have not discussed much of this material in class; one of the purposes of this homework is to make sure that everyone is up to speed on basic data analysis, and the use of Minitab, before we get further into more complex data analysis later on in the semester. You should show your work and justify your statements in this and future homeworks, or you could lose credit.

Homework due: September 24