M358K Sp 06

REVIEW PROBLEMS FOR FIRST EXAM

Please Note: This review sheet is not intended to tell you what will or what will not be on the exam. However, most of these problems have appeared on or are very similar to problems that have appeared on previous exams in this course.

Remember that on the exam you will be expected to give reasons and explain your work clearly.

1. A researcher wishes to study how the average weight y (in kilograms) of children changes during the first year of life. He gathers data from a suitable sample of children, recording each child's weight each month for a year after they are born. He then plots mean weight y against age x (in months) and fits a least squares regression line to the data, with x as the explanatory variable and y as the response variable. He computes the following quantities:

r (= correlation between x and y) = 0.9

x-bar (= mean of the values of x) = 6.5

y-bar (= mean of the values of y) = 6.6

s_x (= standard deviation of the values of x) = 3.6

s_y (= standard deviation of the values of y) = 1.2

a. Based on this information, how much would you expect mean weight to change as age increases one month in this age range? (Explain)

b. What percent of the variation in mean weight is accounted for by regression on age, for this age range? (Explain briefly)

c. If instead of mean weights, we considered the individual weights of all the children in the sample (but still in the same age range), how would you expect the correlation between weight and age to compare with the correlation between average weight and age given here? (Explain)

d. If you just knew x-bar, y-bar, s_x and s_y (i.e., not r or the original data), you would still know one point on the regression line. What is that point?

e. Find an equation for the least squares regression line.

f. A second researcher collects data on another sample of children four months old. She finds their mean weight to be 5 kg. Find the predicted mean weight of four

month old children from the first sample, and find the residual for the observed mean weight in the second sample.

2. The distribution of actual weights of chocolate bars produced by a certain machine is normal with a mean of 8.1 oz. and a standard deviation of 0.1 oz. What weight should be put on the chocolate bar wrappers so that only 1% of bars are underweight? (Explain)

3. The temperature at randomly chosen locations in a kiln used in the manufacture of bricks is normally distributed with a mean of 1000° F and a standard deviation of 50° F. If bricks are fired at a temperature above 1125° F, they will crack. If they are fired at a temperature below 900° F, they will discolor. If bricks are placed randomly throughout the kiln when fired, what proportion will be good (i.e., neither cracked nor discolored)?

4. One hundred volunteers who suffer from severe depression are available for a study. The study will compare the effectiveness of a new drug for treating severe depression with an existing drug for treating this condition. A psychiatrist will evaluate the symptoms of all volunteers after two months to determine if there has been a substantial improvement in the severity of the depression. Describe an appropriate design for the study. Explain why your design is appropriate.

5. True or false. If the statement is false, explain why. (Depending on the statement, this may need to be done by an example showing that the statement is false.)

a. A sample chosen in such a way that every unit in the population has the same chance of being selected is a simple random sample.

b. A normal quantile plot is formed by plotting points (z,x) where z is the normal score $(x-\mu)/\sigma$ of x.

c. If the mean of a distribution equals the median of the distribution, then the distribution is normal.

d. If we regress x on y, then solve for y, we get the same line as when we regress y on x.

e. If two variables have correlation zero, then there is no relationship between them.

f. The largest value in a data set is an outlier.

g. If the correlation between people's memories of what they ate and what they actually ate is .22, then 22% of the people remembered accurately what they ate.

h. If the correlation between the actual weight of a sample of test weights and the weight indicated by a certain scale is 1, then the scale is accurate.

6. For each of the following, state which of correlation, regression, a well designed experiment, or none of these would be most appropriate to study the question. Explain why.

a. Determining how well height at age four predicts adult height.

b. Determining whether a new drug causes blood pressure to lower.

c. Determining the average per person beef consumption in Texas.

7. A marketing research firm wishes to determine if the adult men in a certain small city would be interested in a new upscale men's clothing store. From a list of all 95,481 residential addresses in the city, the firm selects a simple random sample of 100 and mails a brief questionnaire to each of these 100 addresses.

a. What is the population of this study?

b. What is the sample in the study?

c. What is the sampling frame in this study?

8. We have seen the when you do a least squares regression, the sum of the residuals e_i is zero. What can you say about the mean $\bar{e}$ of the residuals?

9. In each of the following examples, explain how the sample chosen or the wording of the question asked probably biased the results of the study.

a. A 1992 Roper poll reported that 22% of Americans say that the Holocaust may not have happened . The actual question asked in the poll was

"*Does it seem possible or impossible to you that the Nazi extermination of the Jews never happened?"*

Twenty-two percent of participants in the poll responded "possible".

b. A study sponsored by American Express Co. and the French Government tourist office reported that old stereotypes about French unfriendliness weren't true. The respondents were over 1000 Americans who had visited France more than once for pleasure over the past two years.

c. A simple random sample of 1200 adult Americans was selected; each person in the sample was asked the following question:

" In light of the huge national deficit, should the government at this time spend additional money to establish a national system of health insurance?"

Thirty-nine of those responding answered yes.

10. Discuss difference between uses of the following words in ordinary usage and in statistics:

a. Correlation
b. Predict
c. Experiment

11. Compare and contrast: the correlation coefficient of a set of two variable data, and the slope of the least squares regression line of one of the variables on the other.

12. A researcher is interested in studying college students' perceptions about whether student drinking is a problem. She will sample students from UT Austin (which has about 50,000 students) and Rice University (which has about 5000 students).

a. i. Suppose simple random samples of 100 students from each school are taken. Will the variability of the sample proportion for the two schools differ? If so, how? Explain briefly.

ii. Suppose that instead simple random samples of 5% of the students from each school are taken. Will the variability of the sample proportion for the two schools differ? If so, how? Explain

b. Suppose 40% of UT students actually believe that student drinking is a problem. Suppose 100 UT students are asked their opinion and 30% of them say they think student drinking is a problem. Identify the *parameter* and the *statistic* in this situation.

13. A random sample of records of sales of 117 homes during a certain period in 1993 in Albuquerque, New Mexico gave price (in thousands of dollars) and size (in square feet). The resulting least squares regression equation (with y = price and x = size) was

$$\hat{y} = 47.82 + 0.061x$$

Explain what the slope of the least squares line says about housing prices and sizes. Be as precise as possible given the above information.

14. For each statement, identify and explain the error in understanding and/or interpretation of correlation:

a. "The correlation of -0.72 between average household income and infant mortality rate shows that there is almost no association between these two variables."

b. "The correlation of 0.83 between gross domestic product and literacy rate shows that countries wanting to increase their standard of living should invest heavily in education."

15. Here is part of the output Minitab will give you if you enter the 1996 mean SAT verbal scores in the 50 states plus the District of Columbia and ask for descriptive statistics:

N	Mean	Median	StDev	Min	Max	Q1	Q3
51	531.90	525.00	33.76	480.00	596.00	501.00	565.00

a. Use this information to construct a boxplot these data.

b. Based on the Minitab output and your boxplot, what can you tell about the shape of the distribution of these data?

16. A few years ago, a group of students in this class was interested in the question of whether people who make illegal right turns on red signal for the turn. They narrowed their project down to study people making illegal right turns on red on weekdays from 8 AM to 5 PM from west-bound twenty-first street onto north-bound Guadalupe. In a preliminary study, they found that six cars made illegal right turns on red at the intersection in a fifteen minute observation period, so they decided that fifteen minute observation intervals were appropriate for collecting their data. They divided the times from 8 AM to 5 PM in one week of weekdays (Monday through Friday) into fifteen-minute intervals and excluded the time slots when no one in the group was able to collect data. (This turned out to be exactly the times when the course met.) They numbered the remaining time intervals and used a random number generator to select 12 of them. They chose an average week that, to the best of their ability to determine, had no unusual events that might disrupt normal traffic patterns. During observations, the observer sat near the sidewalk somewhat out of sight of drivers to prevent any influence on drivers' behavior.

a. What is (are) the population(s) studied in this project?

b. What is (are) the variable(s) studied in this project?

c. What is (are) the parameter(s) studied in this project?

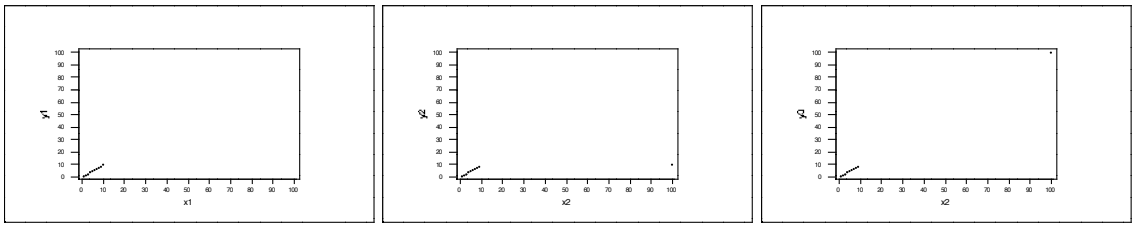
d. What is (are) the sample(s) used in this project?

e. What is (are) the statistic(s) calculated in this project?

f. What is (are) the sampling frame(s) used in this project?

g. Comment on the data collection process. Could it be improved?

17. Below are three plots of 10 data points each:



A B C

Without doing any calculations, answer each of the following questions, *and explain how you know*.

a. How does s_x compare for the three plots?
b. How does s_y compare for the three plots?
c. How does the correlation coefficient r compare for the three plost?
d. How does the slope of the regression line compare for the three plots?

Additional review problems from the textbook:

1.131, 1.137, 1.143
2.109, 2.111
3.81, 3.83, 3.85, 3.89, 3.95