

Stat 31-5 Class Project Part III: Exploring relationship and discussing sampling design

Assigned on Wednesday 10-15-03 and due on Wednesday 10-22-03

Name: \_\_\_\_\_

**Before you start, you might need to manipulate your data**

- If you have **two quantitative variables**, say  $\{Q_1, Q_2\}$  and **one categorical variable**, say  $C$ , in your data, don't change your data and you can start answering the questions.
- If you have **three quantitative variables** in your data set, say  $\{Q_1, Q_2, Q_3\}$ , you will need to manipulate your data as follows to have two quantitative variables and one categorical variable. *Data manipulation:*
  1. Pick a variable say  $Q_3$  to transform into a categorical variable.
  2. Divide  $Q_3$  into 3 categories of approximately equal sizes as follows: sort your data by the values of  $Q_3$ , and divide the data into the three following categories: small values of  $Q_3$  denoted by S, medium values of  $Q_3$  denoted by M, and large values of  $Q_3$  denoted by L.
  3. The new variable taking values S,M, or L for each individual will be your categorical variable  $C$ .

With the two quantitative variables  $\{Q_1, Q_2\}$  and the categorical variable  $C$  you can answer all the following questions.

**Answer the following questions and turn in your answers and the required Excel graphics**

1. Fill in the dots by the appropriate variable's name. Write NA if the statement is 'not applicable' to your problem. More than one statement could be applicable to your problem.
  - (a) You would like to explore the relationship between the two quantitative variables ..... and ..... Either could be explanatory or response.
  - (b) You would like to explore the relationship between the two quantitative variables..... and ..... and/or predict the value of ..... given the value of ..... Your response variable is ..... and your explanatory variable is .....
  - (c) You would like to explore the relationship between the two quantitative variables for each category of the variable .....
2. Scatterplot:
  - (a) Make a scatterplot of your data exploring the relationship between your two quantitative variables. Turn in the Excel sheet with your graphic.
  - (b) Describe the relationship between the two variables, i.e answer the following questions: is the relationship curved or linear?, is it strong or weak?, do you see any granularity in the data?, if the relationship is linear is it positive or negative?, are there any outliers?, are there any clusters in the data? any other main feature of the graphic?
  - (c) Make a scatterplot of your two quantitative variables showing the different categories of your categorical variable in the same plot. Use different colors or different symbols for each category. Turn in the Excel sheet with your graphic.
  - (d) Do the different categories of your categorical variable appear as different clusters or overlap with each other? For further investigation of the relationship, you can choose to answer question 4 either with the complete data set or the data set for each category.
3. Fill in the blanks in all statements
  - (a) The mathematical formula for the correlation  $r$  is .....
  - (b) The mathematical formula for the slope of a linear regression is .....
  - (c) The mathematical formula for the intercept of a linear regression is .....  
The unit of the intercept is the same as the unit of the ..... variable.
  - (d) The ..... is invariant to change of units of your variables whereas the ..... and ..... are not invariant.
4. Numbers describing the relationship and goodness of linear fit.
  - (a) Use the function "=correl" in Excel to find the correlation between your two quantitative variables. Give the value of the correlation.
  - (b) Is the correlation a good measure of relationship between your two variables? why?
  - (c) Find the equation of the least squares regression line using the mathematical formulas in your answers to questions 3.b and 3.c. (Don't recompute the correlation but use the value that you found in 4.a, you can use Excel or your answers in Class Project Part II to find the sample means and sample standard deviations.)
  - (d) Show the least square regression line on your scatterplot and check with Excel that you obtained the correct values for the slope and intercept. Turn in the Excel sheet with on the same graphic the scatterplot, regression line, and equation of the regression line.

- 
- (e) Fill in the dots: ..... percent of the variation in ..... is explained by the least square regression of ..... on .....
- (f) Make a residual plot of the data. Turn in an Excel sheet with the graphic of the residual plot.
- (g) What are your final conclusions about the data: is there a relationship between the two variables?, is the linear fit a good fit?, are there any outliers? are they influential? do you have different clusters? do you suspect the existence of lurking variables? do you believe that a strong association would mean causation? any other thought or suggestion for further investigation?
5. Discussing the design. Fill in the dots
- (a) Your target population is .....
- (b) Your sample (is?/is not?)..... a simple random sample (SRS) of the target population. (You can check the definition of a SRS in your notes or in the book)
- (c) Possible sources of bias in your sampling scheme are (list at least two): ....., .....
- (d) Possible source(s) of bias in the answer to your questions (is?/are?) .....: .....
- (e) Based on your answers to questions 5.a-5.d, you feel (confident?/not confident?)..... about generalizing the results obtained on this sample to your general target population.
6. Open question. If you could redo a sampling or perform an experiment with a limited budget to answer the same questions you had on the target population. How would you go about it?