STA441/1008 Assignment 1

Quiz on Tuesday Jan. 16

This assignment is based on the Chapters 1 and 2 of the online text, and lecture slide sets 1-3. **Chapter 1 is assigned reading**, while Chapter 2 may be helpful and is optional for now.

The non-computer part is preparation for Quiz 1; it is not to be handed in. For the computer part, please bring hard copy of your log file (*not* your program file) and output file to the quiz. You may be asked to answer questions about them, and you may be asked to hand them in with the quiz.

Note that your log file and output file must be produced by the same run of SAS. If they do not, you will get zero for the computer part of the quiz.

When reviewing for the quiz, please pay attention to concepts like Explanatory variable, Response variable, Categorical variable, Quantitative variable, Statistical significance, Significance level of a test, p-value, Definition of "unrelated" and "related" variables, Univariate, Multivariate, Explanatory observations, Repeated measures, Counterbalancing, Experimental versus observational studies, Confounding variables, Placebo effect, Correlation versus causation, Experimenter expectancy, Internal versus external validity.

On the quiz, You could be asked for definitions. You might be asked to make up an original example of a study requiring a t-test, chi-square test of independence, one-way ANOVA or correlation. You could be asked what's the Explanatory Variable, the Response Variable, and how to set up the data file. Here are some sample questions to think about.

1. Invent and briefly describe (in a few sentences at most) original studies with the following characteristics. Do not use any examples from lecture or the class notes. If the requested example is impossible, say so and explain why it is impossible. The word **original** is important. If you give an example that is overly similar to one from lecture or the class notes, your answer will receive a zero. If two people give exactly the same example, they will both get a zero for the question.

   a. A categorical explanatory variable and a continuous response variable.
   b. A continuous explanatory variable and a continuous response variable.
   c. A nominal scale explanatory variable and an ordinal scale response variable.
   d. An explanatory variable that is both quantitative and nominal scale, and a response variable that is continuous.
   e. Two categorical explanatory variables and two categorical response variables.
   f. A single categorical explanatory variable and two quantitative response variables

2. In a study relating IQ score to birth order, which is the explanatory variable and which is the response variable?

3. Make up an original example of a study that is multivariate, and both response variables are categorical.

4. Give an example of a variable for which it would be unreasonable to compute the standard deviation.

5. If p>.05, the results are significant and we can draw conclusions. True or False?

6. Is it possible for a variable to be both categorical and quantitative? Give an example.

7. What is the difference between a statistic and a parameter?

8. In simple regression, if the slope of the least-squares line equals zero, what is the value of the correlation coefficient r?

9. Explain how a single outlier could have a huge effect on the least-squares regression line. Draw a picture to illustrate your argument.

10. A medical researcher conducts a study using twenty-seven litters of cancer-prone mice. Two members are randomly selected from each litter, and all mice are subjected to daily doses of cigarette smoke. For each pair of mice, one is randomly assigned to Drug A and one to drug B. Time (in weeks) until the first clinical sign of cancer is recorded.
    a. What is the explanatory variable (or variables)?
    b. What is the response variable (or variables)?
    c. Indicate how the data file would be set up.
    d. How could the design be modified to allow comparison of 3 drugs and a placebo? Is this still "repeated measures?"
    e. Presumably the mice are so cancer-prone that they all come down with the disease eventually. But this might not happen, especially if one of the drugs is very effective. Discuss two ways of handling the data from a mouse that died of old age, and never showed signs of cancer. Find a problem with both solutions (there *will* be a problem, unless you know about survival analysis.
    f. In this study, suppose the sample means are exactly identical for the various drug treatments. Is it possible for the population means to be different?
11. What does it mean for two variables to be *related* in the population? Your answer must include the word "conditional," or it is wrong.
12. Is it possible to have a study with repeated measures and a categorical response variable? If it is possible, make up an original example. If it is impossible, explain why.
13. Is it possible for a study to be both experimental and observational? Explain.
14. It is well known that people who graduate from university have higher lifetime earnings on average than those who do not. Discuss at least one confounding variable that could have produced this result.
15. Is it possible for Explanatory Variable and Response Variable to be related in the population and unrelated in the sample?
16. Is it possible for Explanatory Variable and Response Variable to be related in the sample but unrelated in the population?
17. Is it possible for Explanatory Variable and Response Variable to be related in the population and related in the sample, but not significantly related?
18. Is it possible for Explanatory Variable and Response Variable to be related in the population and also significantly related in the sample, but in the wrong direction?
19. Suppose that volunteer patients undergoing elective surgery at a large hospital are randomly assigned to one of three different pain killing drugs, and one week after surgery they rate the amount of pain they have experienced on a scale from zero (no pain) to 100 (extreme pain).
    a. Indicate how the data file would be set up.
    b. What is the explanatory variable?
    c. What is the response variable?
    d. What statistical test would you recommend?
    e. Is this an experimental study, or observational?
    f. Why is it important for the patients to be unaware of which drug they are receiving? Relate this to the idea of a confounding variable.
    g. Is it also important for the physicians to remain unaware of what drugs their patients are getting? Why or why not?
    h. Is it also important for the person administering the questionnaire to remain unaware of what drug each patient is getting? Why or why not?
    i. In this study, suppose the population means were exactly identical for the various drug treatments. Would it be possible for Explanatory Variable and Response Variable to still be related in the population? Explain.
    j. What "population" are we talking about here, anyway?
20. In a study relating physical attractiveness to academic performance, six judges rated attractiveness on a 10-point scale, from photos of 100 randomly chosen first-year students. The data file contained 10 variables: Six attractiveness ratings, sex of student, number of credits completed by the end of first year, cumulative Grade Point Average (GPA) at the end of first year, and a binary variable

indicating whether the student was still enrolled at the end of first year. An eleventh variable, mean attractiveness rating, was calculated from the 6 ratings, and was taken to be the definition of "attractiveness."

    a. Is this an experimental study, or observational?

    b. Would it make sense to compute the average correlation of the attractiveness ratings with one another? What would it tell us, if anything? How many numbers would we be averaging?

    c. Which variables are explanatory variables, and which are response?

    d. What statistical test would you recommend for assessing the relationship between attractiveness and GPA?

    e. Give an example of an unmeasured variable that is a potential confounding variable in this study. Explain how it might produce an apparent relationship between attractiveness and GPA even if no real relationship existed.

    f. It is suggested that order of presentation be counterbalanced, so that each student has approximately the same mean order of presentation when averaging over judges. Why is this a good idea?

21. Answer the following questions T for true or F for false; write the letters on the lines. Assume the significance level is $\alpha = 0.05$ in all cases.

    a. _____ In an experimental study, a statistically significant relationship between the explanatory variable and the response variable can provide some evidence of a causal relationship.

    b. _____ In simple regression, a positive regression coefficient b_1 implies that high values of X tend to go with low values of Y and low values of X tend to go with high values of Y.

    c. _____ We observe r = -0.70, p = .009. We conclude that X and Y are unrelated.

    d. _____ An observational study is one in which cases are randomly assigned to the different values of an explanatory variable.

    e. _____ If p < .05 we say the results are statistically significant at the .05 level.

    f. _____ We seek to predict the response variable from the explanatory variable.

    g. _____ We observe r = 0.50, p = .002. This means that 50% of the variation in the response variable is explained by a linear relationship with the explanatory variable.

    h. _____ When a relationship between the explanatory variable and the response variable is statistically significant, we conclude there is no evidence that the two variables are actually related.

    i. _____ The greater the p-value, the stronger the evidence that the explanatory and response variable are related.

    j. _____ We would like to predict type of automobile owned (North American vs. other) from income, education and sex of owner. This is impossible, because the response variable is categorical.

22. In the table below, write the name of the most appropriate (least inappropriate?) statistical method in each cell. Use the tests in the following list. More than one test may be appropriate; in this case, write the names of both tests. I have filled in two of the cells, because the methods are not elementary.

    o Two-sample (independent) t-test

    o Paired (matched) t-test

    o Correlation

    o Simple regression

    o One-way Analysis of variance

    o Chisquared test of independence

	Response Variable		
Explanatory Variable	Categorical: Two Categories	Categorical: More than 2 Categories	Quantitative
Categorical: Two Categories			

Categorical: More than 2 Categories			
Quantitative	*Logistic Regression*	*Extension of Logistic Regression*	

23. The Little Statclass Data set has Sex, Race, Quiz average, Computer assignment average, Midterm score and Final Exam score from students in a statistics class, long ago. Using SAS, make frequency distributions of the Sex and Race, and compute n, mean and standard deviation, minimum and maximum for the other variables. Bring your log file (*not* your program file) and output file to the quiz. You may be asked to answer questions about them, and you may be asked to hand them in with the quiz. For example, what's the lowest midterm score? What's the mean score on the computer assignments? What percentage of students were from Race B?

If you are asked to hand in your log file and results file, your name and student number should be on both printouts. You are allowed to write your name and student number on the printouts in advance, but **do not write anything else on your printouts in advance**.

This should be obvious, but if you are asked for a number from your printout and you don't have a printout, do not answer the question and then pretend you remembered the number. If you do, you will be charged with an academic offence.

This also should be obvious, but you are not allowed to put answers or any other material related to the non-computer questions in comment statements, or otherwise cause such material to appear on your printout.

Note that your log file and output file *must be produced by the same run of SAS*. If they do not, you will get zero for the computer part of the quiz. It is surprisingly easy to mess this up, so please be careful.