Biostatistics 513

Homework 1 - due 4/8/13

Please submit your assignment in Word or PDF format by 9:30 am (PDT) to the Biost 513 Assignment Dropbox at http://canvas.uw.edu/. You may go to "Assignments" at the left side menus on your Canvas page and then follow the instructions to upload your assignment.

NOTE: Unless explicitly stated, direct computer output is not desired. Typically only part of the computer output is asked for (such as a confidence interval) and then proper interpretation of the statistics is requested.

DATA: The data for these exercises can be found on the class web page: http://courses.washington.edu/b513/ in the **Homework** and **Datasets** directories.

StataHelp1: Key Stata commands that are useful for these exercises are described in the text file StataHelp1.txt, also available in the homework directory.

2 x 2 Tables

1. (Pagano and Gauvreau, 1993) In a study of HIV infection among women entering the NYS prison system, 475 inmates were cross-classified with respect to HIV seropositivity and their history of intravenous drug use (Smith et al., 1991, *AJPH Supplement*). Use either direct calculations or STATA to answer the following questions. (Note: If using Stata, compare the cci and csi commands. Which is more appropriate?)

	HIV+	HIV-	Total
IVDU Yes	61	75	136
IVDU No	27	312	339
Total	88	387	475

   a) Test the null hypothesis that there is no association between IVDU and HIV serostatus. State the hypotheses, and write a sentence (or two) that explains the results to a general medical audience.

   b) Give a point estimate and 95% confidence interval for the risk difference and write a sentence (or two) that explains this interval to a general medical audience.

   c) Give a point estimate and 95% confidence interval for the relative risk, comparing IV drug using women to women that are not IV drug users, and write a sentence (or two) that explains this interval to a general medical audience.

   d) Give a point estimate and 95% confidence interval for the odds ratio, comparing IV drug using women to women that are not IV drug users, and write a sentence (or two) that explains this interval to a general medical audience.

   e) Based on these summaries, what do you conclude?

Binary data

2. The data HivnetWide.dat on the course web page contain baseline and follow-up data on a random subset of n = 1000 subjects (from the large cohort of ~5,000 subjects) that participated in an HIV vaccine preparedness substudy. This substudy was designed to assess the effects of the informed consent process on educating participants about the risks and benefits of participation in a phase III HIV vaccine trial. Specifically, half of the participants were given a mock informed consent between the baseline and the month 6 visit. The other half did not receive the mock informed consent. At baseline and followup participants were asked about their understanding of key vaccine trial concepts.

   These data contain a number of items but we are particularly interested in one: **`nurse`** asks whether participants understand randomization, specifically that it isn't the study nurse who decides to which treatment arm one will be assigned (placebo or active product); **`nurse`** was measured at baseline and 6 months (there is also a 12 month visit; ignore this). The other important variable is **ICgroup** which indicates whether or not the participant received the mock informed consent.

   Our analysis of these data will focus on the questions: Did the intervention appear to improve knowledge? Did improvements in knowledge result from "correction" of those subjects that answered incorrectly at baseline and/or "reinforcing" the knowledge of those subjects that answered correctly at baseline? Use STATA to answer the following questions focusing on the **`nurse`** item. Please provide written answers to these questions and only include the computer output as necessary.

   The .do file HivnetWide.do on the course website will read and format the data.

   a) Did randomization appear to balance the groups (treatment group is ICgroup==1, control group is ICgroup==0) with respect to baseline understanding of the nurse item?

   b) One analysis that is valid under randomization would compare the percent answering correctly at month 6 (cross-sectionally) for the intervention and control groups. Formulate the statistical hypothesis that would be used to test for a treatment effect, execute an analysis of the treatment effect, and provide summary rates, confidence intervals and estimates of the magnitude of the intervention effect.

   c) Another analysis would focus on the pre/post analysis for the <u>intervention group only</u> (ICgroup = 1). For the subset of patients that participated in the informed consent process we find:

		Month 6	
		Correct	Incorrect
Month 0	Correct	271	38
	Incorrect	146	45

   What is the hypothesis that is appropriate for these paired data? Execute an analysis of the paired data (hint: use the `mcc` command). Interpret the results.

d) To understand how the intervention may have worked we will subset on those subjects that did not correctly answer the nurse0 item, and within this subset we will compare the response at follow-up, nurse6, for subjects that were in the treatment versus control group. Form the 2 x 2 table that classifies nurse6 by ICgroup restricting to subjects for whom nurse0==0 (incorrect at baseline). What does this summary suggest about one way in which the informed consent process might have "worked"? Use appropriate statistical summaries to support your conclusion.

e) To understand how the intervention may have worked we will subset on those subjects that did correctly answer the nurse0 item, and within this subset we will compare the response at follow-up, nurse6, for subjects that were in the treatment versus control group. Form the 2 x 2 table that classifies nurse6 by ICgroup restricting to subjects for whom nurse0= =1 (correct at baseline). What does this summary suggest about one way in which the informed consent process might have "worked"? Use appropriate statistical summaries to support your conclusion.

Regression Thinking

3. Recall that in Biostat 512 the concept of a linear regression model was introduced and the interpretation of the model parameters was emphasized. Recall that models were specified for the *average* response in the form (e.g., regression model with two predictors)

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

and the interpretation of β_0, β_1, and β_2 was given. In addition, discussion of the interpretation of a model that also includes the interaction term $X_1 * X_2$, was discussed.

For binary data Y, (yes=1, no=0 or presence=1, absence=0), we can still think about the average response, $E[Y]$, and this is exactly the same as the probability of a "1", or P[Y=1]. (Please convince yourself.)

a. In 2(d), we restricted analyses to subjects with nurse0==0. We used the group variable X=ICgroup as our "predictor" and the item Y=nurse6 as the "response". Write the linear model for the average Y as a function of X and interpret the model parameters, being careful to mention that this is restricted to subjects with nurse0==0.

b. In 2(e), we restricted analyses to subjects with nurse0==1. We used the group variable X=ICgroup as our "predictor" and the item Y=nurse6 as the "response". Write the linear model for the average Y as a function of X and interpret the model parameters, being careful to mention that this is restricted to subjects with nurse0==1.

c. You can combine the two models that you have written in 2(a) and 2(b) into a single model with Y = nurse6, X_1=ICgroup, and X_2=nurse0. Write the model and interpret the regression parameters.

d. Although standard linear model thinking is quite useful for structuring the mean for binary data, we probably do not satisfy all the assumptions that are necessary to use standard linear regression methods for inference. Which of the standard assumptions may be violated in this example?