Biostat 513: **Exercise #1** DUE: Friday April 12, 2002

• Reading: Rosner 4th Ed. Chpt. 10 "Hypothesis Testing: Categorical Data" (*optional*)

---

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*). The data file `PRISON.DAT` contains the HIV serostatus in the first column and IV drug use history in the second column. Each row represents the record for one inmate. Note: the data is stored as **string** variables. Use either direct calculations or STATA to answer the following questions.

	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 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 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 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?

2. (HIVNET, 1995) The data `HivnetWide.dat` on the course web page `http://courses.washington.edu/b513` contain baseline and follow-up data on a random subset of $n = 1000$ subjects (from the large cohort of $\approx$ 5,000 subjects) that participated in an HIV vaccine preparedness study. At baseline and follow-up participants were asked about their understanding of key vaccine trial concepts. These data contain the response to two items: `qsafe` asks whether participants understood that a vaccine in a phase III study was not known with certainty to be safe (although having gone through rigorous phase I/II testing); and `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). A subset of approximately 20% of the cohort was asked to participate in a mock informed consent process between the baseline and the month 6 visit. This substudy was designed fully inform participants of the risks and benefit of participation in a phase III study. 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 (and looking at `qsafe` only if you have time/inclination). Please provide written answers to these questions and only include the computer output as necessary.

(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? Provide point estimates, confidence intervals, and tests as appropriate.

(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 intervention group only. For the subset of patients that participated in the informed consent process we find:

		Month 6 incorrect	correct
Baseline	incorrect	45	146
	correct	38	271

What is the hypothesis that is appropriate for these paired data? Execute an analysis of the paired data. Interpret the results both in terms of the results of formal tests, as well as the

estimated time (here time=intervention) comparison. (*optional* perform a similar analysis for the control group to see if there are improvements in knowledge just through study participation and not just through the IC process)

(e) 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 2×2 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.

(f) 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 2×2 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 we specified a model for the average response of the form (for example with 2 predictor variables)

$$\text{average } Y \;=\; E[Y] \;= \beta_0 \;+\; \beta_1 \cdot X_1 \;+\; \beta_2 \cdot X_2$$

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 $X_1 \cdot X_2$ was discussed.

For binary data Y (yes=1, no=0; 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 $P(Y = 1)$ (convince yourself).

(a) In 2(e) we restricted analysis 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(f) we restricted analysis 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) Can you combine the two models that you've written in 2(a) and 2(b) into a single model with Y=`nurse6`, X_1=`ICgroup`, and X_2 = `nurse0`. Write this model and interpret the parameters.

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