Biost 517
Emerson, Fall 01
November 2, 2001

Biost 517
Applied Biostatistics I

Midterm Examination

Name: _____

Instructions: Please provide concise answers to all questions. Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.

The examination is closed book and closed notes. If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

1. Suppose we are interested in studying whether higher values of Prostate Specific Antigen (PSA) are useful in the diagnosis of prostate cancer. For this problem we will regard that there is some "gold standard" for the diagnosis of prostate cancer (perhaps digital rectal exam and transrectal ultrasound directed needle biopsy) that is used when classifying patients as to their prostate cancer. Consider the following study designs for hypothetical studies done in King County:

   A. We randomly sample 200 men aged 50 - 80 years who have been recently diagnosed with prostate cancer and 400 men of similar ages who do not have prostate cancer. Each man has blood levels of PSA measured.

   B. Using hospital records of recently measured blood levels of PSA of men aged 50 - 80 years, we randomly select 500 men with PSA levels less than 4.0 and 500 men with PSA levels greater than 4.0. Each man is then examined for prostate cancer using the "gold standard" for diagnosis.

   C. We sample 1,000 men drawn randomly from the population of 50 - 80 year old men in King county. Each man has blood levels of PSA measured and is examined for prostate cancer via the "gold standard".

   a. (3 points) Which of the above study designs can provide an estimate of the prevalence of prostate cancer in the population of 50 - 80 year old men in King County?

   b. (3 points) Which of the above study designs can provide an estimate of the prevalence of PSA values less than 4.0 in the population of 50 - 80 year old men in King County?

   c. (3 points) Which of the above study designs can provide an estimate of the proportion of men with prostate cancer who have a PSA value greater than 4.0?

d. (3 points) Which of the above study designs can provide an estimate of the proportion of men without prostate cancer who have a PSA value less than 4.0?

e. (3 points) Which of the above study designs can provide an estimate of the proportion of men with PSA values above 4.0 who actually have prostate cancer?

f. (3 points) Which of the above study designs can provide an estimate of the proportion of men with PSA values below 4.0 who do not have prostate cancer?

g. (3 points) Which of the above study designs can provide information regarding an association between elevated PSA levels and prostate cancer? Justify your answer.

h. (3 points) Which of the above study designs can tell whether prostate cancer causes an elevation of prostate specific antigen in the blood?

i. (3 points) Which of the above study designs would be the easiest to perform logistically?

Appendix A contains descriptive statistics from an early phase clinical trial of beta carotene in cancer prevention. Fifty (50) patients were randomized to receive either 0, 15, 30, 45, or 60 mg/day of beta-carotene supplementation. The goal of the study was to assess how beta-carotene supplementation affected levels of beta-carotene and vitamin E in plasma. Measurements of plasma beta-carotene and vitamin E levels were to be obtained at randomization and after 3 and 9 months of treatment. An additional measurement was to be made 3 months after treatment stopped (12 months after randomization). For each observation in the dataset, the following data are available The following variables are available:

$PTID$	patient's identification number
AGE	patient's age in years
$RACE$	patient's race: 1= Asian, 2= Black, 3= White, 4= Other
$MALE$	patient's sex: 0= female, 1= male
$WEIGHT$	patient's weight in pounds
$DOSE$	treatment group (0, 15, 30, 45, or 60)
$MONTH$	time since randomization in months
$CAROT$	plasma beta-carotene level
$VITE$	plasma vitamin E level

2. (10 points) Based on the descriptive statistics presented in Appendix A, do any of the variables appear prone to outlying values? Explain your answer briefly.

3. 15 points How would you use the descriptive statistics presented in Appendix A? That is, how would such statistics aid you in the analysis of these data to answer the scientific question? Briefly explain.

Appendix B contains descriptive statistics from a study of sexual discrimination in salaries paid to university faculty. All nonadministrative faculty currently employed in 1995 were measured for the following variables

$EMPLID$	employee's identification number
$FEMALE$	employee's sex: 0= male, 1= female
$YRDEG$	year employee obtained highest degree
$FIELD$	employee's field (1= fine arts, 2= general, 3= professional)
$RANK95$	employee's rank (1= assistant prof, 2= assoc prof, 3= full prof)
$SLRY95$	employee's monthly salary in 1995
$TASSOC$	time spent as associate professor at the university (Assistant professors will have 0 for this variable, associate professors will have the years since they were promoted to associate, and full professors will have the time it took them to get promoted.)

4. 3 pts each part For each of the following descriptive statistics presented in Appendix B, identify the variables for which the specified statistic provides no scientifically meaningful descriptions of the sample. For each such variable, very briefly explain why not (just a few words should suffice to justify your entire answer).

   a. Mean

   b. Standard deviation

   c. minimum

d. maximum

e. 25th percentile

f. median

g. 75th percentile

5. (10 points) How would any of your answers to question 4 change if we were focusing on comparing distributions across sexes? Explain. (Ignore $FEMALE$ for this question.)

6. Suppose we are interested in comparing the time to promotion from Associate Professor to Full Professor for females and males.

   a. (10 points) What descriptive statistics would you want to see? How would they be computed?

   b. (BONUS: 10 points) How might time trends in promotion policies affect the analysis you suggested in part a? That is, what key assumptions for the analysis of these data might be invalidated by time trends in the policies regarding criteria for promotion to full professor?

7. (10 points each part) Appendix B also contains mean salary levels by faculty rank and faculty field.

   a. Is there evidence of an association between faculty rank and faculty salary? Briefly explain the evidence.

b. Is there evidence of an association between faculty field and faculty salary? Briefly explain the evidence.

c. Is there evidence of that faculty field modifies the association between faculty rank and faculty salary? Briefly explain the evidence.

8. (5 points each part) Suppose we study the association between mental functioning and age. Two independent researchers have performed studies and found the same estimated slope in a linear regression of mental function tests on age. However, researcher A found a correlation of -.17 between age and mental function tests, and researcher B found a correlation of -.42 between age and mental function tests. For each of the following scenarios, specify whether the study conditions might explain why such different correlations would be observed in the presence of similar slope estimates. Briefly justify your answer.

   a. Researcher A: 100 subjects of ages 50-100; all educational levels sampled

      Researcher B: 100 subjects of ages 50-100; only considered college graduates

   b. Researcher A: 100 subjects of ages 50-100; all educational levels sampled

      Researcher B: 100 subjects of ages 70-80; all educational levels sampled

   c. Researcher A: 100 subjects of ages 50-100; all educational levels sampled

      Researcher B: 50 subjects of ages 50-100; all educational levels sampled