Biost 517
Applied Biostatistics I

Midterm Examination Key
October 30, 2006

Name: ____ Disc Sect: M W F

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. You may use calculators, but you may not use any special programs written for programmable calculators.

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.

Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor on Wednesday.

PLEDGE:

On my honor, I have neither given nor received unauthorized aid on this examination:

Signed:____

Problems 1 – 7 relate to a hypothetical study of a new marker protein CBP (Completely Bogus Peptide) as a prognostic factor in patients with cancer. The following variables are available:

- **age=** patient's age at diagnosis in years
- **male=** indicator that patient's sex is male
- **type=** type of cancer: 1= colon, 2= lung, 3= brain, 4= breast, 5= liver
- **stage=** staging of cancer at diagnosis: 1= small tumor without local invasion, 2= tumor that has spread only into surrounding tissue, 3= tumor that has spread to lymph nodes, 4= tumor that has spread to distant organs
- **CBP=** levels of CBP measured in the blood in milligrams per deciliter
- **futime=** time of last follow-up (since cancer diagnosis and CBP measurement) in months
- **status=** patients' status at last follow-up: 0= alive, 1= dead

The following table contains descriptive statistics on the sample.

	N	Msng	Mean	SD	Min	25th %ile	Mdn	75th %ile	Max
Age	133	0	70.08	5.15	57.46	66.44	70.32	73.44	83.53
Male	133	0	0.56	0.50	0	0	1	1	1
Type	133	0	1.79	1.12	1	1	1	2	5
Stage	129	4	2.41	1.13	1	1	2	3	4
CBP	133	0	33.8	57.77	2.44	11.73	22.37	36.36	601.22
FUtime	133	0	21.04	11.72	0.15	12.07	20.86	31.00	48.50
Status	133	0	0.59	0.49	0	0	1	1	1

1. (6 points) Consider the patient's age.
   a. Which of the descriptive statistics mean, median, min, max, standard deviation, and quartiles would be scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

   Ans: All of them are OK. This is a continuous, quantitative variable.

   b. Which of those descriptive statistics could be useful when trying to compare distributions across populations. Briefly explain your reasons.

   Ans: All but the minimum and maximum. The sampling distribution of the minimum and maximum are too heavily influenced by the sample size. *(I also note that comparisons of the SD would be looking at spread, while comparisons on all others would be looking at location.)*

   c. Do the descriptive statistics provide evidence that the distribution of age is skewed? Briefly explain your reasons.

   Ans: No. The distribution of age looks pretty symmetric: Mean and median are approximately the same, and the median is the midpoint of the range.

2. (6 points) Consider the patient's sex.
   a. Which of the descriptive statistics mean, median, min, max, standard deviation, and quartiles would be scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

   Ans: Only the mean is of much interest for this binary variable, but all of them are OK.

   b. Which of those descriptive statistics could be useful when trying to compare distributions across populations. Briefly explain your reasons.

   Ans: Only comparison of the mean would make much sense.

   c. Do the descriptive statistics provide evidence that the distribution of sex is skewed? Briefly explain your reasons.

   Ans: This is largely an uninteresting question for a binary variable: Skewness would be a factor of the size of the mean..

3. (6 points) Consider the patient's type of cancer.
   a. Which of the descriptive statistics mean, median, min, max, standard deviation, and quartiles would be scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

   Ans: Only the sample size and the number of missing. The min and max might pick out some egregious coding errors. But this is an unordered categorical variable.

   b. Which of those descriptive statistics could be useful when trying to compare distributions across populations. Briefly explain your reasons.

   Ans: None of them would generally be acceptable for comparisons, because this is an unordered variable.

   c. Do the descriptive statistics provide evidence that the distribution of type of cancer is skewed? Briefly explain your reasons.

   Ans: This is an irrelevant question for unordered variables.

4. (6 points) Consider the patient's stage of cancer.
   a. Which of the descriptive statistics mean, median, min, max, standard deviation, and quartiles would be scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

   Ans: Neither the mean nor SD are of interest on this qualitatively ordered variable. The others are OK.

b. Which of those descriptive statistics could be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: The quantiles (25th, 50th, 75th) are certainly acceptable for an ordered variable. Comparisons of the mean would give qualitative indications of shifts in distribution, though the interpretation of the exact magnitude of any difference would be difficult to judge.

c. Do the descriptive statistics provide evidence that the distribution of type of cancer is skewed? Briefly explain your reasons.

Ans: This is not too relevant a question on an ordered categorical variable.

5. (6 points) Consider the patient's measurement of the marker CBP.

a. Which of the descriptive statistics mean, median, min, max, standard deviation, and quartiles would be scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: All of them are OK. This is a continuous, quantitative variable.

b. Which of those descriptive statistics could be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: All but the minimum and maximum. The sampling distribution of the minimum and maximum are too heavily influenced by the sample size. *(I also note that comparisons of the SD would be looking at spread, while comparisons on all others would be looking at location.)*

c. Do the descriptive statistics provide evidence that the distribution of CBP is skewed? Briefly explain your reasons.

Ans: Yes, very much so. The SD is larger than the mean of this positive random variable. The mean is larger than the median. The maximum is vastly further from the median than is the minimum.

6. (6 points) Consider the patient's time until death.

a. Which of the descriptive statistics mean, median, min, max, standard deviation, and quartiles would be scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: None of these descriptives (except N and number missing) are scientifically relevant, because some of the observations are censored. (KM-based estimates would need to be used.)

b. Which of those descriptive statistics could be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: None are suitable for this censored variable.

c. Do the descriptive statistics provide evidence that the distribution of time to death is skewed? Briefly explain your reasons.

Ans: This cannot be judged due to the censoring.

7. (6 points) Consider the patient's survival status.

a. Which of the descriptive statistics mean, median, min, max, standard deviation, and quartiles would be scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: None of these descriptives (except N and number missing) are scientifically relevant, because the survival status is measured over different timeframes. (KM-based estimates would need to be used.)

b. Which of those descriptive statistics could be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: None are suitable due to the censoring.

c. Do the descriptive statistics provide evidence that the distribution of survival status is skewed? Briefly explain your reasons.

Ans: This is an irrelevant question for this binary indicator of censoring status.

8. Suppose we are interested in studying whether a fast saliva test for antibodies to HIV can accurately diagnose a person infected with HIV. The "gold standard" for the diagnosis of HIV infection is a blood test that typically requires several days in order to obtain results. Consider the following study designs for hypothetical studies done at an HMO:
   - **Study A**: We sample 1,000 patients drawn randomly from the adult members of the HMO. Each patient has both the saliva test and the blood test performed.
   - **Study B**: Using hospital records of patients who recently had the saliva test performed, we sample 300 patients who had a positive saliva test and 700 patients who had a negative saliva test. We then perform the blood test on all of these patients.
   - **Study C**: Using hospital records of patients known HIV infection status as determined by the blood test, we sample 500 patients who are known to be HIV positive and 500 patients who are known to be HIV negative. We then perform the saliva test on all of these patients.

   a. (4 points) Which of the above study designs can provide an estimate of the prevalence of HIV infection among the HMO participants?

Ans: Only study A uses the cross-sectional sampling necessary to estimate the overall prevalence of HIV infection.

   b. (4 points) Which of the above study designs can provide an estimate of the prevalence of positive saliva tests among the HMO participants?

Ans: Only study A uses the cross-sectional sampling necessary to estimate the overall prevalence of test positivity.

   c. (4 points) Which of the above study designs can provide an estimate of the proportion of HIV positive patients at the HMO who will have a positive saliva test?

Ans: This can be answered by the cross-sectional sampling (study A) or the study in which sampling was stratified by HIV status (study C).

   d. (4 points) Which of the above study designs can provide an estimate of the proportion of HIV negative patients at the HMO who will have a negative saliva test?

Ans: This can be answered by the cross-sectional sampling (study A) or the study in which sampling was stratified by HIV status (study C).

   e. (4 points) Suppose we want to estimate what proportion of the saliva test positive patients will actually be HIV positive by the blood test. Which study designs can provide such an estimate?

Ans: This can be answered by the cross-sectional sampling (study A) or the study in which sampling was stratified by test positivity (study B).

   f. (4 points) Suppose we want to estimate what proportion of the saliva test negative patients will actually be HIV negative by the blood test. Which study designs can provide such an estimate?

Ans: This can be answered by the cross-sectional sampling (study A) or the study in which sampling was stratified by test positivity (study B).

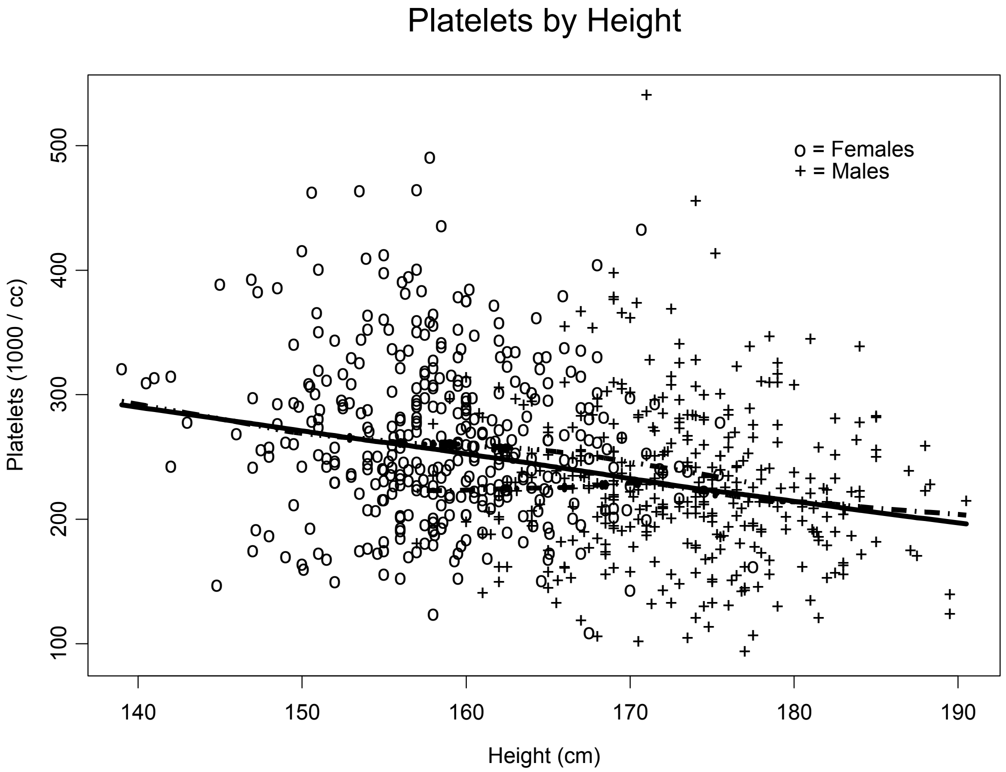
   g. (4 points) Which of the above study designs can provide information regarding an association between a positive saliva test and a positive blood test? Justify your answer.

Ans: All of them are OK. In study B we will compare the probability of HIV+ between patients positive and negative for saliva test. In study C we will compare the probability of positive saliva test between patients HIV+ and HIV-. In study A we could use either of those approaches.

   h. (4 points) Which of the above study designs would be the easiest to perform logistically?

Ans: My guess is that it would be easiest to do study C, because we could most easily identify patients by HIV status, and we could ensure adequate sample sizes in each group. The cross-sectional study might identify very few HIV+ patients.

9. The following scatter plot displays measurements of platelet counts (a cell fragment necessary for clotting blood) versus height for 735 elderly patients. Different symbols are used for each sex, and lowess smooths are superimposed on the plot for the entire sample, as well as each sex stratum.



a. (10 points) What observations would you make about this descriptive analysis?

<u>Ans</u>**: There does not appear to be any marked outliers. There is an overall trend toward decreasing levels of platelets among the taller patients. A straight line would appear to fit that trend reasonably well. The variability of the data appears relatively constant across height groups.**

b. (5 points) Would you expect the sample correlation between platelet count and height to be positive, near zero, or negative in the combined sample?

<u>Ans</u>**: The downward slope would suggest a negative correlation.**

c. (5 points) What would you guess the actual value of the sample correlation between platelet count and height would be in the combined sample? (You get this answer right, so long as it agrees with your answer to part b. I am just sampling the class's ability to guess the correlation.)

<u>Ans</u>**: Turns out the correlation is -0.29.**

d. (10 points) If we were to compute the correlations for each sex separately, how do you think they would differ from the correlation in the combined sample? Explain your reasoning.

<u>Ans</u>**: The slopes of the lowess curves look approximately the same for both sexes. Similarly, the spread of the data with height groups looks about the same for each sex separately, as well as in the combined sample. However, the variability of height is much less for each sex group compared to the combined sample. Hence, we would expect the correlation to be closer to zero in each of the sex strata than in the combined sample.**

10. The following table provides descriptive statistics about starting monthly salaries (dollars) for assistant professors hired at a particular university. Statistics are provided for all academic fields combined, as well as within strata identified by Arts, Sciences, and Professional schools.

		N	Mean	SD	Min	25th %ile	Mdn	75th %ile	Max
All Fields	**Male**	77	3895	832	2556	3171	3688	4603	6356
	Female	63	3412	698	1938	2955	3341	3711	5222
Arts	**Male**	14	3138	234	2694	2998	3106	3348	3559
	Female	27	3072	397	2418	2667	3263	3357	3711
Sciences	**Male**	26	3537	629	2556	3110	3379	3918	5305
	Female	27	3437	691	1938	2980	3313	3873	5222
Professional	**Male**	37	4433	744	2913	3962	4538	4718	6356
	Female	9	4360	551	3586	3764	4651	4731	5010

a. (10 points) When all fields are considered together, do the descriptive statistics suggest an association between sex and starting salary? How would you quantify any such association?

Ans: Yes. The mean salary for men is $483 per month higher for men than women.

b. (10 points) Is there evidence that academic field confounds the description of an association between sex and starting salary? Briefly describe the issues you consider in answering this question, providing descriptive statistics in support of your conclusion as appropriate.

Ans: Academic field is a strong predictor of salary independent of sex. For instance, looking at males: males in professional fields average $1,295 more per month than males in the arts, and males in the sciences average $399 more per month than males in the arts. Also, academic field and sex are associated in the sample: 27/41 faculty members in the arts are female (nearly a 2:1 ratio of F:M), while only 9/43 faculty members in the professional fields are female (approximately a 1:4 ratio of F:M). The decision for confounding would then rest with whether we thought that any sex discrimination was being effected through hiring practices in the different fields (e.g., do we pay people in the arts less only because that field is mostly women?). Overall, I would think we would first consider this confounding, because we do tend to recognize that fields receive different salaries for economic reasons separate from the sex makeup of the workforce.

c. (10 points) Using the above descriptive statistics, how would you quantify the evidence for or against possible sex discrimination in starting salaries?

Ans: I would have accepted you arguing for either an adjusted analysis due to the confounding, or for an unadjusted analysis because you worried about the causal pathway. In the case of the adjusted analysis, I would take the average of the stratum specific differences: (66 + 100 + 73) / 3 = 79.7, so I would report that men average $79.7 per month higher than women in the same academic field .