

Biost 517
Applied Biostatistics I

Midterm Examination Key
November 2, 2007

Name: _____ Disc Sect: M W F

Instructions: Please provide concise answers to all questions. The exam is worth a total of 137 points.

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 Monday.

PLEDGE:

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

Signed: _____

Problems 1 – 9 relate to a study of patients with a particular disease of the liver, primary biliary cirrhosis (PBC). The following variables are available:

- **enroll**= Date of enrollment in the study (MMDDYY format)
- **age**= Age (years)
- **sex** = Sex (0=male, 1=female)
- **albumin**= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
- **bili**= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
- **cholest**= Serum cholesterol (mg/dl)
- **edema**= Presence of edema (swelling of extremities) (0= no, 1=yes)
- **sgot** = Serum SGOT (an enzyme found in liver cells) (U/l)
- **spiders** = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
- **stage**= Stage of disease (1= best, 2, 3, or 4= worst)
- **obstime**= Observation time from enrollment until death or until end of study, whichever comes first (years)
- **status**= Survival status at time of last observation time (0=still alive, 1= dead)

The following table contains descriptive statistics on the sample.

Variable	N	Mean	SD	Min	25 th %ile	Median	75 th %ile	Max
enroll	418	68475	35275	10192	32791	70891	100791	123090
age	418	50.7	10.4	26.3	42.8	51	58.3	78.4
sex	312	0.885	0.32	0	1	1	1	1
albumin	418	3.5	0.42	1.96	3.24	3.53	3.77	4.64
bili	418	3.22	4.41	0.3	0.8	1.4	3.4	28
cholest	284	370	232	120	249	310	400	1775
edema	418	0.12	0.325	0	0	0	0	1
sgot	312	122.6	56.7	26.4	80.6	114.7	151.9	457.3
spiders	312	0.288	0.454	0	0	0	1	1
stage	312	3.032	0.878	1	2	3	4	4
obstime	418	5.49	3.08	0.11	3.26	5.04	7.40	12.47
status	418	0.385	0.487	0	0	0	1	1

1. (6 points) Consider the sample means presented in the above table.
- a. For which of the variables would the mean **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: *enroll* is essentially a nominal variable, because the MMDDYY format does not allow any sensible ordering. *stage* is an ordered categorical variable, and standard arithmetic operations do not necessarily make any sense. *obstime* and *status* record the censored time to death, and using standard descriptive statistics does not address any scientific questions.

- b. For which of the variables would the mean **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: The sample mean is not useful for comparing distributions of *enroll*, *obstime*, and *status* for the same reasons that it is not useful descriptively.

(Note: On the other hand, the sample mean could be used to detect shifts in distribution for *stage*, even though the quantification of any such shift based on the sample mean will be problematic.)

2. (6 points) Consider the sample medians presented in the above table.
- a. For which of the variables would the median **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: *enroll* is essentially a nominal variable, because the MMDDYY format does not allow any sensible ordering. *obstime* and *status* record the censored time to death, and using standard descriptive statistics does not address any scientific questions. *sex*, *edema*, and *spiders* are binary variables for which the sample median provides very little information, though it is not inappropriate to consider it.

(Note: The sample median is valid for any ordered variable. *stage* is an ordered categorical variable, and hence the sample median is well-defined for that variable.)

- b. For which of the variables would the median **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: The sample median is not useful for comparing distributions of *enroll*, *obstime*, and *status* for the same reasons that it is not useful descriptively. Similarly, the sample median is largely useless for comparing the binary variables *sex*, *edema*, and *spiders* across groups.

3. (6 points) Consider the sample standard deviations presented in the above table.
- For which of the variables would the standard deviation **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: Because it is based on the mean squared distance from the average measurement, the sample standard deviation is not useful for the same variables and the same reasons that the sample mean would not be useful: *enroll* is nominal; *stage* is ordered categorical; *obstime* and *status* represent a censored time to event. In addition, the standard deviation does not really provide much information for the binary variables *sex*, *edema*, and *spiders*.

- For which of the variables would the standard deviation **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: The sample standard deviation is not useful for comparing distributions of *enroll*, *obstime*, and *status* for the same reasons that it is not useful descriptively. Similarly, there would be no advantage in comparing the distribution of the binary variables *sex*, *edema*, and *spiders* using the standard deviation, because those standard deviations are wholly determined by their means (which are in turn just the proportions).

*(Note: On the other hand, the sample standard deviation could be used to detect changes in the variability for *stage*, even though the quantification of any such change based on the sample standard deviation will be problematic.)*

4. (6 points) Consider the sample minima and maxima presented in the above table.
- For which of the variables would the minimum and maximum **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: *enroll* is essentially a nominal variable, because the MMDDYY format does not allow any sensible ordering. *obstime* and *status* record the censored time to death, and using standard descriptive statistics does not address any scientific questions. *sex*, *edema*, and *spiders* are binary variables for which the sample minimum or maximum provides very little information, though it is not inappropriate to consider it.

- For which of the variables would the minimum and maximum **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: The minimum and maximum is clearly not useful to compare distributions of variables for which they are not useful descriptively. Furthermore, because the sampling distributions of these statistics are heavily influenced by the sample size, they are generally not useful when trying to make inference across distributions.

5. (6 points) Consider the sample 25th and 75th percentiles presented in the above table.
- For which of the variables would the 25th and 75th percentile **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

Ans: The 25th and 75th percentiles are useful exactly when the median would be. Hence, the answer is the same as that to problem 2a.

- b. For which of the variables would the 25th and 75th percentile **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

Ans: The 25th and 75th percentiles are useful exactly when the median would be. Hence, the answer is the same as that to problem 2a.

6. (10 points) Where relevant, indicate which of the above variables appear to have a skewed distribution due to outlying values. Briefly explain your reasons.

Ans: *bili* shows a standard deviation larger than the mean, which for this positive random variable would suggest the presence of a markedly skewed distribution. Also, the median and mean are not equal, and the distance between the median and the minimum is markedly less than the distance between the maximum and the median. The degree of the skewness implied by these descriptive statistics would make the presence of outlying values highly probable.

***cholest* shows a standard deviation larger than half the mean, which, while not as striking as that for *bili*, is suggestive of a skewed distribution for a positive random variable. Also, the median and mean are not equal, and the median is not midway between the minimum and maximum. The degree of the skewness implied by these descriptive statistics would make the presence of outlying values highly probable.**

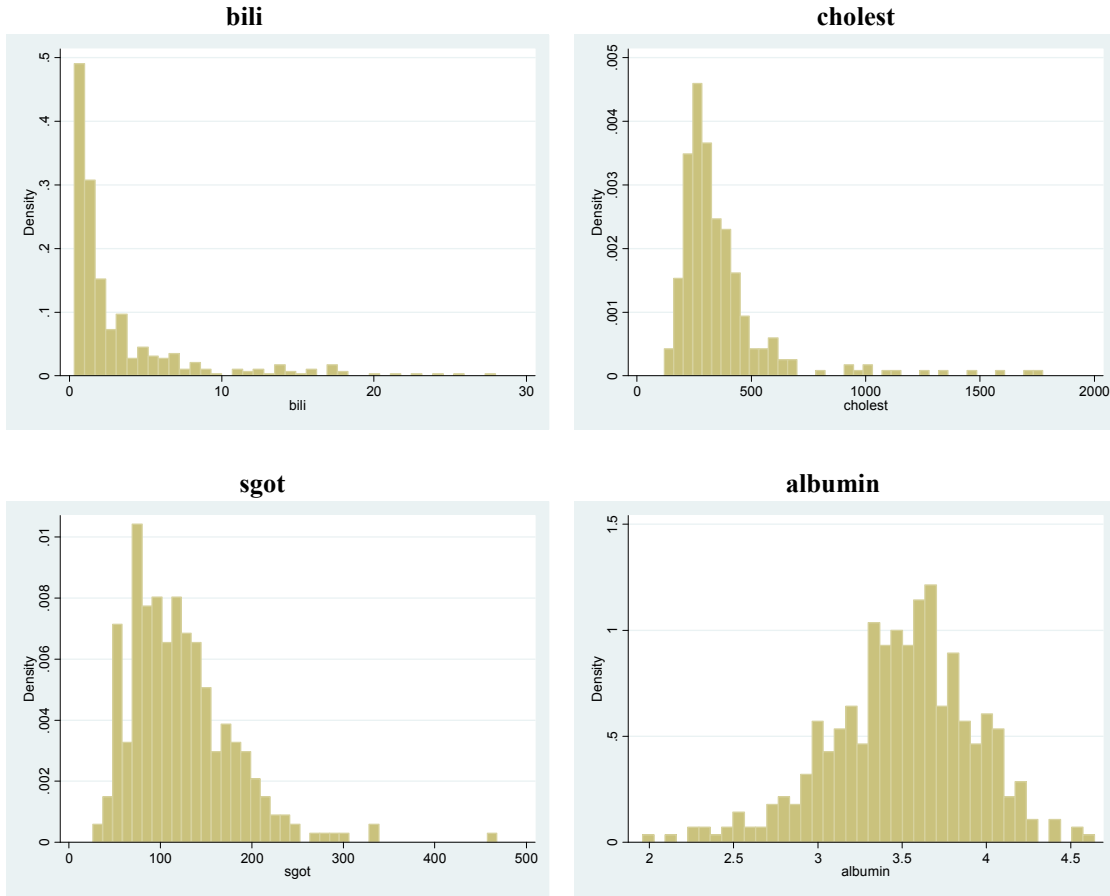
***sgot* shows a standard deviation larger than one-third the mean, which, while not as striking as that for *bili* or *cholest*, is suggestive of a somewhat skewed distribution for a positive random variable. The mean is only slightly higher than the median, but the maximum is markedly further from the median than is the median. The degree of the skewness implied by these descriptive statistics would make the presence of outlying values probable.**

***albumin* shows a minimum that is further from the mean than is the maximum. This is not very striking, and I am not really impressed by the possibility of low outlying values that would be problematic. But it is likely that there is some amount of skewness to the left.**

*Note: Our interest in looking for outliers is to 1) look for possible data entry errors, and 2) to be forewarned of cases that might have extreme "leverage" in analyses. I note that none of the "outlying" values in this dataset are data entry errors. Highly leveraged predictors have the capability to exert undue influence in an analysis if they also happen not to follow general trends that were observed for the response in the rest of the data. Admittedly, a part of my answers given above are based on my experience in judging when "outlying" values might be prone to be highly leveraged. I am most worried about *bili*, *cholest*, and *sgot*, and not so worried about *albumin*.*

It is not appropriate to try to assess outlying values of nominal variables using the descriptive statistics provided in the table. It is generally not too useful to try to use such criteria for ordered categorical variables, either. It is generally silly to talk about outlying values for binary variables, as the proportion would tell you everything you wanted to know. The relevance of outlying values for the censored observations is generally lacking without knowing whether any extremely large values are censored or not.

*Just for interest's sake I provide histograms of the variables identified above, as well as for *albumin* which might have been prone to very small values rather than very large values, because it was the distance between the median and the minimum that was slightly larger than the distance between the median and the maximum. From these plots it is clear that *bili*, *cholest*, and *sgot* are indeed subject to the presence of some "outlying" cases. Personally, the low *albumin* cases do not worry me from the standpoint of being possibly influential.*



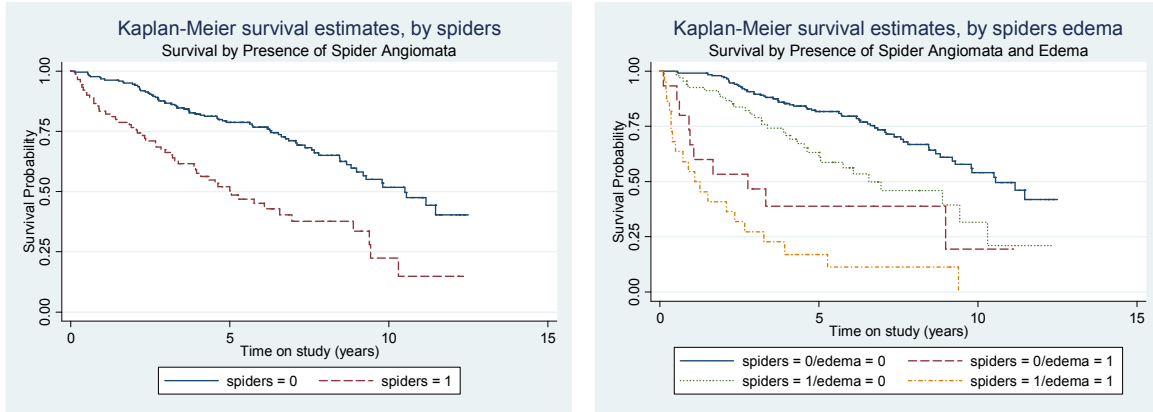
7. The following table presents descriptive statistics for selected variables according to whether the patient did or did not have spider angiomas present.

Variable	N	Mean	SD	Min	25 th %ile	Median	75 th %ile	Max
<i>Patients without angiomas present</i>								
age	222	50.4	10.6	28.9	42.6	50.1	57	76.7
sex	222	0.856	0.352	0	1	1	1	1
albumin	222	3.58	0.4	2.1	3.35	3.6	3.85	4.64
bili	222	2.43	3.57	0.3	0.7	1.1	2.3	28
edema	222	0.068	0.252	0	0	0	0	1
obstime	222	6.01	3.01	0.11	3.69	5.92	8.19	12.47
status	222	0.329	0.471	0	0	0	1	1
<i>Patients with spider angiomas present</i>								
age	90	49.2	10.5	26.3	41.6	49.5	56	78.4
sex	90	0.956	0.207	0	1	1	1	1
albumin	90	3.37	0.42	1.96	3.12	3.42	3.63	4.19
bili	90	5.3	5.84	0.5	1.3	3.2	6.5	24.5
edema	90	0.244	0.432	0	0	0	0	1
obstime	90	4.21	2.87	0.14	2.05	3.94	6.07	12.32
status	90	0.578	0.497	0	0	1	1	1

a. (10 points) How would you use the above descriptive statistics to assess whether the presence of spider angiomas is associated with shortened survival?

Ans: Because our measurements of time to death are censored for some subjects, I would not (and should not) use the above descriptive statistics to try to assess any such association. I would instead use Kaplan-Meier estimates or other analytic techniques appropriate for censored data.

8. The following are results from a Kaplan-Meier analyses of the time to death within strata defined by the presence of spider angiomata, the presence of edema, or all combinations of those two signs of liver disease.



	Spider Angiomata Absent			Spider Angiomata Present		
	n	Survival Probability		n	Survival Probability	
		2 Year	5 Year		2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

- a. (10 points) Based on the above statistics, would you conclude that there is overall an association between the presence of spider angiomata and the probability of survival? Provide statistics to quantify your answer.

Ans: From the lefthand panel displaying curves stratified by the presence of spider angiomata, it is clear that there is a tendency for patients without spider angiomata to survive longer than those who do have them. Choosing the difference in 5 year survival probability to quantify the effect, I note that there is a $0.788 - 0.520 = 0.268$ absolute difference in the probability of surviving 5 years in favor of the group without spider angiomata.

(Because the problem did not specify the particular probability measure to be used to quantify the association, you were free to choose whichever seemed most appropriate. You could, of course, have chosen the two year survival probability, and you could have expressed the comparison as a ratio of the probabilities instead of the difference. It would not be correct to have used the estimates within the strata defined by edema: The question did not specify any sort of stratum specific effect.)

- b. (10 points) Based on the above statistics, would you conclude that the presence of edema modifies any association between the presence of spider angiomata and survival? Provide statistics to quantify your answer.

Ans: Choosing the difference in 5 year survival probability to quantify the effect, I note that in the group of patients having no edema, there is a $0.817 - 0.613 = 0.204$ absolute difference in the probability of surviving 5 years in favor of the group without spider angiomata, while in the group of patients having edema, there is a $0.379 - 0.171 = 0.208$ absolute difference in the probability of surviving 5 years in favor of the group without spider angiomata. Hence, by these measures, the association between survival and the presence of spider angiomata is strikingly similar, and I would not say that there is effect modification.

(Again, because the problem did not specify the particular probability measure to be used to quantify the association, you were free to choose whichever seemed most appropriate. Had you chosen to quantify the magnitude of the association by the difference in two year survival probabilities, you would also have found very similar measures of association. And I note that you could have assessed the existence of any effect modification by considering the difference across edema strata within groups having similar presence or absence of spider angiomas: If edema does not modify the association between survival and spider angiomas, then we also know that the presence or absence of spider angiomas does not modify any association between survival and edema.

I do note that had you chosen to quantify an association by looking at the ratio of five year survival probabilities, then we would have found that among patients with no edema, the five year survival is $0.817 / 0.613 = 1.33$ fold higher for patients without spider angiomas, while among patients with edema, the five year survival is $0.379 / 0.171 = 2.26$ fold higher for patients without spider angiomas. Hence, using this measure of association, we clearly find a difference between the edema and no edema strata, and we would decide that edema modified the association between survival and the presence/absence of spider angiomas.

Take home message: Deciding whether there is or is not effect modification can be affected by your choice of summary measure to compare across groups and the measure (difference or ratio) you use for that comparison.

You would, of course, receive full credit for any correct statement you made, so long as you justified your response by presenting the quantification of the effect across strata.)

- c. (10 points) Based on the above statistics, would you conclude that the presence of edema confounds any association between the presence of spider angiomas and survival? Provide statistics to quantify your answer.

Ans: From the sample sizes in each stratum, we see that 9 of 211 (or 4.3%) of patients with spider angiomas have edema, while 10 of 71 (or 14.1%) of patients without spider angiomas have edema. Thus there does seem to be an association between edema and spider angiomas in the sample.

Furthermore, from the righthand panel, it is clear that in each stratum defined by presence or absence of spider angiomas, the survival curve for patients with edema is below that of the patients without edema. This suggests that edema is associated with survival after adjusting for spider angiomas. This can be quantified by noting that among patients without spider angiomas the difference in five year survival probability across edema strata is $0.817 - 0.379 = 0.438$. Similarly, in the patients with spider angiomas, the difference in five year survival probability is $0.613 - 0.171 = 0.442$.

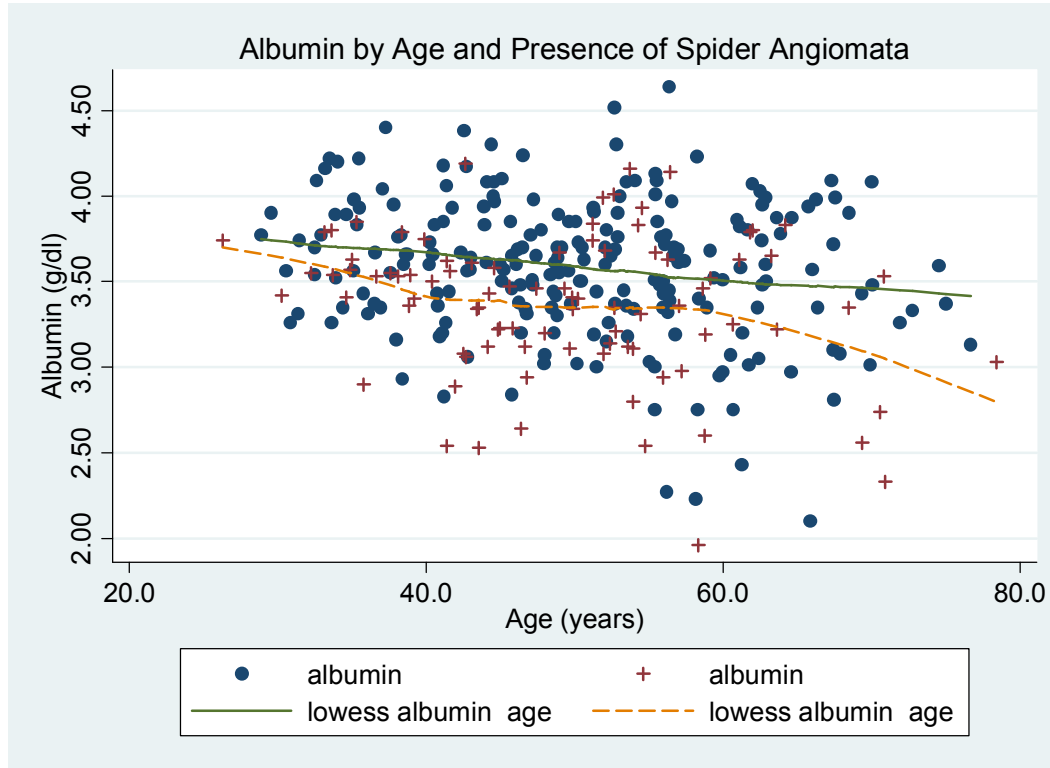
Thus, providing the association between survival and edema is causal in nature and not in the causal pathway of interest, I would contend that edema does confound the association between survival and spider angiomas.

(Edema is sometimes a sequela of advanced liver disease, and thus when viewed as a surrogate for such advanced liver disease, we would regard the association as causal. I note that the development of edema is associated with the liver's inability to make albumin, while the development of spider angiomas is believed to be associated with the liver's inability to break down estrogens. Hence, there is not really any strong argument for one of these risk factors to be in the causal pathway of the other. But if you stated that there was no confounding due to the causal pathway, that was an acceptable answer to me.

Proportions are means, and because I was using the differences of means as my measure of association, I can assess the presence of confounding by comparing the results of an unadjusted analysis (as in part a) and the results of stratum specific analyses (as were described in my answer to part b). In this case, I found a difference of five year survival probabilities of 0.268 in the unadjusted analysis, but stratum specific differences of approximately 0.206 (taking the average across the edema strata). This difference in estimates is evidence of confounding. It should be noted however, that when using odds ratios or hazard ratios as our

measure of association, we cannot rely on a comparison of unadjusted and adjusted analyses to identify the presence of confounding. In those settings, we need to rely on the approach based on first principles.)

9. The following scatter plot displays measurements of serum albumin levels versus age for the sample. Different symbols are used according to whether the patient had spider angiomata (+) or did not have spider angiomata (●). Lowess smooths are superimposed for each stratum defined by the presence of spider angiomata (solid=no angiomata, dashed= spider angiomata present).



A scatterplot of albumin levels by age. Patients without spider angiomata are represented with solid dots (●), with a lowess line displayed as a solid line. Patients having spider angiomata are represented with plus (+), and a lowess line displayed as a dashed line.

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

Ans: I would make the following observations :

- there are no obvious outliers,
- there is a general tendency toward lower albumin levels among the older age groups, with perhaps more of a downward trend among those patients having spider angiomata,
- the trend in each group would appear to be well approximated by a straight line,
- there does appear to be greater variability of measurements about the means in the older age groups (with that same tendency evident in both the patients with and without spider angiomata),
- the distribution of ages appears roughly similar for patients with or without spider angiomata (and this is confirmed by the table given in problem 7 above), and

- **there does appear to be a tendency toward lower albumin levels among patients having spider angiomas when compared to patients of the same age who do not have spider angiomas.**

(You were expected to comment on at least the first four of these.)

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

Ans: As the slope is downward to the right, the correlation will be negative. *(The actual correlation is -0.18. Typically, a correlation this close to 0 is hard to detect without a smooth.)*

- c. (10 points) If we were to compute the correlations for each stratum separately (i.e., for those with and those without spider angiomas), how do you think they would differ from the correlation in the combined sample? Explain your reasoning.

Ans: The variability of age is approximately the same in each stratum and the combined sample, so this would not tend to make the stratum specific correlation very different from that in the combined sample. The slope in the spider angiomas stratum is maybe a bit steeper than that in the no spiders stratum. This might make the correlation in the spiders group more negative than that in the combined group, which would in turn be more negative than that in the no spiders group, because the slope in the combined group should look something like an average of the two strata specific slopes (with a slight weighting toward the slope seen in the large group.) Because there is vertical separation between the lowest smooths for the two groups, this does argue that the within age group variation of albumin will be less for the individual strata than it will be in the combined sample. This last aspect will tend to make the correlation more extreme (further from 0 in the negative direction) for the individual strata than for the combined sample. *(The actual correlations are -0.21 in the no spiders group and -0.23 in the spiders group.)*

10. Suppose we are interested in studying whether expression of gene DCC can accurately predict the presence of metastases from colon cancer (“metastases” are instances in which the cancer has spread far from its original site). The “gold standard” for the diagnosis of metastases would be based on extensive radiologic examination and surgical exploration. Consider the following study designs for hypothetical studies done at an HMO:

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- a. (4 points) Which of the above study designs can provide an estimate of the prevalence of metastases among all colon cancer patients?

Ans: Only study B (the cross-sectional study).

- b. (4 points) Which of the above study designs can provide an estimate of the prevalence of positive DCC expression among all colon cancer patients?

Ans: Only study B (the cross-sectional study).

- c. (4 points) Which of the above study designs can provide an estimate of the patients having a positive DCC test among the patients with metastatic colon cancer? What is this probability usually called?

Ans: The sensitivity of the test can be estimated from either study B (the cross-sectional study) or study A (the study with stratified sampling based on disease status).

- d. (4 points) Which of the above study designs can provide an estimate of the patients having a negative DCC test among the patients without metastatic colon cancer? What is this probability usually called?

Ans: The specificity of the test can be estimated from either study B (the cross-sectional study) or study A (the study with stratified sampling based on disease status).

- e. (4 points) Suppose we want to estimate what proportion of the DCC positive patients will actually have metastatic colon cancer. Which study designs can provide such an estimate? What is this probability usually called?

Ans: The predictive value of the positive (or positive predictive value) of the test can be estimated from either study B (the cross-sectional study) or study C (the study with stratified sampling based on test result).

- f. (4 points) Suppose we want to estimate what proportion of the DCC negative patients will actually be free of metastases. Which study designs can provide such an estimate? What is this probability usually called?

Ans: The predictive value of the negative (or negative predictive value) of the test can be estimated from either study B (the cross-sectional study) or study C (the study with stratified sampling based on test result).

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

Ans: Any of the study types can assess an association, because an association between metastases and positive DCC test would be defined by any of the following

- $\Pr(\text{Pos DCC} \mid \text{Metastases}) \neq \Pr(\text{Pos DCC} \mid \text{No Metastases})$ (or $\text{Sensitivity} \neq 1 - \text{Specificity}$) as can be evaluated in studies B or A;
- $\Pr(\text{Neg DCC} \mid \text{No Metastases}) \neq \Pr(\text{Neg DCC} \mid \text{Metastases})$ (or $\text{Specificity} \neq 1 - \text{Sensitivity}$) as can be evaluated in studies B or A;
- $\Pr(\text{Metastases} \mid \text{Pos DCC}) \neq \Pr(\text{Metastases} \mid \text{Neg DCC})$ (or $\text{PV}^+ \neq 1 - \text{PV}^-$) as can be evaluated in studies B or C;
- $\Pr(\text{No Metastases} \mid \text{Neg DCC}) \neq \Pr(\text{No Metastases} \mid \text{Pos DCC})$ (or $\text{PV}^- \neq 1 - \text{PV}^+$) as can be evaluated in studies B or C.

(Note that if one of the above criteria is true, they all must be. That is why we can assess association from cross sectional studies, cohort studies, or case-control studies.)

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

Ans: As a general rule, with a rare disease, the case-control study is the easiest to perform. This would typically be the most common type of study to do in this case, because we would identify the patients through a registry, and minimize the number of genetic analyses we would have to do.

DISTRIBUTION OF GRADES:

n	Mean	SD	Max	Percentiles								
				90 th	80 th	70 th	60 th	50 th	40 th	30 th	20 th	10 th
64	105.4	16.9	134.0	125.1	120.4	116.1	110.6	106.0	102.0	96.9	92.0	86.6