

**Biost 517
Applied Biostatistics I**

**Final Examination Key
January 17, 2002**

(As always, this key contains more detail than I necessarily expected from you. Usually (but not always) the extra detail is put in regular italics.)

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.

Problems 1 - 5 refer to a study of the association between findings on magnetic resonance imaging (MRI) of the brain and cognitive function and survival in the elderly. The variables available in this data set include

- **Age** = the patient's age in years
- **Male** = an indicator of the patient's sex (0 = female, 1= male)
- **Marital** = a code indicating the patient's marital status (1= married, 2= widowed, 3= divorced, 4= never married, 5= other)
- **Health** = a code indicating the patient's self-reported health status (1= excellent, 2= very good, 3= good, 4= average, 5= poor)
- **DSST** = score on a test of mental function (possible scores 0 - 150)
- **Infarct** = an indicator of MRI evidence of a brain lesion due to lack of oxygen (0= none, 1= infarct present)
- **Obstime** = time of follow-up in days from start of study until death or the time of data analysis, whichever comes first
- **Death** = indicator that a death was observed

	n	mean	std dev	min	25%-ile	median	75%-ile	maximum
age	58.00	75.34	5.22	67.00	71.25	75.00	78.00	90.00
male	58.00	0.60	0.49	0.00	0.00	1.00	1.00	1.00
marital	58.00	1.38	0.95	1.00	1.00	1.00	1.00	5.00
health	58.00	2.50	1.05	1.00	2.00	2.00	3.00	5.00
DSST	58.00	63.53	29.85	15.00	40.00	57.00	86.50	130.00
infarct	58.00	0.43	0.50	0.00	0.00	0.00	1.00	1.00
obstime	58.00	1539.22	378.29	231.00	1473.25	1610.50	1773.25	2103.00
death	58.00	0.19	0.40	0.00	0.00	0.00	0.00	1.00

1. The above table provides the sample size, sample mean, median, standard deviation, minimum, maximum, 25th and 75th percentiles for the above data. For each of the above variables briefly indicate which of the above descriptive statistics are not of particular scientific interest for this study. Also indicate any other descriptive statistics you might wish to look at.

a. Age

Ans: All of the above descriptive statistics are meaningful for this uncensored, continuous random variable.

b. Male

Ans: The sample mean, because it is the proportion of males, is of interest for this binary variable. The remaining statistics are not particularly interesting (though not incorrect).

c. Marital

Ans: None of the above descriptive statistics are meaningful for this unordered categorical (nominal) random variable. I would rather see a frequency table giving the proportion in each category.

d. Health

Ans: The mean and standard deviation are not interpretable for this ordered categorical variable. All others are interpretable. Given the small number of categories, I might still prefer a frequency table.

e. DSST

Ans: All of the above descriptive statistics are meaningful for this uncensored, continuous random variable.

f. Infarct

Ans: The sample mean, because it is the proportion of patients having infarcts, is of interest for this binary variable. The remaining statistics are not particularly interesting (though not incorrect).
(Some students thought this variable might measure incident infarcts during the study. In fact, the MRI exam was performed at the start of the study, but I gave credit if you considered this to be the indicator of a censored event.)

g. Obstime

Ans: None of the above descriptive statistics are meaningful for this censored, continuous random variable. Instead, I would want to see Kaplan-Meier estimates of survival at various time points and/or quantiles of the survival distribution estimated from the Kaplan-Meier functions.

h. Death

Ans: None of the above descriptive statistics are meaningful for this binary variable indicating events over varying time periods. Instead, I would want to see Kaplan-Meier estimates of survival at various time points and/or quantiles of the survival distribution estimated from the Kaplan-Meier functions.

2. For each of the following scientific questions, indicate two possible statistical approaches that you might use for statistical inference. Be sure to identify what aspect of the distribution (e.g., a population parameter) that you are comparing, and indicate whether your method of inference can include a confidence interval as well as a test of a hypothesis.

Ans: There are, of course, many options that could be used for each answer, thus I provide more than two for each part. I accepted any answers that showed that you really knew how you would do the test. Tests that I list in parentheses are actually valid, but you would have to be able to justify the

use of that test in the situation, as it would be widely regarded as quite nonstandard in the situation.

a. Does the distribution of age differ by sex?

Ans: We are comparing the distribution of a quantitative, continuous random variable across two groups. Possibilities include:

Comparing the difference of the means using the t-test allowing for the possibility of unequal variances. There is a corresponding confidence interval.

Comparing the difference of the means using the t-test presuming equal variances. The confidence intervals that correspond to this test are only correct if the variances are truly equal under all alternatives for the mean. (*In a randomized experiment, we might be willing to assume equal variances under the null hypothesis that the intervention does not affect the distribution in any way. However, if there is any possibility that the variances could differ across the groups, at the stated level of confidence we can really only regard that rejection of the null indicates a difference in the distributions. In that setting, we could not really trust the CI. The t-test which allows for unequal variance is not as limited in this regard.*)

Comparing the ratio of geometric means using the t-test allowing for the possibility of unequal variances on log transformed data. There is a corresponding confidence interval. (*Note that we can use this test here, because age is always positive.*)

Comparing the ratio of geometric means using the t-test presuming equal variances on log transformed data. The confidence intervals that correspond to this test are only correct if the variances are truly equal under all alternatives for the mean. (*The comments about the issues surrounding the presumption of equal variances hold here as well, though it pertains to the question of equal variances on the log transformed data. Also, we again can use this test only in the situation where our data is constrained to be positive.*)

Comparing the difference or ratio of medians using large sample tests with standard errors estimated from bootstrapping. There are corresponding confidence intervals.

Testing whether the probability that a randomly chosen male would be older than a randomly chosen female using the Wilcoxon rank sum test. There is no confidence interval which corresponds to this test unless you make strong assumptions about the groups having similar shapes for their distribution.

Testing whether the proportion of males who exceed some threshold (say 70 years old) is greater than the proportion of females who exceed that same threshold. Testing this dichotomization of age could proceed under any of the tests described under part b.

Testing for any difference in the cumulative distribution functions using the Kolmogorov-Smirnov test. There is no confidence interval which corresponds to this test unless you make strong assumptions about how the shapes of the distributions might vary under the alternatives.

(Testing the ratio of hazard functions using the logrank test. There is a confidence interval appropriate for the hazard ratio under the

(rather strong) assumption that the distributions would exhibit "proportional hazard" across the sexes. (Usually this analysis is used only with censored data, however, it works just fine with uncensored data. The definition of the hazard function is a little more easily understood in the setting of measuring time to an event: The hazard is the rate with which an event will occur in the next instant conditional upon not having had an event yet. That is, among survivors up to now, what is the rate of death in the next instant. When using this with variable age, the hazard function has a far better mathematical than real-world definition.))

(Testing the mean difference between the age of a randomly chosen male and a randomly chosen female using the paired t test. There is a corresponding confidence interval. (Using this test in this setting of unmatched data is HIGHLY unorthodox. The major problems you would face are determining how to pair up the men and women in order to compute the differences, and which observations to discard if the sample sizes for the two groups were not equal. However, if you were to just randomly match up men and women in pairs, this test would give you accurate inference in large samples. I note that when using this test in large independent samples, you would tend to get the same test as the t-test which allows for unequal variances. Furthermore, because the difference of means is the same as the mean of the difference, the scientific interpretation is exactly the same for this test as for the t-test which allows unequal variances. Hence, I see no reason to ever use the paired t-test on unmatched groups.))

(Testing that the median difference between the age of a randomly chosen male and a randomly chosen female is zero using the sign test. The corresponding confidence interval is truly looking at the probability that the difference in age between a randomly chosen woman and a randomly chosen man would be below zero, rather than producing a confidence interval for the median difference. (Using this test in this setting of unmatched data is HIGHLY unorthodox. The major problems you would face is determining how to pair up the men and women in order to compute the differences, and which observations to discard if the sample sizes for the two groups were not equal. However, if you were to just randomly match up men and women in pairs, this test would give you accurate inference in large samples. Unlike with the paired t-test, however, this test does not correspond to any of the above tests for independent samples. In particular, the median difference between individuals is not the same as the difference in medians between the groups. To see this, consider the hypothetical data in which the men were {71, 73, 77} and the women were {72, 79, 77}. The median age is 73 for men and 77 for women for a difference of medians of 4. But if we randomly match up pairs of men and women, taking all possibilities we have a distribution of differences {-5, -2, -1, 0, 1, 4, 6, 6, 8}, which corresponds to a median difference of 1. Thus, if you really want to know whether the median difference is 0, we really have no test defined beyond the sign test. A better approach would be to bootstrap, however.))

(Testing for a difference between the distributions using the Wilcoxon signed rank test. I know of no easy description of the aspect of the distribution that this test is comparing. Thus, there would not seem to be a confidence interval without making strong assumptions about how alternatives would affect the shape of the

distribution. (Using this test in this setting of unmatched data is *HIGHLY* unorthodox. The major problems you would face is determining how to pair up the men and women in order to compute the differences, and which observations to discard if the sample sizes for the two groups were not equal. However, if you were to just randomly match up men and women in pairs, this test would give you accurate inference in large samples. I know of no corresponding test that we routinely use in unmatched samples. Or rather, I don't know that any test that we routinely use in unmatched samples corresponds to this unorthodox use of the Wilcoxon signed rank test.)

b. Does the distribution of infarcts on MRI exam differ by sex?

Ans: We are comparing the distribution of a binary random variable across two groups. Possibilities include:

Comparing the difference of the proportion having infarcts using the Z-test for binomial proportions. There is a corresponding confidence interval. (This is the same test as the chi squared test. I did not give credit for both answers if you gave this answer and an answer saying you would compare proportions using the chi squared test.)

Comparing the ratio of the odds of having infarcts using the chi squared test. There is a corresponding confidence interval. (This is the same test as Z test for binomial proportions, but the confidence intervals are for different population parameters. Hence, I give credit for both answers if you gave this answer and an answer saying you would compare proportions using the chi squared test.)

Comparing the difference of the proportion having infarcts using Fisher's exact test. There is no corresponding confidence interval.

Comparing the ratio of the odds of having infarcts using Fisher's exact test. There is a corresponding confidence interval.

Modifications of any of the above tests which adjust for the discrete nature of the outcome and take the worst case value for the nuisance parameter of the underlying proportion when the distributions are equal. Confidence intervals are possible if create modified tests for nonzero null hypotheses as well.

c. Does the distribution of marital status differ by sex?

Ans: We are comparing the distribution of an unordered categorical variable across two groups. Possibilities include:

Comparing the frequency distribution for the two groups using the chi squared test for independence. No general confidence interval is possible.

Dichotomizing the data (e.g., married versus not married) and using any of the tests described in part b.

d. Does the distribution of self-reported health status differ by sex?

Ans: We are comparing the distribution of an ordered categorical variable across two groups. Possibilities include:

Comparing the difference in means across the two groups using the t-test which allows unequal variances. A corresponding confidence interval can be computed, but it is not easily interpreted because the variable itself is not quantitative. We do not really know how to interpret a mean by itself, but we do know that a difference in the

means shows a difference in the distribution, and the direction of the difference can be used as a description of a tendency toward higher or lower values. (We could of course use the *t*-test which presumes equal variances, but I would be even more reluctant to do so in this setting of a finite number of categories. At some level of shift to higher values, everyone would be in the same category and have zero variance.)

(Use the Wilcoxon rank sum test. (This would be a little unorthodox only because the way we compute the standard error for the sum of the ranks does not behave real well in the presence of a lot of ties in the data. If we statisticians would clean up the way we compute standard errors, this would be perfectly okay to use.))

Comparing the frequency distribution for the two groups using the chi squared test for independence. No general confidence interval is possible. (This test makes no use of the ordering of the categories, and thus would not be particularly powerful when trying to detect tendencies to higher or lower values in one group.)

Dichotomizing the data as above or below some level and using any of the tests described in part b.

(Using any of the tests describe in part a. (As a general rule, those tests listed in part a but not explicitly listed here are even more unorthodox in this setting.))

e. Does the distribution of DSST scores differ by sex?

Ans: We are comparing the distribution of a quantitative, continuous random variable across two groups. Possibilities include all of the possibilities listed in part a.

f. Does the survival time differ by sex?

Ans: We are comparing the distribution of a censored quantitative, continuous random variable across two groups. Possibilities include:

Comparing the proportion surviving past some threshold (say 3 years) using Kaplan-Meier survival estimates and corresponding standard errors. Confidence intervals are possible.

Comparing the hazard ratio between the two groups using the logrank test. Confidence intervals are possible under the (relatively strong) assumption of proportional hazards.

Using the modified Wilcoxon rank sum test appropriate for censored variables. No confidence intervals have been described.

Using a modification of the Kolmogorov-Smirnov test appropriate for censored variables. No confidence intervals have been described.

3. Suppose the average DSST is computed separately for males and females in the combined sample and also in strata according to the presence of infarcts. The following table contains the sample mean, standard error, and 95% confidence intervals for those means.

	Mean	Std Err	95% CI Low	95% CI Hi
Females	58.30	6.67	42.26	74.34
Males	66.97	4.79	55.74	78.20
No Infarcts				
Females	85.40	9.08	61.01	109.79
Males	83.48	3.86	74.19	92.77
Infarcts				
Females	37.46	3.65	28.12	46.80
Males	35.33	3.30	26.77	43.89

- a) Based on the results in the above table, how would you characterize the evidence of an association between DSST scores and sex in the combined sample?

Ans: Males were observed to have a higher average DSST (66.97) than females (58.30), though we do not seem to have enough information to state that this represents a true difference in the population, because there is large overlap in the CI with the estimate for the males' average DSST included in the CI for the females' average. (When we have overlap to the degree that the CI for one group includes the estimate for the other group, we can be confident that the difference is not statistically significant. Less overlap than that would require us to explicitly test for a difference. That could be done here by noting that the standard error for the difference would be the square root of the sum of the squared standard errors for the two groups:

$$\sqrt{6.67^2 + 4.79^2} = 8.21 \Rightarrow (66.97 - 58.30)/8.21 = 1.06 < 1.96$$

Dividing the difference of the means by its standard error results in 1.06, which is less than 2, and thus not significant at a two-sided 0.05 level.)

(A key point to note in this answer is that I provide a scientific characterization of the association, as well as a statistical statement about my confidence in what these results tell me about the population at large.)

- b) An analysis was performed to compare the prevalence of infarcts between males and females. The difference in the prevalence of infarcts in females minus the prevalence in males was 0.22 with a 95% confidence interval of -0.036 to 0.481. Based on this information, would you conclude that there was sufficient evidence to claim an association between sex and prevalence of infarcts in the population?

Ans: Since the confidence interval contains 0, we can with 95% confidence state that there is not sufficient evidence in this sample to rule out the possibility that there is no true difference between the sexes with respect to the prevalence of infarcts in the population. (I do note that this CI is extremely wide. We have not ruled out that the difference in prevalence might be as much as 48%. That is a huge difference scientifically.)

- c) Based on the results in the above table, would you regard that there was a statistically significant association between presence of infarcts and average DSST scores in the population? Explain your reasoning.

Ans: The table did not present the exact analysis that might have been best in this situation, which would have been an analysis with both sexes combined. However, for each sex there is a huge difference in the average DSST score across the groups defined by infarct status, with no overlap between the CI. On the other hand, within the groups defined by infarct status the differences between the sexes are more modest. Thus I feel confident that the sex stratum specific results (which are highly statistically significant) would be an accurate representation of what an unadjusted analysis might show. So, yes, there does seem to be sufficient evidence to state with 95% confidence that there is an association between infarcts and DSST in the population.

- d) Based on the results in the above table, is there effect modification by infarct status on any association between DSST scores and sex? Explain your reasoning.

Ans: The difference between average DSST for females minus males is 1.92 in subjects with infarcts and 2.13 in subjects without infarcts. This does not look like a difference in association between the strata, so I would conclude that there is not much evidence to suggest effect modification.

- e) Based on the results in the above table, is there evidence of confounding by infarct status on the association between DSST scores and sex? Explain your reasoning. How does your answer relate to the findings of part b and c above?

Ans: When the analysis did not consider infarct status, the analysis suggested men did better on DSST than women by about 8 points on average. When analyses were performed within strata defined by infarct status, the analysis in each stratum suggested women did better on DSST than men by about 2 points. Given this reversal of effect between the unadjusted and adjusted analyses, we can state that there does appear to be confounding. *(In fact, because we were using difference in means as our measure of association, any substantial difference in the estimates from the adjusted and unadjusted analyses is sufficient to infer confounding in the absence of effect modification. That is, we do not have to have that the direction of association is reversed. Had we been using the odds ratio, we would have to have had a reversal of association to be certain that confounding was the cause of a difference in the unadjusted and adjusted analyses.)*

- f) What analysis would you regard as the most appropriate way to assess an association between sex and DSST scores in this sample?

Ans: I would prefer an adjusted analysis which looked at a weighted average of the stratum specific estimates of association.

4. Suppose that the distribution of age was compared between the sexes using a t test for equal variances (yielding a P value of 0.04) as well as a t test allowing unequal variances (yielding a P value of 0.09). Give two possible interpretations for the seeming discrepancy between these results.

Ans: The t test which presumes equal variances can be anti-conservative when testing means in a setting in which the variances

are actually unequal. Hence, the statistical significance (living and dying by .05) of the t-test for equal variances may really be reflective of unequal variances (and perhaps equal means).

The t test which allows equal variances is not as efficient a test when the variances are truly equal. Hence, if the variances were truly equal, we may just be seeing a higher P value because we are using a more conservative calculation of the degrees of freedom in the t test which allows unequal variances.

5. Below are 7 scatterplots. List the plots in order according to lowest (most negative) to highest (most positive) correlation. (In all cases, the scale for the x and y axes are the same.)

Ans: From lowest (most negative correlation) to highest (most positive) correlation, the answer is

$$E < B < D < F=G < A < C$$

(The key things we look for are the slope of the line, the variance of the predictor X, and the variance of the response Y within groups that have the same value of X. Thus we can determine the above ordering by considering pairs of plots:

E and B have about the same (negative) slope and the same variance of X, but E has less variation of Y within groups defined by X. Thus the correlation in plot E is further from 0. As the slope is negative, that means that the correlation in plot E is more negative than the correlation in plot B.

B and D have about the same (negative) slope and the same variation of Y within groups defined by X, but plot D has sampled X over a lower range, thus leading to a lower variance of the predictor X. Thus plot D must have a correlation closer to 0 than plot B. Since both slopes are negative, that means that the correlation in plot B is more negative than the correlation in plot D.

The best fitting straight line to the data in plots F and G would be a flat line (zero slope) in each case. Thus the correlation would be close to zero. Thus the correlation in plots F and G are approximately equal to each other, above all the plots that had negative slopes (E, B, D), and below all the plots that had positive slope (A, C).

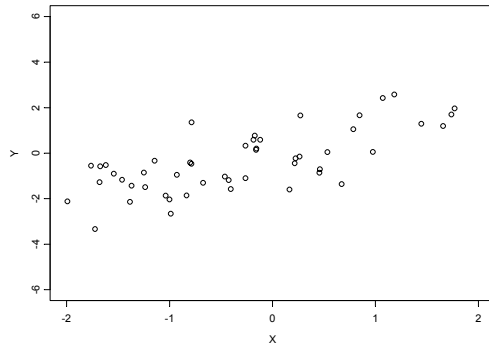
A and C have about the same variation of Y within groups defined by X, and the predictor X has about the same distribution (and hence variance) in each plot. But plot C has a steeper (more positive slope). Thus the correlation in plot C will be more positive than the correlation in plot A.)

(Plots on next page)

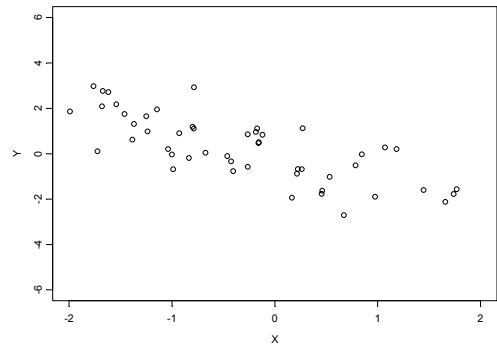
Distribution of grades: Max possible 100; Highest achieved 100

Percent:	<u>10</u>	<u>20</u>	<u>30</u>	<u>40</u>	<u>50</u>	<u>60</u>	<u>70</u>	<u>80</u>	<u>90</u>
Percentile:	66	71		77	80	82		86	91

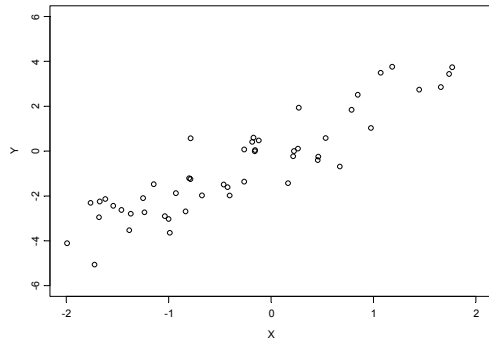
Plot A



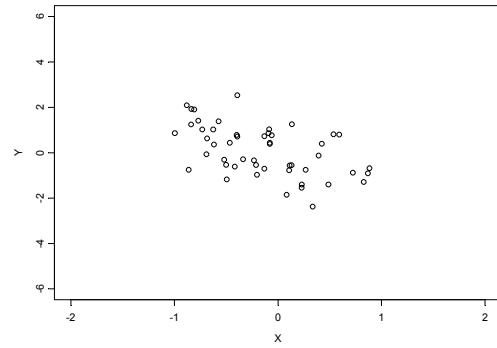
Plot B



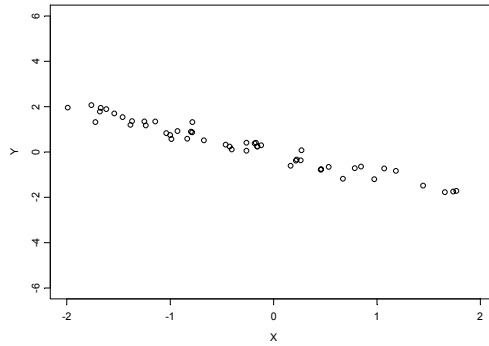
Plot C



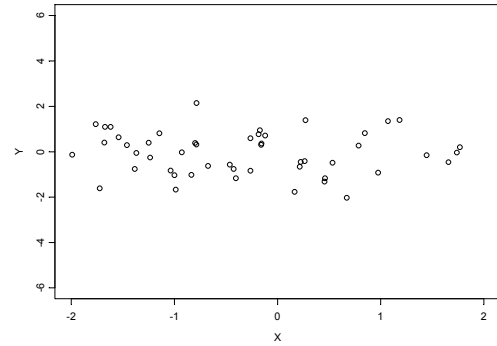
Plot D



Plot E



Plot F



Plot G

