Biost 514
Emerson, Fall 00
December 12, 2000

Biost 514
Biostatistics I

Final Examination Key

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.

You may reference any statistical output you used to analyze the DFMO data, but otherwise 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.

In answering the following questions, you may use the following table of critical values for the standard normal and t distributions. (Note: Should you desire critical values for some other distribution, you may feel free to use an appropriate approximation derived from this table, so long as you state why that approximation might be appropriate.)

			Quantiles			
Distribution	.80	.90	.95	.975	.99	.995
t distribution						
df= 20	0.860	1.325	1.725	2.086	2.528	2.845
df= 22	0.858	1.321	1.717	2.074	2.508	2.819
df= 24	0.857	1.318	1.711	2.064	2.492	2.797
df= 26	0.856	1.315	1.706	2.056	2.479	2.779
df= 28	0.855	1.313	1.701	2.048	2.467	2.763
df= 30	0.854	1.310	1.697	2.042	2.457	2.750
df= 50	0.849	1.299	1.676	2.009	2.403	2.678
df= 52	0.849	1.298	1.675	2.007	2.400	2.674
df= 54	0.848	1.297	1.674	2.005	2.397	2.670
Standard Normal	0.842	1.282	1.645	1.960	2.326	2.576

Problems 1 - 5 refer to the phase IIb study of DFMO that we have analyzed in class. Appendix A presents a brief description of the data relevant to this exam, and Appendices B and C present selected results of analyses of these data.

1. Suppose we wish to assess how the average spermidine level changed over 12 months of treatment with dose 0.4 of DFMO by comparing the baseline and 12 month measures in that dose group.

   a. (12 points) Briefly describe four different hypothesis tests you might use to detect whether the distribution of spermidine measurements tended to change over time in the highest dose group. In each case, provide a five word or less description of how the standard error of the test statistic would be computed.

Ans: Many tests are possible. Below are a few that I might seriously consider.

- **paired t-test using** $spd12$ **and** $spd0$ **(equivalent to one sample t test on** $spddiff = spd12 - spd0$**).**
- **sign test**
- **signed rank test**
- **one sample t test on** $spdratio = spd12/spd0$ **(null hypothesis is that the average ratio is 1).**
- **one sample t test on log transformed** $spdratio$ **(null hypothesis is that the average log ratio is 0).**
- **permutation test on the difference or ratio of** $spd12$ **and** $spd0$**, where the values of** $spd12$ **and** $spd0$ **are randomly permuted within individuals.**

   b. (8 points) Using the results presented in Appendix B, provide an analysis that addresses this question. State the conclusion you would reach giving both a statistical and scientific interpretation.

Ans: I will use the confidence interval that corresponds to the paired t-test, as that is the only analysis that I can perform using the results presented in Appendix B. Hence, my 95% confidence interval for the change in spermidine level is

$$-1.756 \pm 2.092 \times \frac{2.171}{\sqrt{20}} = (-2.77, -0.74)$$

where the critical value $t_{19,.975} \doteq 2.092$ **was obtained by extrapolation from the table on the first page of the exam (the true value is 2.093). Thus the observed data is typical of what might be expected if 12 months of treatment with dose 0.4 of DFMO truly caused a decrease in the average spermidine level anywhere between 0.74 and 2.77. Because 0 is not in that confidence interval, I conclude that the average spermidine levels truly dropped after 12 months of treatment. (However, because this analysis does not take into account possible time trends in the spermidine levels in the placebo group, I am not yet confident enough to ascribe such a change as being due to DFMO rather than some trend with, say, age.)**

2. Suppose we wish to compare the dose 0 and dose 0.4 treatment groups with respect to average spermidine level after 12 months on study.

   a. (15 points) Briefly describe five different hypothesis tests you might use to detect whether DFMO tends to result in a lower spermidine level when compared to no intervention. In each case, provide a five word or less description of how the standard error of the test statistic would be computed.

Ans: There are of course a great many tests that are possible. Below I list some of the ones that I might consider among those we covered this quarter.

- **unequal variance t-test comparing mean** $spd12$ **across groups computing standard error from formula for standard errors of the mean.**
- **equal variance t-test comparing mean** $spd12$ **across groups computing standard**

error from formula for standard errors of the mean.

- **Wilcoxon test of $spd12$ across groups with standard error computed from permutation distribution.**
- **compare median $spd12$ across groups with standard error from bootstrapping.**
- **compare mean $spd12$ across groups with standard error from permutation test (Theory shows that this is nearly equivalent to the t-test).**
- **log transform $spd12$ and then compare means of the log transformed variables using t-test with equal or unequal variances. (Of course, the Wilcoxon test and tests on medians would be the exact same whether the data were log transformed or not.)**
- **compute $spd12low$ as an indicator that $spd12 < 2$ (I pick the number 2 arbitrarily– I have no scientific knowledge to base my choice on), and then perform a chi-square test (test for binomial proportions) comparing the treatment groups with respect to this binary variable (standard error comes from mean-variance relationship).**
- **compute $spddiff = spd12 - spd0$ for each subject, and then use any of the first five tests given above. (We would not want to use a log transform of $spddiff$ because of negative values.) (This is not really the best way to account for the initial BP measurement. A better approach is to adjust for the initial value as a covariate in a multiple linear regression.)**
- **chi-square test on a variable indicating whether $spddiff$ is positive (standard error is computed from the mean-variance relationship of the binomial)**
- **compute $spdratio = spd12/spd0$ and then use any of the first six tests to compare $BPratio$ across groups.**

b. (10 points) Using the results presented in Appendix B, provide an analysis that best addresses this question. State the conclusion you would reach giving both a statistical and scientific interpretation.

Ans: **My top choice would be to base a confidence interval on the results of the regression analysis which adjusted for baseline values:**

$$-0.6786 \pm 2.017 \times 0.1657 = (-1.01, -0.34)$$

where the critical value $t_{45,.975} \doteq 2.017$ was obtained by interpolation from the table on the first page of the exam (the true value is 2.014). Thus an individual treated with dose 0.4 of DFMO for 12 months is estimated to have on average a spermidine level 0.6786 lower than a placebo treated individual who had the same value at baseline. The data we observed is what would might be typically observed when the true difference between the two groups in average spermidine levels is somewhere between -1.01 and -0.34. Such results are statistically significant: P= .0002 from the regression analysis.

An alternative analysis (and the second best choice from what was possible with the given data) would be the confidence interval that corresponds to the two sample t-test for unequal variances:

$$1.950 - 3.256 \pm 2.092 \times \sqrt{\frac{0.799^2}{20} + \frac{1.314^2}{28}} = (-1.95, -0.67)$$

where I used the critical value $t_{19,.975}$ for conservatism (the actual degrees of freedom using typical approximations would have been somewhere between 19 and 27). Thus the observed data is typical of what might be expected if 12 months of treatment with dose 0.4 of DFMO truly caused the average spermidine level to be anywhere between 0.67 and 1.95 lower than that in a placebo group. Because 0 is not in that confidence interval, I conclude that the average spermidine levels truly dropped after 12 months of treatment.

Yet another analysis could compare the high dose and placebo group with respect to the change in spermidine over 12 months. This would be effected by using the data for *spddiff* **in the confidence interval that corresponds to the two sample t-test for unequal variances:**

$$-1.756 + 0.041 \pm 2.092 \times \sqrt{\frac{2.171^2}{20} + \frac{1.533^2}{28}} = (-2.90, -0.53)$$

where I again used the critical value $t_{19,.975}$ **for conservatism. Thus the observed data is typical of what might be expected if 12 months of treatment with dose 0.4 of DFMO truly caused the average change in spermidine level to be anywhere between 0.53 and 2.90 lower than that in a placebo group. Because 0 is not in that confidence interval, I conclude that the average spermidine levels were truly lowered by 12 months of treatment. (You can see that this confidence interval is wider than the first two. As noted in the key to homework 7, this is because there is not a high correlation between successive spermidine measurements in the same individual.)**

3. Of the answers you provided in problems 1 and 2, which do you prefer to answer the scientific question of whether DFMO affects spermidine levels? Why?

Ans: (10 points) The approach in problem 2 is best (and the first of those solutions is to be preferred), because it allows for possible changes in spermidine levels that might be due to the experimental setting or due to such biological phenomena as age (all subjects are a year older), seasonal changes in diet (though these are usually presumed to be on a yearly cycle), or secular trends in the use of other cancer prevention strategies.

The types of analyses presented in problem 1 are <u>NOT</u> the appropriate way to analyze data from a randomized study. Instead, we look to comparisons across randomized treatment groups.

4. Given in Appendix C are descriptive statistics of putrescine levels after 6 months of treatment.

a. (5 points) What is a 95% confidence interval for the average putrescine level after 6 months of treatment with placebo (dose 0)?

Ans: A 95% confidence interval is computed as

$$1.055 \pm 2.045 \times \frac{1.593}{\sqrt{30}} = (0.46, 1.65)$$

where the critical value $t_{29,.975} \doteq 2.045$ **was obtained by extrapolation from the table on the first page of the exam (the true value is 2.045).**

b. (5 points) What is a 95% confidence interval for the average putrescine level after 6 months of treatment with DFMO at dose 0.4?

Ans: A 95% confidence interval is computed as

$$0.333 \pm 2.064 \times \frac{0.432}{\sqrt{25}} = (0.15, 0.51)$$

where the critical value $t_{24,.975} = 2.064$ **was obtained from the table on the first page of the exam.**

c. (5 points) Based on your answers to parts (a) and (b), what can you conclude about the statistical significance of the observed difference in average putrescine levels for the dose 0 and dose 0.4 groups after 6 months of treatment? Based on those results, do you have evidence that the average putrescine level is truly different? Explain.

Ans: The two intervals overlap, and we might at first think that such overlap means that we cannot reject the null hypothesis of equivalence between the two groups with respect

to the average putrescine value. As seen below, however, that is not a hard and fast rule. It is true that if the two confidence intervals do not overlap, we can presume statistically significant differences, however the converse is not true. When confidence intervals for two means overlap, it is still sometimes the case that the difference in the means is statistically significantly different from 0.

d. (5 points) What is a 95% confidence interval for the difference between the dose 0 and dose 0.4 groups with respect to average putrescine levels after 6 months of treatment?

Ans: A 95% confidence interval is computed as

$$0.333 - 1.055 \pm 2.064 \times \sqrt{\frac{0.432^2}{25} + \frac{1.593^2}{30}} = (-1.35, -.10)$$

where the critical value $t_{24,.975} = 2.064$ was used for conservatism as in problem 2.

e. (5 points) Based on your answers to part (d), what can you conclude about the statistical significance of the observed difference in average putrescine levels for the dose 0 and dose 0.4 groups after 6 months of treatment?

Ans: Because 0 is not in the confidence interval, I would conclude that the treatment with DFMO results in a lower average putrescine level.

f. (5 points) What would you conclude about the statistical significance of the observed difference in average putrescine levels for the dose 0 and dose 0.4 groups after 6 months of treatment? That is, which is the preferred way to answer such a question: using the answers to parts (a) and (b), or the answer to part (d)? Why?

Ans: The analysis in part (d) directly compares the two groups and accounts for the variability in the estimation of the difference in means. It is thus to be preferred to the idea of constructing two separate confidence intervals.

g. (5 points) Using the results presented in Appendix B, can you compute a confidence interval for the change in average putrescine level after 6 months of treatment in the 0.4 dose group? If so, provide the confidence interval. If not, why not?

Ans: The data measured over time in the same dose group is correlated. We thus cannot compute the standard error for the difference in means from the standard deviations of the measurements at each time. We would need to account for the matched data, and no information about the correlation between successive measurements in an individual is provided.

5. (10 points) When we perform a t-test comparing the dose 0 and dose 0.4 groups with respect to putrescine levels after 12 months of treatment, we obtain a P value of 0.4436, and when we perform a Wilcoxon rank sum test on that same data, we obtain a P value of 0.0255 (analyses not shown). Provide two distinct possible explanations for such seemingly contradictory results.

   - **For distributions having heavy tails, the Wilcoxon test is more efficient than the t-test, so there will be a tendency for the Wilcoxon test to be more statistically significant than the t-test in situations where the shape of the distribution is the same in each group and that shape has heavy tails.**
   - **The Wilcoxon test is testing whether a randomly chosen value from one group has a probability greater than .5 of being larger than a randomly chosen value from the other group. It is possible that this is true for the population, but that the average measurement (which is what is being tested by the t-test) is the same for both groups.**

Problems 6 - 8 consider a hypothetical study of the effect of transcendental meditation (TM) on high blood pressure and its sequelae. Subjects with high blood pressure (systolic blood pressure greater than 160 mm Hg) were randomized to TM or a no intervention group. Study entry occurred between January 1, 1990 and December 31, 1995, with data analysis taking place on December 31, 1998. No patients were lost to follow-up. Data is available on the following variables

AGE = subject's age in years at study entry

$OCCUP$ = occupation (1= clerical, 2= laborer, 3= professional)

$ENTRY$ = time of study entry measured in years since January 1, 1990

TM = indicator of treatment group (0 = no intervention, 1= TM)

$BPinit$ = initial blood pressure at study entry

$BPfinal$ = blood pressure 2 years after randomization

$ObsTime$ = observation time from study entry to stroke or end of study, whichever comes first

$Stroke$ = indicator that subject experienced a stroke on study (0= no stroke, 1= stroke)

6. (20 points) For each of the above variables, tell what descriptive statistics you might provide to describe the sample.

AGE **Mean, median, standard deviation, interquartile range all make sense for this continuous variable.**

$OCCUP$ **Frequency table and mode is all that makes sense for this categorical variable.**

$ENTRY$ **Mean, median, standard deviation, interquartile range all make sense for this continuous variable. (This variable is probably not of too much scientific interest, though it might be important when considering time trends in the data.)**

TM **Frequency table or proportion treated with TM is all that makes sense for this binary variable (I guess we could also report odds, but not many people would really find that useful).**

$BPinit$ **Mean, median, standard deviation, interquartile range all make sense for this continuous variable.**

$BPfinal$ **Mean, median, standard deviation, interquartile range all make sense for this continuous variable.**

$ObsTime$ **Kaplan-Meier estimates of probability of time to stroke using both this variable and *Stroke* (No descriptive statistics are particularly useful for either this variable or *Stroke* when considered alone).**

7. Suppose a new variable $NORMAL$ is created such that $NORMAL = 1$ if $BPfinal \leq 120$ and $NORMAL = 0$ otherwise (so $NORMAL$ is an indicator that the subject's blood pressure has dropped to the normal range). The following table presents the odds ratio comparing TM group to the no intervention group, and a confidence interval for that odds ratio.

	Odds	95% Conf
Occupation	Ratio	Interval
All Combined	1.96	1.56, 2.76
Clerical	1.08	0.52, 2.23
Laborer	1.29	0.77, 2.16
Professional	1.46	0.67, 3.06
Stratified	1.30	0.90, 1.86

a. (5 points) When the data are analyzed with all occupations combined, is there evidence of a statistically significant association between treatment and a decrease in blood pressure to the normal range? State your reasoning.

Ans: Because the 95% confidence interval does not include the null hypothesis of an odds ratio of 1, there is a statistically significant association at the .05 level.

b. (5 points) Is there strong evidence of an interaction between treatment and occupation on the effect of transcendental mediation on blood pressure? State your reasoning.

Ans: To judge the interaction, we consider whether the odds ratios within strata all look to be about the same. While they range from 1.1 to 1.5, the width of the confidence intervals within the strata suggests that we do not have strong evidence that the odds ratios are truly different across strata: There is very large overlap across the stratum specific confidence intervals.

c. (5 points) Is there strong evidence that the analysis of the association between treatment and blood pressure is confounded by occupation? State your reasoning.

Ans: To judge confounding, we consider whether the odds ratios within strata all look to be about the same as the odds ratio when all strata are combined. They will not in general be exactly the same, but in the absence of confounding the odds ratio for the combined data might be expected to look like an average of the stratum specific odds ratios. (A technical note that we did not cover in class in very much detail: in binary data, a strongly predictive stratification variable might cause the combined odds ratio to be attenuated toward the null hypothesis compared to the stratified odds ratio, though it will not generally affect the statistical significance of the analyses very much.) In this data, the combined data odds ratio is more extreme than any of the stratum specific odds ratios, and thus we can confidently state that there is confounding.

d. (5 points) When the analysis is adjusted for occupation, is there evidence of a statistically significant association between treatment and a decrease in blood pressure to the normal range?

Ans: Because the 95% confidence interval does include the null hypothesis of an odds ratio of 1, there is not a statistically significant association at the .05 level.

e. (5 points) Suppose 50% of the subjects in the no intervention group had a final blood pressure in the normal range. What was the odds of a normal blood pressure in that group? Using the estimated odds ratio for the combined sample, what is the odds of a normal final blood pressure in the TM group? What is the proportion of individuals in the TM group that had a normal final blood pressure?

Ans: Letting p be the probability of an event, the odds is $p/(1-p)$. Thus if $p = .5$ in the no intervention group, the odds is $o = .5/.5 = 1$. Now if the odds ratio is $OR = 1.96$, then the odds in the TM group is $o = 1.96$, and inverting the formula, we find $p = o/(1+o) = 1.96/2.96 = 0.662$.

f. (5 points) Suppose 90% of the subjects in the no intervention group had a final blood pressure in the normal range. What was the odds of a normal blood pressure in that group? Using the estimated odds ratio for the combined sample, what is the odds of a normal final blood pressure in the TM group? What is the proportion of individuals in the TM group that had a normal final blood pressure?

Ans: If $p = .9$ in the no intervention group, the odds is $o = .9/.1 = 9$. Now if the odds ratio is $OR = 1.96$, then the odds in the TM group is $o = 9 \times 1.96 = 17.64$, and inverting the formula, we find $p = o/(1+o) = 17.64/18.64 = 0.946$.

8. Consider the following analyses of the incidence of strokes.

A. the treatment groups are compared with respect to *ObsTime* using a t test.

B. the treatment groups are compared with respect to *Stroke* using a chi square test.

C. the data set is restricted to those subjects who entered the study prior to December 31, 1993, and a new variable is created indicating whether a stroke occurred within 5 years of study entry (the new variable *Stroke*5 would be 1 if *Stroke* $= 1$ and *ObsTime* < 5, and it would be 0 otherwise). The treatment groups are then compared with respect to *Stroke*5 using a chi square test.

a. (10 points) Very briefly state why each of the above analyses is not the best one to use.

A. *ObsTime* **is censored in that it is sometimes measuring time to data analysis rather than time to stroke. In general, such a test is testing equality of both the censoring distribution and the time to the event. It is true that if the censoring time distribution is equal in the two groups (such as might be the case in a randomized clinical trial with censoring due only to continued stroke-free status at time of analysis) this analysis will not give a biased result, the interpretation of, say, the average** *ObsTime* **has no scientific relevance. Such a test would also not be using the data most efficiently.**

B. The *Stroke* **variable is measured over varying time frames depending on the censoring distributions. Again, if the censoring distributions were known to be equal, this would not be incorrect, just inefficient.**

C. Because everyone was followed for at least five years, this would be a valid test. However, it would still be inefficient, because it does not use all the data.

b. (5 points) What is a more appropriate analysis to use?

Ans: The logrank test or some other test appropriate for censored data would be better. (Note that one such possibility could be to use the difference in the Kaplan-Meier estimates for the survivor probability to some arbitrary time point. The standard error for the difference can be computed using the estimated standard errors of the survivor estimates for the independent groups.)

Grade distribution: (165 Possible)

Mean (Std Dev):	**121 (23.7)**
Median:	**124**
Highest:	**154**
IQ range:	**106, 139**