Biost 517
Applied Biostatistics I

Midterm Examination Key
November 14, 2002

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

1. Consider a hypothetical study of health care utilization in which the billing records for 500 randomly selected members of a health maintenance organization (HMO) were examined for the period Jul 1, 2000 to Jun 30, 2001. Of particular interest was the usage of radiologic imaging resources for different patient demographics, presenting complaint, etc. For each of the selected subjects, the following variables were measured on each patient visit.
   - *ptidno*= patient identification number uniquely identifying each patient
   - *date*= date of visit in MMDDYY format
   - *age*= age of patient in years
   - *male*= sex of patient in coded format (0= female, 1= male)
   - *clinic*= code for clinic visited (0= family practice, 1= pediatrics, 2= medicine, 3= surgery, 4= obstetrics, 5= emergency room)
   - *temp*= patient's temperature in degrees Celsius at that visit
   - *sbp*= patient's systolic blood pressure in mm Hg at that visit
   - *radimage*= indicator that radiologic imaging was used at that visit (0= no, 1= yes)

   The following table presents descriptive statistics for the dataset.

	n	msng	mean	std dev	min	25%-ile	median	75%-ile	maximum
ptid	702	0	955.4	564.4	11.0	444.3	952.5	1421.5	2000.0
date	702	0	67893	34434	10101	40426	71151	100400	122900
age	702	0	53.25	29.87	0.03	31.90	60.32	77.14	99.61
male	702	0	0.47	0.50	0	0	0	1	1
clinic	702	0	2.26	1.76	0	1	2	4	5
temp	702	58	37.29	1.02	36.00	36.63	37.11	37.63	40.97
sbp	702	71	124.59	24.70	90.04	105.09	121.90	135.68	198.37
radimag	702	0	0.32	0.46	0	0	0	1	1

a. For each of the variables given above, indicate the descriptive statistics that are not of scientific use to answer any scientific question.

Ans: For the nominal variable of ptid *and* clinic *neither the mean, standard deviation, nor any of the quantiles are of scientific value. The variable* date *does represent a quantitative variable (measured on an interval scale), but the coding of the variable prevents the usual descriptive statistics from being of much use (note 12/29/00 is considered a maximum, while 01/01/01 is considered a minimum). For the continuous variables* age, temp, *and* sbp, *all of the descriptive statistics are informative. For the binary variables* male *and* radimag, *the mean tells us the proportion of male-visits and visits with radiologic imaging, but the standard deviation and quantiles do not present additional information beyond that (and hence are pretty boring).*

(Note that the missing data in this data set is quite unlikely to be censored data. The term "censored data" applies only to a special type of missing data in which the exact measurement is unknown, but that it is known that the exact measurement occurs in some limits. The variable temp, *for instance, would not be a "censored" variable if the missing data were missing just because the measurement was never made. It would be "censored" if the missing data occurred because 40 degrees Celsius was the maximum temperature that could be recorded by the thermometers in use, and all subjects with missing data were missing because their temperature was greater than 40 degrees Celsius.)*

b. How would you use the above statistics to estimate the average age of patients enrolled in the HMO? Briefly explain the issues you need to consider.

Ans: I neither could nor would. The issue is that we randomly chose 500 subjects, and then collected data on the visits that those 500 subjects had. We might imagine that there could have been subjects with no visits. If there were, it is clear that any subject who had no visits was apparently not in the data (as could be determined from the fact that every case in the data had a nonmissing clinic value). Furthermore, some subjects are apparently represented more than once in the data set (as could be determined from the fact that there were 702 cases in the data from at most 500 subjects). We would need a different data set to estimate the average age, because it is highly likely that number of visits per year is associated with age.

(I note that the type of missing data mechanism present in this dataset for the number of visits is called a "truncated distribution". We know how many visits people had, providing they had 1 or more visit in a year. We do not know how many of the selected 500 people had zero visits, though we could of course figure that out by counting the number of distinct values for variable ptid. *It does sometimes happen that we get a data set in which we never get to see how many subjects had zero visits. And, of course, merely knowing how many subjects were omitted from our data does not help us estimate their ages.)*

c. How would you use the above statistics to estimate the percentage of patients at the HMO who had radiologic imaging during 2001? Briefly explain.

Ans: Same answer as above.

(But here we could use the data set (but not the presented descriptive statistics) to nearly answer the question: If I knew how many distinct values of ptid *were associated with at least one visit which involved radiologic imaging, then I could divide that number by 500 to produce an estimate of the percentage of patients imaged during the 12 months studied. Of course, the 12 months studied only included 6 months of 2001, so we would have to decide whether the 2000-2001 data would be enough. I did not expect you to consider all these issues, only the issue regarding the over- and under-representation of patients in the descriptive statistics.)*

d. How would you use the above statistics to estimate the percentage of patient visits to the HMO that involve radiologic imaging? Briefly explain.

Ans: The mean for variable radimag *is interpretable as the proportion of patient visits for the 500 patients which involved radiologic imaging. Because the 500 patients were a random sample, it follows that the 702 patient visits are also a random sample of all patient visits. Thus I estimate 32% of patient visits involve radiologic imaging.*

2. Consider a hypothetical study of cardiovascular disease in which data was gathered on 100 subjects who were recently diagnosed with a heart attack (myocardial infarction, MI) and 500 patients of comparable age and sex who were seen at the Emergency Room for medical reasons other than cardiovascular disease. Of interest is whether marijuana use is associated with increased risk of heart attacks. Data available for the study includes the following
   - *subjid*= unique subject identification number
   - *age*= age of subject in years
   - *male*= sex of subject in coded format (0= female, 1= male)
   - *chol*= serum cholesterol value in mg/dl
   - *sbp*= patient's systolic blood pressure in mm Hg at that visit
   - *mi*= indicator that subject was diagnosed with an MI (0= no, 1= yes)
   - *marij*= indicator that subject smoked marijuana within three hours of arriving at the emergency room

   The following table presents descriptive statistics for the dataset.

	n	Msng	mean	std dev	min	25%-ile	median	75%-ile	maximum
ptid	600	0	995	573	1	507	998	1482	2000
age	600	0	60.27	6.79	38.59	55.61	60.03	64.99	80.76
male	600	0	0.68	0.47	0.00	0.00	1.00	1.00	1.00
chol	600	0	222.56	33.76	159.00	208.00	216.00	236.50	395.00
sbp	600	0	134.58	30.92	90.04	107.56	129.12	160.47	199.37
mi	600	0	0.17	0.37	0.00	0.00	0.00	0.00	1.00
marij	600	0	0.23	0.42	0.00	0.00	0.00	0.00	1.00

a. Based on the descriptive statistics presented above, might any of the above variables appear to have substantial outliers? Explain your reasoning.

Ans: The maximum cholesterol value is approximately 170 points above the mean. As the standard deviation is approximately 34, this 5 SD departure is markedly greater than the 2 SD difference between the minimum and the mean. Hence I would be a little suspicious of an outlying value, even though there is not a huge difference between the mean and the median, nor is the standard deviation particularly large for the mean. I am not too worried about either age or systolic blood pressure when using these same criteria.

(Note that with a sample size of 600, we would expect the minimum and maximum values to be greater than 2 SD from the mean even when the data were normally distributed. That is, with a normal distribution, 5% of the data (or about 30 observations with a sample size of 600) lies more than 2 SD from the mean. With other shapes of distributions it can be even more. With 600 observations from a normal distribution we might expect the minimum and maximum to be approximately 2.9 SD from the mean.)

b. From the above descriptive statistics, can you estimate the incidence of MIs in the population? If so, what is your estimate? If not, why not.

Ans: Because our sampling scheme fixed the number of MI and non-MI patients, we cannot estimate the prevalence of MIs in the population. (This was a case-control study.)

c. From the above descriptive statistics, can you estimate the prevalence of smoking marijuana in the population? If so, what is your estimate? If not, why not.

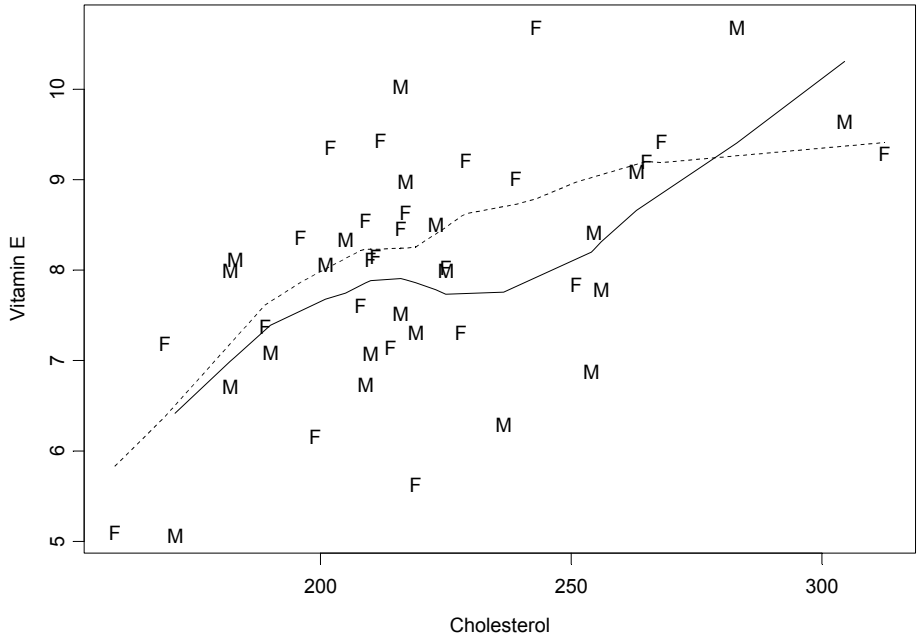
Ans: Our sampling scheme did not fix the proportion of marijuana smokers, but if there is an association between MIs and marijuana use, the sampling scheme which fixed the numbers of MI patients would lead to a biased estimate of marijuana use. And one would presume that the investigators were at least entertaining the idea that there would be such an association.

d. From this study, what statistics would you use to estimate the association between marijuana use and MIs?

Ans: Due to the sampling scheme, we would have to estimate the prevalence of marijuana use within groups defined by MI status. We could then use those estimates to compute the difference in prevalences or the ratio of prevalences between groups. We could also use the prevalences to compute the odds of marijuana use within each group and then compute the odds ratio across groups. This latter approach has the advantage that the population odds ratio comparing odds of marijuana use across MI groups is exactly equal to the population odds ratio comparing odds of MI across marijuana use groups. Furthermore, if MIs are relatively rare, the population odds ratio comparing odds of MI across marijuana use groups is approximately equal to the incidence ratio comparing incidence of MI across groups defined by marijuana use.

(We would not, however, want to base our inference on the incidence of MIs across groups defined by marijuana habits, because those quantities are not estimable from the case-control study design.)

3. Below is a scatterplot of serum cholesterol measurements and plasma vitamin E levels for 46 healthy subjects. Points are labeled according to sex, and lowess curves for each sex stratum are superimposed on the graph (broken line is females, solid line is males).



a. Is there evidence of an association between cholesterol and vitamin E levels? Explain your reasoning.

Ans: There does appear to be a clear positive slope to the linear trend in the plot of vitamin E levels versus cholesterol levels. Hence, the distribution of plasma vitamin E does seem to vary according to serum cholesterol level signifying an association between these variables.

b. Is there evidence of an association between sex and vitamin E levels after adjusting for cholesterol levels? Explain your reasoning.

Ans: The separation between the lowess curves drawn for the sex strata would be an indication of an association between sex and vitamin E level after adjusting for cholesterol. This is very much a judgement call, but I am not too impressed with there being a difference between the curves beyond what might be expected due to random chance. That is, I would suspect that if the truth were that the curves would lie exactly on top of each other, we might see graphs like this pretty often. I am, of course, allowing statistical precision enter into my decision. You could have answered this either way, providing you explained your reasoning.

c. Is there evidence that sex modifies the association between cholesterol and vitamin E levels? Explain your reasoning.

Ans: This decision should be based on whether the lowess curves are roughly parallel. To the extent that I would entertain the notion that the two curves were not coincident, I would personally think that the curves were reasonable parallel. Again, I am letting my perception of statistical precision sway my thinking. You would have gotten full credit so long as you noted that you were answering the question based on parallelicity of the curves.

4. For the data presented in problem 3, the correlation is estimated at r = 0.559. What would be the likely effect on the correlation if I were to modify the study design in the following ways?

a. Restrict eligibility for the study to subjects having cholesterol between 160 and 220.

Ans: This decreases the variance of the predictor of interest cholesterol. That would tend to decrease the correlation toward zero.

(If you thought that the slope of the curves were markedly steeper in the lower portion of the cholesterol distribution, the restriction to that lower range would tend to drive the correlation closer to 1. The net effect would in this data set is toward less of a correlation, however. The correlation is 0.37 when the range of cholesterol is restricted to the range 160 – 220.)

(In any case, changing the sample size does not have any systematic effect on correlation.)

b. Restrict eligibility for the study to subjects taking vitamin E supplements, rather than including subjects regardless of vitamin E intake.

Ans: This would tend to decrease the variability of vitamin E levels within groups defined by cholesterol level. That tends to increase the correlation toward 1.

5. The following table contains descriptive statistics for FEV (l/sec) in a population aged 65 – 100. Descriptive statistics are presented for the combined sample, as well as within strata defined by smoking history (ever versus never). Also presented are descriptive statistics defined by smoking history within each sex.

	n	msng	mean	std dev	min	25%-ile	median	75%-ile	maximum
All	735	10	2.21	0.69	0.41	1.75	2.16	2.65	4.47
All Nonsmok	322	4	2.22	0.66	0.57	1.79	2.15	2.57	4.47
All Smokers	414	8	2.20	0.71	0.41	1.70	2.16	2.69	4.21
Nonsmok Females	200	4	1.94	0.43	0.57	1.62	1.99	2.22	2.86
Smoking Females	171	4	1.77	0.45	0.57	1.52	1.79	2.07	2.93
Nonsmok Males	123	1	2.67	0.72	0.58	2.27	2.66	3.03	4.47
Smoking Males	244	5	2.49	0.71	0.41	2.08	2.55	2.95	4.21

a. Is there evidence of an association between FEV and smoking? Provide descriptive statistics in support of your answer.

Ans: *No. The mean FEV for smokers was .02 l/sec lower than that for nonsmokers. Such a difference was not so great that I would consider this evidence for an association.*

(If you said you thought .02 was sufficiently nonzero as to indicate an association, that was okay by me.)

(You could have compared medians or some other quantile. However, you only got half credit if you adjusted for sex in this question.)

b. Is there evidence that the association between FEV and smoking is confounded by sex? Provide descriptive statistics in support of your answer.

Ans: *Yes. There is strong evidence that sex is associated with FEV: Among nonsmokers, males average an FEV of 2.67 while females average 1.94. A similar difference exists between the sexes among smokers. This, of course, fits in well with our understanding about the relationship between sex and body size and body size and FEV.*

There is also strong evidence that sex is associated with smoking. 171 / 371 females smoke, while 244 / 367 males smoke.

Sex is most certainly not in any causal pathway of interest between smoking and differences in FEV, hence the above observations are sufficient for me to conclude that sex confounds the detection of an FEV – smoking association.

(A common symptom of confounding is that an unadjusted analysis differs substantially from an analysis adjusted for the confounder. In this example, we see that the difference between average FEV in nonsmokers and that in smokers is 0.18 l/sec for males and 0.17 l/sec for females. These stratified estimates of association are remarkably similar to each other, but they are quite different from the unadjusted analysis reported in part a. This would be enough to make me suspect confounding when using any summary measure as the basis for measuring association. When using the difference of means, this is actually diagnostic for confounding, though as noted in the key to Homework #5 from 2001, it would not necessarily prove confounding if we were using odds ratios as our measure of association. In any case, for this exam, I gave full credit if you pointed out the difference between the stratified and unadjusted analysis results. Be forewarned, however, that were I to ever use the odds ratio as a measure of association in such a problem, I would not have given full credit for such an argument.)

c. Is there evidence that the association between FEV and smoking is modified by sex? Provide descriptive statistics in support of your answer.

Ans: No. The difference between average FEV in nonsmokers and that in smokers is 0.18 l/sec for males and 0.17 l/sec for females. These stratified estimates of association are remarkably similar to each other, thus the effect of smoking on FEV is not modified by sex.

(Note that I chose to measure the association between smoking and FEV by the difference in average FEV between smokers and nonsmokers. My finding of no association is of course dependent upon the measure of association chosen. Had I decided to use ratio of mean FEV, I would have found that male smokers' average FEV was 2.49 /2.67 = 0.933 that of male nonsmokers and that female smokers' average FEV was 1.77 / 1.94 = 0.912 that of female nonsmokers. The answer would then depend upon whether I thought that 0.933 was sufficiently different from 0.912 to regard it as effect modification.)

d. What statistic would you present to describe the association between FEV and smoking? Provide the sentence you would use to report the results of your analysis.

Ans: While we found that the average FEV of 2.20 l/sec in smokers was only negligibly less than the average FEV of 2.22 l/sec in nonsmokers, this seeming lack of association may have been due to the fact that a disproportionate number of smokers were male and males, due to their tendency toward larger body size, naturally average higher FEV. When we compare smokers to nonsmokers of the same sex, we find that smokers average 0.175 l/sec lower FEV.

(In the remainder of the course we will consider whether observed differences are "statistically significant". I note that in this data, the sex-adjusted association between smoking and FEV is statistically significant: If there were no true association between smoking and FEV, there is less than a 0.02% chance of observing sex adjusted differences in average FEV that are as great as 0.175 l/sec.)

6. The following table presents the 2.5th percentile and 97.5th percentile for the distribution of sample means obtained from a sample of size *n* from a lognormal distribution having a mean of 10 and standard deviation of 10. From this table, what sample size would be required to be able to decide with 95% accuracy that an observed sample mean of 7.2 was not consistent with the data coming from a lognormal distribution with the stated mean and standard deviation?

n	2.5%ile	97.5%ile
1	1.44	31.52
4	3.63	21.54
9	5.25	17.87
16	6.23	14.88
25	6.72	14.06
36	7.13	13.45
49	7.56	13.00
64	7.92	12.98
81	8.00	12.14
100	8.25	12.22

Ans: The above table represents the "central 95%" of study results when computing sample means from lognormally distributed data having a mean of 10 and standard deviation of 10. That is, if I only gather a single observation, with 95% probability that observation will be between 1.44 and 31.52. As 7.2 falls in that range, I certainly cannot call 7.2 an unusual observation for a single measurement.

Similarly, if I gather 4 observations and take the sample mean, with 95% probability the observed sample mean will be between 3.63 and 21.54. Again, a sample mean of 7.2 is not that unusual.

Continuing in this vein I find that it is not until I collect data on 49 subjects that an observed sample mean of 7.2 would fall outside the central 95% of the sampling distribution for the sample mean of data drawn at random from a lognormal distribution having mean 10 and standard deviation 10.

(This is the conceptual way we go about choosing sample size for a scientific study.)

Grade distribution:

Highest possible: 125
Highest achieved: 124
Mean: 94.3
Std. Deviation: 20.5

Percentiles:

10th	20th	30th	40th	Median	60th	70th	80th	90th
60	76	87	92	97	102	109	113	118