Biost 514
Medical Biometry II

Midterm Examination Key

Name: _____

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. Suppose we are interested in studying the possible association between hormone replacement therapy and heart attacks. Consider the following study designs for hypothetical studies done in postmenopausal women in King County:

   A. We sample 5,000 women drawn randomly from the population of postmenopausal women in King county. Each woman is asked about use of hormone replacement therapy (yes/no) and history of heart attacks (yes/no).

   B. We randomly sample 2,500 women who use hormone replacement therapy, and 2,500 women who do not use hormone replacement therapy. Each woman is asked about her history of heart attacks (yes/no).

   C. We randomly sample 1,000 women who have had heart attacks and 4,000 women who have not had heart attacks. Each woman is asked about use of hormone replacement therapy (yes/no).

   a. (5 points) For each of the above studies, provide the name given in class for that type of study design.

   Ans: Study A is cross-sectional, study B is a cohort design, and study C is a case-control design.

   b. 5 points) Which of the above study designs can provide an estimate of the prevalence of heart attacks in the population of postmenopausal women in King County?

   Ans: The prevalence of heart attacks can be estimated only in the cross-sectional study (study A). From the cohort study (study B), we can estimate the prevalence of heart attacks among those taking HRT and among those not taking HRT. In the case-control study (study C), the proportion of patients having heart attacks was fixed by design.

   c. 5 points) Which of the above study designs can provide an estimate of the prevalence of hormone replacement therapy in the population of postmenopausal women in King County?

   Ans: The prevalence of HRT use can be estimated only in the cross-sectional study (study A). From the case-control study (study C), we can estimate the prevalence of HRT among those having had heart attacks and among those not having had heart attacks. In the cohort study (study B), the proportion using HRT was fixed by design.

d. 5 points) Which of the above study designs can provide information regarding an association between hormone replacement therapy and heart attacks? Justify your answer.

Ans: All three of the study designs can be used to detect an association between HRT and heart attacks. An association exists if heart attacks and HRT are not independent random variables. If they are independent, then Pr(heart attacks — HRT) = Pr(heart attacks — no HRT). These probabilities can be estimated from both the cross-sectional (study A) and the cohort (study B) studies. Similarly,if heart attacks and HRT are independent, then Pr(HRT use — heart attacks) = Pr(HRT use — no heart attacks). These latter probabilities can be estimated from both the cross-sectional (study A) and the case-control (study C) studies.

e. 5 points) Which of the above study designs can tell whether hormone replacement therapy causes heart attacks?

Ans: As they are all observational studies, none of them can tell whether HRT causes heart attacks, or, for that matter, whether heart attacks causes HRT use. Causation can only be determined from an interventional experiment, if then.

f. (Bonus: 10 points) Letting D= disease (heart attacks) and E= exposure (hormone replacement therapy), show that the ratio (odds of disease among exposed : odds of disease among unexposed) is equal to the ratio (odds of exposure among diseased : odds of exposure among nondiseased). (Recall the definition of conditional expectation is $Pr(D|E) = Pr(D,E)/Pr(E)$, and the definition of odds is $p/(1-p)$.)

Ans: Odds of disease among exposed is

$$\frac{Pr(D|E)}{Pr(\overline{D}|E} = \frac{Pr(D,E)}{Pr(E)}\frac{Pr(E)}{Pr(\overline{D},E)} = \frac{Pr(D,E)}{Pr(\overline{D},E)}$$

Similarly, odds of disease among nonexposed is

$$\frac{Pr(D|\overline{E})}{Pr(\overline{D}|\overline{E}} = \frac{Pr(D,\overline{E})}{Pr(\overline{E})}\frac{Pr(\overline{E})}{Pr(\overline{D},\overline{E})} = \frac{Pr(D,\overline{E})}{Pr(\overline{D},\overline{E})}$$

Thus the odds ratio comparing odds of disease among exposed to odds of disease among nonexposed is

$$\frac{Pr(D,E)Pr(\overline{D},\overline{E})}{Pr(\overline{D},E)Pr(\overline{D},\overline{E})}$$

Odds of exposure among diseased is

$$\frac{Pr(E|D)}{Pr(\overline{E}|D} = \frac{Pr(D,E)}{Pr(D)}\frac{Pr(D)}{Pr(D,\overline{E})} = \frac{Pr(D,E)}{Pr(D,\overline{E})}$$

Similarly, odds of exposure among nondiseased is

$$\frac{Pr(E|\overline{D})}{Pr(\overline{E}|\overline{D}} = \frac{Pr(\overline{D},E)}{Pr(\overline{D})}\frac{Pr(\overline{D})}{Pr(\overline{D},\overline{E})} = \frac{Pr(\overline{D},E)}{Pr(\overline{D},\overline{E})}$$

Thus the odds ratio comparing odds of exposure among diseasedsed to odds of exposure among nondiseased is

$$\frac{Pr(D,E)Pr(\overline{D},\overline{E})}{Pr(\overline{D},E)Pr(\overline{D},\overline{E})}$$

The equality of the two odds ratios is obvious by inspection.

(Note: The odds ratio is a suitable measure by which to judge an association, because if two variables are not associated the odds ratio is 1. It should also be noted that for a rare disease, the probability of not having the disease is very nearly 1, and the odds of having the disease is very nearly the probability of having the disease. In such a case, the odds ratio computed from either a cohort or a case-control study can be viewed as an estimate of the ratio of probabilities of disease (often called the rate ratio). Of course, if an event is not rare, then the odds of the event does not approximate the probability of the event. In this case, the odds ratio still can be used to assess associations, but can not be used to estimate the relative increase in the probability of disease.)

Appendix A contains hypothetical descriptive statistics from a hypothetical study of fiber supplementation as a preventive agent in colon cancer. The following variables are available:

$RACE$	patient's race: 1= Asian, 2= Black, 3= White, 4= Other
$WEIGHT$	patient's weight in pounds
$EDUC$	education: 1= did not graduate high school, 2= highschool graduate, 3= some college, 4= college graduate
TX	treatment group (1= placebo, 2= fiber)
$TIME$	time of last follow-up in years
$DEAD$	patient's status at last follow-up: 0= alive, 1= dead

2. (3 points each) For each of the following descriptive statistics presented in Appendix A, identify which of the above variables the specified statistic provides no scientifically meaningful descriptions of the sample. For each such variable, very briefly explain why not (just a few words should suffice to justify your entire answer).

   a. Sample mean

   Ans: The sample mean is only of scientific interest for uncensored numeric variables measured on either an interval or ratio scale, or a binary variable coded as 0 or 1. Thus the sample mean is not of scientific interest for the unordered variable $RACE$, the ordered categorical (non-interval measurement) $EDUC$, the censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time. While the sample mean of binary variable TX is indicative of the proportion of treated patients, the coding is such that 1 would have to be subtracted from that sample mean in order to get that proportion.

   b. Standard deviation

   Ans: The standard deviation is only of scientific interest for uncensored numeric variables measured on either an interval or ratio scale. Thus the standard deviation is not of scientific interest for the unordered variable $RACE$, the ordered categorical (non-interval measurement) $EDUC$, the binary variable TX, the censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time.

   c. minimum

   Ans: The minimum is only of scientific interest for uncensored ordered variables. Thus the minimum is not of scientific interest for the unordered variable $RACE$, the censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time. It is not a particularly exciting descriptive statistic for the binary variable TX.

   d. maximum

   Ans: The maximum is only of scientific interest for uncensored ordered variables. Thus the maximum is not of scientific interest for the unordered variable $RACE$, the

censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time. It is not a particularly exciting descriptive statistic for the binary variable TX.

e. 25th percentile

Ans: Percentiles are only of scientific interest for uncensored ordered variables. Thus the 25th percentile is not of scientific interest for the unordered variable $RACE$, the censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time. It is not a particularly exciting descriptive statistic for the binary variable TX.

f. median

Ans: Percentiles are only of scientific interest for uncensored ordered variables. Thus the median is not of scientific interest for the unordered variable $RACE$, the censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time. It is not a particularly exciting descriptive statistic for the binary variable TX.

g. 75th percentile

Ans: Percentiles are only of scientific interest for uncensored ordered variables. Thus the 75th percentile is not of scientific interest for the unordered variable $RACE$, the censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time. It is not a particularly exciting descriptive statistic for the binary variable TX.

h. mode

Ans: The mode as reported in Appendix A (the sample mode) is only of scientific interest for uncensored discrete variables. (Even then, it is not as interesting as knowing the frequency of the observation.) Thus the mode is not of scientific interest for the continuous variable $WEIGHT$, the censored variable $TIME$, or the variable $DEAD$ which is measured over variable periods of time.

3. (3 points each) Suppose we are interested in detecting associations between sex and each of the variables listed in Appendix A. For each of the following descriptive statistics, identify those variables which might be compared in a scientifically meaningful way using that statistic. Very briefly justify your answer (just a few words should suffice to justify your entire answer).

a. Sample mean

Ans: The sample mean will provide a meaningful comparison of distributions for any variable for which it is a scientifically meaningful descriptive statistic, as well as any uncensored ordered categorical variable. Thus we might want to use the sample mean to compare the distributions of $WEIGHT$ (a continuous variable measured on a ratio scale), $EDUC$ (an ordered categorical variable), and TX (a binary variable).

(Note that the reason that ordered categorical variables can be included here is because it is often of scientific interest to know whether there is a tendency for one group to have higher measurements than another. The sample mean will generally reflect any such tendency, even though the exact difference in means may not be of particular interest for variables not measured on an interval scale. While the sample mean could be used to detect differences between the distributions of unordered categorical variables which have been coded as numbers, the scientific interpretation as a shift toward higher values is lacking.)

b. Standard deviation

Ans: The standard deviation will provide a meaningful comparison of distributions for

any variable for which it is a scientifically meaningful descriptive statistic, as well as any uncensored ordered categorical variable. Thus we might want to use the standard deviation to compare the distributions of $WEIGHT$ (a continuous variable measured on a ratio scale), $EDUC$ (an ordered categorical variable).

c. minimum

Ans: The minimum is not a particularly useful measure for comparing distributions because its sampling distribution is heavily dependent on the sample size. Furthermore, it is unlikely that discrete variables having few possible categories would show different minimums. Nonetheless, it could make scientific sense to use the minimum for any uncensored ordered variable. It therefore could be used on the continuous variable $WEIGHT$ and and the ordered categorical variable $EDUC$. It is not likely to be of much interest for the binary variable TX, though it would not be wrong.

d. maximum

Ans: The maximum is not a particularly useful measure for comparing distributions because its sampling distribution is heavily dependent on the sample size. Furthermore, it is unlikely that discrete variables having few possible categories would show different minimums. Nonetheless, it could make scientific sense to use the maximum for any uncensored ordered variable. It therefore could be used on the continuous variable $WEIGHT$ and and the ordered categorical variable $EDUC$. It is not likely to be of much interest for the binary variable TX, though it would not be wrong.

e. 25th percentile

Ans: Percentiles could be used to compare any uncensored ordered variable, though they would not be likely to be of much use for discrete variables having few possible categories, and especially not binary variables. The 25th percentile could be used for comparing the distribution of the continuous variable $WEIGHT$ and the ordered categorical variable $EDUC$.

f. median

Ans: Percentiles could be used to compare any uncensored ordered variable, though they would not be likely to be of much use for discrete variables having few possible categories, and especially not binary variables. The median could be used for comparing the distribution of the continuous variable $WEIGHT$ and the ordered categorical variable $EDUC$.

g. 75th percentile

Ans: Percentiles could be used to compare any uncensored ordered variable, though they would not be likely to be of much use for discrete variables having few possible categories, and especially not binary variables. The 75th percentile could be used for comparing the distribution of the continuous variable $WEIGHT$ and the ordered categorical variable $EDUC$.

h. mode

Ans: The mode is a valid measure for all variables except censored variables and continuous variables (the problem with the latter is that the mode is defined for the density rather than the observations). Thus it could conceptually be used to compare distributions of $EDUC$, $RACE$, and TX. It would not be a good measure for $WEIGHT$, because that is a continuous random variable. Nor would it be appropriate for $TIME$ (due to its censoring) or $DEAD$ (due to it being measured over variable times.

(Note that the mode is not really a very useful measure for comparing distributions due to the difficulty in characterizing its sampling distribution.)

4. (5 points each) Suppose we are interested in comparing the treatment groups in the above hypothetical study with respect to the following variables. In each case what summary measure of the distribution would you choose to compare across treatment groups. Very briefly state your reasoning and provide a brief description of how the summary measure would be computed (you need not give a formula, but make sure I know what you have in mind).

   a. race

   Ans: Race is an unordered categorical variable. Thus the frequency distribution, as computed by counting the observations in each category, is really the only possibility. This might be compared for all categories, or just as a dichotomized variable (e.g., comparing the proportion that are nonwhite).

   b. weight

   Ans: Weight is a continuous variable, and I would typically use the mean or median to compare the distributions. The sample mean is a little less variable for data that are not prone to large outliers. Furthermore, it is easier to compute the sum of the observations and divide by the sample size than it is to sort them and find the middle value. Thus the mean would be my top choice

   c. education

   Ans: Education is an ordered categorical variable. We could comparing the frequency distribution as we did with race, however, I would rather take advantage of the ordering and look for tendencies for one group to be more educated than the other using the sample mean.

   (I should note that some people look at a more specialized analysis called the proportional odds model for ordered categorical data. It is harder to compute and more difficult to understand, and furthermore the underlying assumption of that model is a pretty strong assumption.)

   d. time to death

   Ans: The measurements of time to death is subject to censoring in this data set. Thus the distribution of time to death would have to be compared across groups using methods suitable for censored data. One such possibility would be to compute the survival probability estimates using Kaplan-Meier curves, and to compare the estimated probability of survival at specific time points, say 10 years.

5. (10 points each) Appendix B contains selected results from an analysis of a hypothetical dataset exploring an association between FEV and smoking in 11 - 14 year olds. Use these results to answer the following questions.

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

   Ans: There is a slight suggestion of a trend toward 0.16 l/sec higher FEV in smokers (sample mean 3.44) than in nonsmokers (sample mean 3.28). (This difference may just due to random sampling error.)

   b. How does your answer change when you consider measures of association in each age group? Provide descriptive statistics to support your answer.

   Ans: Within each age group, the differences in sample means between smokers and nonsmokers shows the reverse trend, with lower average FEV in smokers than in nonsmokers. In 11 year olds, the average FEV is 0.40, 0.40, 0.41, and 0.34 l/sec higher in nonsmokers than smokers for 11, 12, 13, and 14 year old, respectively.

c. What do the results in parts (a) and (b) suggest about the role age might play in the estimation of an association between smoking and FEV?

Ans: The fact that similar degrees of association are observed across all age groups, but that the degree of association when all ages are combined is indicative of the smoking-FEV association being confounded by age. This would be caused by their being an association between age and smoking in the sample (as clearly evidenced by the fact that approximately one-fourth of nonsmokers are 14 in the sample, but one-half the smokers are 14) and an association between age and FEV (as clearly evidenced by the average FEV in nonsmokers increasing from 3.44 to 3.56 to 5.20 to 5.86 as ages range from 11 to 14).

d. (Bonus) Consider the association between sex and height in the combined sample as well as in the strata defined by age. What does the data suggest about the role age might play in the estimation of an association between sex and height?

Ans: In comparing the average difference in height between the sexes, there is a suggestion of a changing degree of association by age. The average difference in height (males - females) is 0.18, -0.92, 3.58, and 7.23 inches for 11, 12, 13, and 14 year olds, respectively. The fact that the effect of sex is not constant across age strata suggests that there is an age-sex interaction in the average height of children. This is also described by saying that age modifies the effect of sex on height.

6. (10 points each) Appendix C contains results from simulated studies in which the response variable X has a t distribution with k degrees of freedom. The t distribution is a symmetric distribution having in all cases a median of 0 and a kurtosis (tendency to extreme values) which decreases with the degrees of freedom. The t distribution with 1 degree of freedom is the Cauchy, which has such heavy tails that it does not have a mean. The t distribution with infinite degrees of freedom is the Normal distribution.

For each value of k, 1,000 studies were simulated in which the sample size was $n = 30$. For each such sample, the sample mean $\overline{X}$ and the sample median X_M were calculated, and these measurements then constituted a data set from which descriptive statistics could be computed. (These data sets are thus estimating the sampling distribution for the sample mean and sample median across a large number of repeated experiments.)

Appendix C then presents for each choice of μ the mean, standard deviation, and selected percentiles of the distribution of those statistics obtained across the 1,000 simulated replications.

a. Based on the statistics presented in Appendix C, does it appear that the sampling distribution for the sample mean and sample median have the same central tendencies for samples from the t distribution? State your reasoning.

Ans: In every case, the average of the sample means across the 1,000 studies and the average of the sample medians across the 1,000 studies is very close to the true value for the median of 0.

(Note that I told you that the distributions had median 0 and were symmetric. Thus, so long as the mean exists, the mean of the distribution is in each case also 0. Because the sample mean and the sample medians tend to be on average equal to the true mean and median, we call those estimators unbiased. That is, an unbiased estimator is one that when computed for a large number of samples will have expectation equal to the true population parameter.

b. Based on the statistics presented in Appendix C, what can you say about the relative sampling variability of the sample mean and sample median across samples as the tendency to heavy tails (kurtosis) increases?

Ans: When the degrees of freedom are small (and the tendency to extreme values is high), the standard deviation of sample means is higher than the standard deviation of sample medians. For instance, the standard deviation of sample means

is .519, .323, and .264 for 2, 3, and 4 degrees of freedom, respectively, while the standard deviation for the sample medians is .249, .242, and .242, respectively, for those same cases.

When the degrees of freedom are high (and the tendency to extreme values is lowe), the standard deviation of sample means is lower than the standard deviation of sample medians. For instance, the standard deviation of sample means is .205, .192, and .191 for 10, 20, and 30 degrees of freedom, respectively, while the standard deviation for the sample medians is .233, .225, and .226, respectively, for those same cases.

If it is safe to generalize from these distributions (and it is), then it would appear that higher kurtosis leads to markedly greater variability in the sampling distribution of the sample means than in the sample medians. (It is noteworthy that the variability of the two statistics is about the same with 5 degrees of freedom, which still has relatively heavy tails.

c. If you are interested in estimating the central tendency of a distribution, what issues must you consider as you choose a particular descriptive statistic?

Ans: First and foremost, of course, you consider the scientific question. A part of that may be consideration of whether you want "outliers" to affect the summary measure of the distribution.

In those situations where there is some scientific flexibility in the choice of summary measures, then we can consider the statistical issues. (One such setting would be where you consider a "shift alternative" in which the distribution of the response always has the same shape, but it is shifted between groups. In that situation, either the mean or the median will capture the "shift".) Based on the above, when either the median or mean would address the scientific question of interest, it would probably be most precise to use the sample median to describe the "typical" measurement when the distribution is prone to extreme measurements.

(It should be noted, however, that it is entirely possible that one group might have a higher mean but lower median than the other.)

APPENDIX A

Descriptive statistics for hypothetical study of fiber in colon cancer prevention.

	n	Msng	Mean	SD	Min	25%	Mdn	75%	Max	Mode
RACE	200	0	2.99	0.48	1.00	3.00	3.00	3.00	4.00	3.00
WEIGHT	200	0	149.75	18.07	85.00	138.00	148.00	161.00	200.00	138.00
EDUC	200	0	2.62	0.91	1.00	2.00	3.00	3.00	4.00	3.00
TX	200	0	1.54	0.50	1.00	1.00	2.00	2.00	2.00	2.00
TIME	200	0	11.59	5.35	0.01	7.64	12.25	16.36	19.98	10.66
DEAD	200	0	0.47	0.50	0.00	0.00	0.00	1.00	1.00	0.00

APPENDIX B

Selected descriptive statistics from a hypothetical data set studying the association between smoking and FEV in children. Available data include sex, age, height, smoking status (yes/no), and FEV.

Sample Mean and Standard Deviation of FEV
by Smoking Status
(All Ages Combined and Stratified by Age

	Nonsmokers				Smokers			
Age	n	msng	mean	std dev	n	msng	mean	std dev
all	164	0	3.28	2.28	36	0	3.44	2.23
11	47	0	3.44	1.83	4	0	3.04	1.66
12	38	0	3.56	1.89	7	0	3.16	2.33
13	39	0	5.20	2.15	7	0	4.81	1.84
14	40	0	5.86	2.30	18	0	5.52	2.12

Sample Mean and Standard Deviation of Height
by Sex
(All Ages Combined and Stratified by Age

	Females				Males			
Age	n	msng	mean	std dev	n	msng	mean	std dev
all	110	0	62.00	2.03	90	0	64.25	2.52
11	25	0	59.54	0.96	26	0	59.72	1.03
12	26	0	61.16	1.20	19	0	60.24	0.94
13	23	0	62.38	0.97	23	0	65.96	1.05
14	36	0	64.07	1.15	22	0	71.30	1.07

APPENDIX C

Distribution of statistics from 1,000 simulated samples of size 30 from t distributions having degrees of freedom k. For each distribution, 1,000 random samples were drawn, and the sample mean and median were computed. Presented in the following table are the descriptive statistics (mean and standard deviation) of the sample means and sample medians across the 1,000 replicated studies.

k	Mean $\overline{X}$	SD of $\overline{X}$	Mean X_M	SD of X_M
2	-0.029	0.519	-0.005	0.261
3	-0.015	0.323	-0.010	0.249
4	-0.010	0.264	-0.010	0.242
5	-0.011	0.244	-0.005	0.242
7	0.000	0.213	-0.009	0.229
10	-0.006	0.205	-0.004	0.233
20	0.003	0.192	0.000	0.225
30	-0.002	0.191	0.002	0.226