Biost 517: Applied Biostatistics I

Emerson, Fall 2007

Homework #2 Key

October 16, 2007

Written problems: To be handed in at the beginning of class on Wednesday, October 10, 2007.

> *On this (as all homeworks) unedited Stata output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

1. The class web pages contain descriptions of two datasets
    - Chemosensitizer data (chemo.doc)
    - Mayo PBC data (liver.doc)

   a. For each of the described scientific questions, briefly characterize the type of statistical question to be answered. That is, using the classification presented in class, characterize the problem as clustering of cases, clustering of variables, quantifying distributions within groups, comparing distributions across groups, or prediction, identifying any variable whose distribution is of interest and any groups that might be being compared.

Answer:

For the chemosensitizer dataset, the distribution of the counts of surviving cell colonies is compared first across dose of doxorubicin to create the IC$_{50}$, a measure of the effect of doxorubicin on cell survival. The sensitization factor for a particular chemosensitizer, is then a comparison (using a ratio) of the effect of doxorubicin in two different subgroups: one with the chemosensitizer and one without. In this sense, the chemosensitizer is modifying the effect of doxorubicin. Finally, our interest is in comparing the sensitization factor across experiments done in 10% calf serum and those done in 100% human serum. So we are thus considering how effect modification might differ according to the value of the fourth variable—such constitutes a three way interaction between doxorubicin concentration, chemosensitizer, and serum. So our primary question corresponds to the question labeled 4c in the lectures: a comparison of associations between variables across subgroups.

For the PBC data, the first question is one of comparing the distributions of survival across groups defined by treatment status (D-penicillamine vs placebo)—so question type 4b as we discussed in lecture.. The second question is concerned with predicting the time of survival for individual patients according to the various laboratory values. This then is a question of type 5 from class lecture.

   b. For each of the datasets, classify the available measurements with respect to the statistical role they might play in answering the scientific question. That is, using the classification presented in class, identify which variables might be outcome measurements, predictors of interest, subgroup identifiers for interactions, potential confounders, precision variables, surrogates for the response, or irrelevant.

Answer:

For the chemosensitizer data set, the counts of surviving cell colonies is the response variable, while the other three variables (doxorubicin concentration, type of chemosensitizer, and serum used) all model the effect modification.

For the PBC data, the response variables are *obstime* and *status*, which together contain the information about the possibly right censored time of survival. The predictor of interest for the first question is the variable indicating treatment. All other variables are of scientific interest due to the possibility that they might be predictive of survival (note that all of the other variables were measured prior to the start of treatment). Randomization precludes any systematic association in the sample between those variables and the treatment. Hence, we are not worried about confounding, and our only interest in the other variables are as precision variables.

With regard to the prediction question posed for the PBC dataset, our response variable is still the possibly censored survival times as recorded using *obstime* and *status*. All other variables are of interest as variables that might increase the precision of our predictions, either singly or jointly. (In prediction problems, the questions of confounding and effect modification are largely not of interest.)

c. For each of the datasets, classify the available measurements with respect to the type of measurement: qualitative versus quantitative, unordered versus partially ordered versus ordered, discrete versus continuous, and interval versus ratio.

Answer:

For the chemosensitizer dataset, the cell colony counts are discrete count data, the doxorubicin concentration is a quantitative, ratio variable (even though only limited concentrations were considered in this experiment, we know that other concentrations exist and we also know that there is a well-defined zero value), the treatment variable recording the chemosensitizers is a nominal (unordered categorical) variable, and the indicator of serum used is a binary variable (though this is a somewhat moot point for a binary variable, it should be recognized that the numerals '10' and '100' are merely labels, rather than being any sense of a numerical measurement of some quantity).

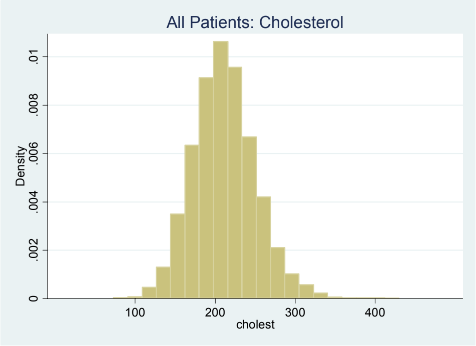
For the PBC dataset, quantitative variables measured on a ratio scale include *age, albumin, alkphos, bili, cholest, platelet, protime, sgot,triglyc,* and *urinecu.* The variables *ascites, edema, hepmeg, sex, spiders,* and *treatment* are binary variables. *Edmadj* and *stage* are ordered categorical variables. The variables *obstime* and *status* represent the censored quantitative variable measuring the time to survival and the binary variable indicating the complete observations, respectively.

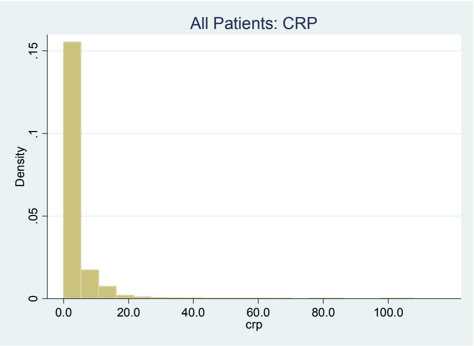
This problem deals with a data set containing various measurements made on a sample of generally healthy elderly adults. The primary goal in assembling this particular data set was to investigate the role of chronic inflammation in patient survival. The data (inflamm.txt) and documentation (inflamm.doc) can be found on the class web pages.

a. The variable *ttodth* represents an incomplete measurement of the time from study enrollment to a patient's death. That is, for some patients, *ttodth* contains the number of days between study enrollment and death, and for other patients *ttodth* contains the number of days between study enrollment and "locking" of the database for data analysis. Such data is called "right censored", because when the variable *death=0*, we only know that the patient survived longer than the time recorded in *ttodth*. We do not know the exact timing of the patient's death. In the prefatory remarks to this problem, I suggested that you create a variable *surv4yr* indicating whether a patient has survived at least 4 years. Why is this variable valid scientifically? Provide descriptive statistics justifying your answer.

Answer: **The possibililty of censored observations means that we must be careful in interpreting the values of the variable *ttodth*. That is, we must look at the value of the variable *death* to see whether *ttodth* is the actual survival time or merely the time that the patient was last known to be alive. Now, prior to the earliest time of censoring, we can trust the values of *ttodth* to represent a true survival time. So we examine the distribution of *ttodth* in groups defined by the value of *death*. When we do that, we find that the minimum value of *ttodth* in the subjects still alive at last follow-up is slightly over 4 years. Hence, anyone with *ttodth* less than 4 years must have died at that time, and everyone with *ttodth* greater than 4 years (1,461 days) must have survived at least 4 years. (See the annotated Stata log for the Stata commands that I used to ascertain this.)**

b. Using the two laboratory values of cholesterol and C reactive protein generate the following descriptive statistics for each group defined by whether or not they survived for 4 years:
   - Histogram
   - Number of cases with missing data
   - Mean
   - Geometric mean
   - Median
   - Mode (it suffices to take an approximate mode from a histogram)
   - Standard deviation
   - Variance
   - Minimum and maximum
   - Range (the difference between minimum and maximum)
   - 25th, 75th percentiles
   - Interquartile range (the difference between 25th and 75th percentiles)
   - Proportion of cases with "high" laboratory values (as defined above)





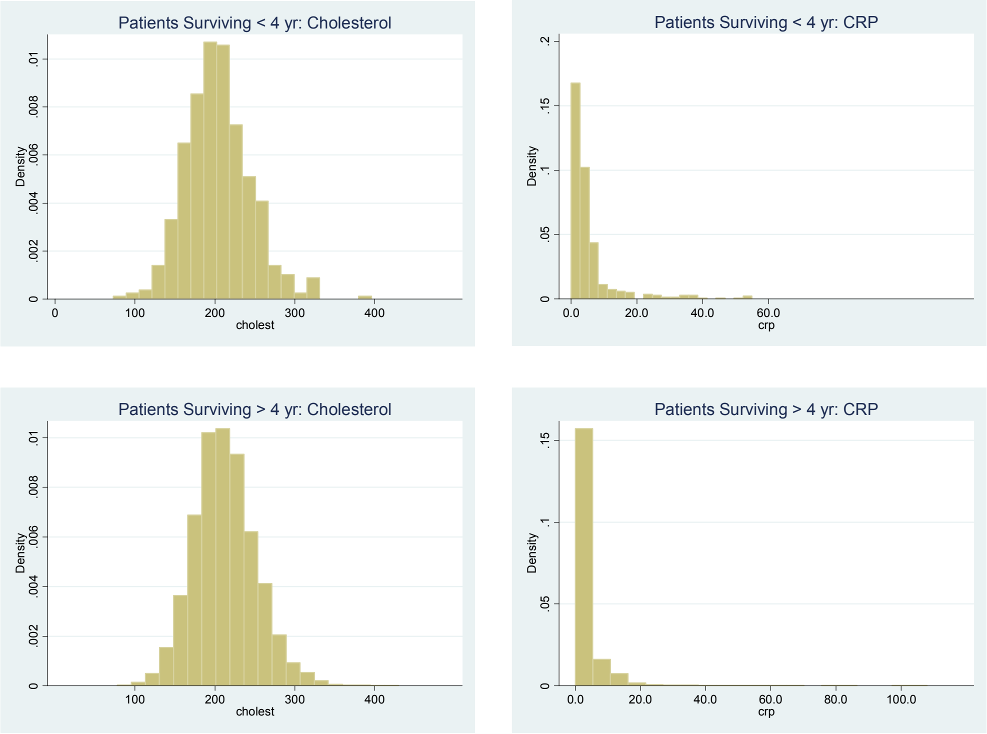


Figure 1: Distribution of serum cholesterol and blood C reactive protein for all patients combined as well as within groups defined by four year survival status.

Table 1: Descriptive statistics for serum cholesterol and blood C reactive protein for all patients combined as well as within groups defined by four year survival status.

	N msng	N	Mean	SD	Geom Mean	Min	25th %ile	Mdn	75th %ile	Max	IQ Range	Range	Prop High[2]
	Subjects Surviving Less Than 4 Years												
Cholesterol (mg/dl)	9	486	204	41	200	73	176	202	229	396	53	323	0.126
C Reactive Protein (mg/l)	11	484	5.4	8.1	3.0[1]	0	1	3	6	55	5	55	0.378
	Subjects Surviving More Than 4 Years												
Cholesterol (mg/dl)	38	4467	213	39	209	78	187	211	236	430	49	352	0.158
C Reactive Protein (mg/l)	56	4449	3.4	5.9	2.0[1]	0	1	2	3	108	2	108	0.223
	All Subjects												
Cholesterol (mg/dl)	47	4953	212	39	208	73	186	210	236	430	50	357	0.154
C Reactive Protein (mg/l)	67	4933	3.6	6.2	2.1[1]	0	1	2	3	108	2	108	0.238

[1] Measurements of 0 were assumed to be below a lower limit of detectability of 1 mg/l. These nondetectable levels were imputed to be 0.5 mg/l only when computing the geometric mean.

[2] "High" values were defined as serum cholesterol values greater than 250 mg/dl and C reactive protein levels greater than 3 mg/l.

Answer:

The figures and table above provide the descriptive statistics requested. The mode for cholesterol can be seen from the graph to be approximately 200 mg/dl for all patients combined, with a slightly lower mode for patients dying within 4 years and slightly higher for patients surviving at least 4 years. The mode for CRP was found to be 1 mg/l for all patients combined, as well as for groups defined by four year survival status.

Special attention should be paid to the fact that 21 subjects surviving less than 4 years had a C reactive protein measurement of 0, as did 407 subjects surviving 4 or more years. The geometric mean can only be computed using positive valued random variables. There are several rational approaches that could be taken here:

- **Report that it does not make sense to take the geometric mean of a variable that can take on values of 0 (or negative values), and do not report anything.**
- **Presume that all subjects with a reported value of 0 actually had some small positive value above the limit of detectability. We could then "impute" what the true value might be for those subjects by**
    - **Assigning them all to be the same as the lowest reported value (in this case, 1 mg/l). This does not seem as rational, because they were all probably below that level.**
    - **Assigning them all to be equal to one-half the lowest reported value (which in this case would lead to a choice of 0.5 mg/l). This might seem rational if you thought that the low measurements were uniformly distributed between 0 and 1. The imputed value of 0.5 is then the average value.**
    - **Assigning them all to be equal to the midpoint of a range that would not be rounded to the lowest reported value. That is, if we imagined that each CRP measurement were rounded, then all measurements truly between 0.5 and 1.5 would have been reported as 1 mg/l. Measurements below the detectable**

limit would then be between 0 and 0.5. We could then impute those values as 0.25 mg/l.

- **Assign values to individuals randomly from the interval thought to be below the lower limit of detectability. We could use a random number generator to come up with values between either 0 and 1 or between 0 and 0.5, depending upon how you thought the lower limit of detectability was implemented. Complicated statistical models could be derived to decide which subjects were most likely to have the lowest values, or we could just assume they were uniformly distributed over the interval. I note that for this exercise, all that would have mattered is what the geometric mean of those random values was within each survival group, because we were not doing more complicated analyses trying to assess associations between, say, CRP and cholesterol.**

In computing the value presented in the table, I chose the second of the above options for measurements below the limit of detectability. Of course, I have no way of knowing which was the best way, but the method I took is one commonly used method. Whatever you choose, that decision should be made prior to examining the resulting statistics. It is possible for your choice to have huge impact on the statistics, and you should avoid any “data-driven” choices.

For each laboratory test, how would you answer the question regarding whether measurements made on longer surviving patients tend to be “better” or “worse” than those made on patients surviving less than 4 years?

<u>**Answer:**</u>

Patients dying within four years seem to have tended toward lower cholesterol values, no matter whether we use the mean, geometric mean, median, mode, the 25th percentile, the 75th percentile, minimum, maximum, or the proportion having values greater than 250 mg/l to summarize the distribution.

Similarly, patients dying within four years seem to tend toward higher CRP levels for most of the summary measures, though for this measurement we do note that the maximum CRP value is found among those patients surviving at least four years. The sample maximum however is less useful as a summary measure for the purposes of comparing distributions, because it is heavily dependent upon the sample size. In this problem, the much larger sample size for patients surviving at least 4 years would lead us to expect a tendency toward more extreme values.

2. Suppose you are an unethical researcher who takes the ill-advised position of siding with my daughter in her unending quest to “improve” my diet. You thus want to “prove” that death within 4 years is associated with higher serum cholesterol (and thereby condemn me to eating broccoli at least 8 times per week).

a. Alter one cholesterol measurement (tell which case you use by row number and tell how you change that cholesterol measurement) in such a way that would have the mean cholesterol for patients dying within four years at least 10 mg/dl higher than the mean cholesterol for patients surviving longer than 4 years.

<u>**Answer:**</u> **As shown in table 1, the subjects dying early had a mean cholesterol level that is about 9 mg/dl lower than the mean for the patients surviving four years. I thus want to raise the average for the poor survivors by 20 mg/dl. As there are 486 such subjects, all I need to**

do is increase the cholesterol level of one subject in the poor surviving group by 486 x 20 mg/dl = 9720 mg/dl. I arbitrarily decided to increase the value for the patient who previously had the lowest serum cholesterol in that group (id= 2924). The following table presents selected descriptive statistics following this change. Note that I obtained the desired effect on the sample mean, but did not change the other measures of location (geometric mean, median, proportion with measurements over 250 mg/dl) in any meaningful way.

surv4yr	N	mean	sd	min	p25	p50	p75	max	prop high	geom mn
0	486	224	437	93	176	202	229	9793	0.126	202
1	4467	213	39	78	187	211	236	430	0.158	209
Total	4953	214	142	78	186	210	236	9793	0.154	208

b. Alter one cholesterol measurement (tell which case you use by row number and tell how you change that cholesterol measurement) in such a way that would have the geometric mean cholesterol for patients dying within four years at least 10 mg/dl higher than the geometric mean cholesterol for patients surviving longer than 4 years.

<u>**Answer:**</u> **As shown in table 1, the subjects dying early had a geometric mean cholesterol level that is about 9 mg/dl lower than the geometric mean for the patients surviving four years. I suspect the approach taken by most of you would have been trial and error: Guessing some large number to use in the poor survivors and checking to see when the geometric mean would be sufficiently higher. Alternatively, we could easily use some very small number (very close to zero) in the long survivors to achieve the same thing. I do note that I could have figured out how much to increase the value of one measurement in a manner much like I did for the arithmetic mean. But because the geometric mean is a multiplicative measure, I would need to consider the ratio not the difference. For the brave of heart, I describe the process below. If you want to skip the rationale, note at least the magnitude of the value used to achieve the desired results.**

I can satisfy the requirements of this problem if I increase the geometric mean in the poor survivors to 220, which is a 1.1 fold increase in the current geometric mean. I can do this by increasing each value by a factor of 1.1, or a single value by a factor of $1.1^{486} = 1.309 \times 10^{20}$, since there are 486 subjects in the group. Again, I arbitrarily decided to increase the value for the patient who previously had the lowest serum cholesterol in that group (id= 2924). The following table presents selected descriptive statistics following this change. Note the desired effect on the geometric mean, and the huge effect on the sample mean and sample standard deviation. The other measures of location (median, proportion with measurements over 250 mg/dl, quartiles) are not affected in any meaningful way.

surv4yr	N	mean	sd	min	p25	p50	p75	max	prop high	geom mn
0	486	1.97E+19	4.33E+20	93	176	202	229	9.55E+21	0.126	220
1	4467	213	39	78	187	211	236	430	0.158	209
Total	4953	1.93E+18	1.36E+20	78	186	210	236	9.55E+21	0.154	210

I can also satisfy the requirements of this problem if I decrease the geometric mean in the long survivors to 190, which represents a geometric mean only 0.9 times as high as the current geometric mean. I can do this by decreasing each value by a factor of 0.9, or a single value by a factor of $0.9^{4467} = 3.99 \times 10^{-205}$, since there are 4467 subjects in the group. I note, however, that Stata has trouble representing a number that small in its functions, so it would just change the measurement to 0. So I can't get Stata to do the problem this way. I

could perform the problem on the log scale (create a variable logchol = log(cholest)) and change one measurement and achieve the result.

There is a key issue here: We would all recognize that 10^{21} is a huge outlier relative to the next highest value of 430. But on the log scale, 10^{-205} is a huge outlier relative to the next lowest value of 78, even though it might not look that bad on the untransformed scale. Hence, when using the geometric mean with values below the detectable limit, we must be very careful in imputing small values: We might be creating large outliers.

c. Alter one cholesterol measurement (tell which case you use by row number and tell how you change that cholesterol measurement) in such a way that would have the median cholesterol for patients dying within four years at least 10 mg/dl higher than the median cholesterol for patients surviving longer than 4 years. If it is not possible, explain why not.

Answer: **This cannot be done. The median is the value that half the measurements exceed and that exceeds the other half of the measurements. By changing one measurement, I can at most shift the median to the value of the measurement that is immediately less than the median or the measurement that is immediately higher than the median.**

Now in the poor survivors, the median was 202 mg/dl. But there were 4 measurements at 201 mg/dl, 2 measurements at 202 mg/dl, and 7 measurements at 203 mg/dl. At best, by changing one measurement I could only manage to get the median to be 203 mg/dl in that group. Similarly, the subjects surviving at least 4 years, the median was 211 mg/dl. But there were 49 subjects how had that value, 44 who had 210 mg/dl and 57 who had 212 mg/dl. By changing one value, the most I could hope to change the median would be to increase it to 212 mg/dl (and even this might not be possible).

d. What does the above say about the influence that an outlier can have on the group mean, geometric mean, or median?

Answer: **Clearly, the mean is highly influenced by an outlying value. The geometric mean is less influenced, and in fact it is relatively not influenced by large outliers until they become absurdly extreme. (But as noted above, we must be careful to judge small outliers on the log scale when using the geometric mean.) The median is generally impervious to the effect of outliers.**

So then we have the question: Do we want a summary measure influenced by outliers or not? This must be answered scientifically first: Many treatments and/or risk factors have greatest effect on the most extreme subjects. If we choose to perform comparisons on measures that are unaffected by outliers, we might miss the effect. But on the other hand, the presence of outliers greatly decreases our statistical precision. So if several summary measures are equally relevant scientifically, then when measurements are prone to large outliers (e.g., laboratory measurements in diseased patients), we might want to consider geometric means or medians, rather than means.