Biost 517: Applied Biostatistics I
Emerson, Fall 2012

Homework #2 Key
October 10, 2012

1. The class web pages contain descriptions of two datasets
    - Salary data (salary.doc)
    - Beta carotene data (carot.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 salary dataset, the primary question relates to comparing the distribution of monthly salary between men and women. In order to distinguish current discriminatory practices from historical discrimination or personal preferences, it will likely be necessary to make the comparisons across sexes when controlling for factors such as field, year hired, etc.

For the beta carotene dataset, the primary question relates to comparing the distribution of plasma beta carotene levels across dose groups.

    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 salary dataset, the monthly salary will be the outcome variable and sex will be the predictor of interest. Degree, year of degree, year hired, field, and administrative duties are primarily potential confounders, because they are causally associated with salary and are likely to be distributed differently between men and women due to historical discrimination and/or differences in the personal preferences of men and women. I note that these variables may, however, in part represent a mechanism of discrimination. For instance, underrepresentation of women in some fields might be indicative of current discrimination. Because rank is intimately related to salary and is one possible mechanism for discrimination, rank would largely be regarded as a surrogate for the response. The id variable is only used for identifying measurements made on the same subject.

For the beta carotene dataset, the post randomization plasma beta carotene and vitamin E measurements are the outcomes. The measurements made at 9 months are probably of greatest interest, in which case the other measurements should be regarded as surrogates for the response. Dose is the predictor of interest. As this is a clinical trial, we do not need to worry so much about confounding (though the sample size is indeed relatively small). The baseline plasma beta carotene and vitamin E levels may provide added precision, as might sex, age, cholesterol, weight, body mass index, and percent body fat. These last 6 variables might also be of interest for exploratory analyses considering effect modification, but if such analyses had not been pre-specified, then I would not do them.

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 salary dataset, the monthly salary, year of degree, and year hired are all quantitative variables. Conceptually, all of those variables could be regarded as continuous. Salary is a ratio variable, because it has a natural zero. Year of degree and hire are merely interval scale variables, because there is no natural zero. (We could convert both of these variables to the number of years ago that the degree was obtained or the faculty member was hired, and that number would be on a ratio scale.) Rank is an ordered categorical variable. Field and degree are nominal (unordered categorical) variables. All categorical variables are discrete. Administrative duties is a discrete, binary variable, and it can be regarded as ordered if we want to.

For the beta carotene dataset, all the plasma beta carotene and vitamin E levels are continuous variables measured on a ratio scale. This is also true of age, dose, weight, BMI, cholesterol, and percent body fat. (Note that we only sampled dose at five levels, but conceptually there are an infinite number of doses possible.) Sex is a discrete, binary variable that we can regard as ordered when it is convenient to do so.

2. This problem deals with a data set containing laboratory data from a clinical trial of methotrexate (MTX) in primary biliary cirrhosis (PBC). The data and documentation can be found on the class web pages. The file mtxlabs.txt can be downloaded and read into Stata using the command (typed all on one line)

```
infile case ptid str16 rdate tx week ondrug bili alb ptinr fvc fvcpred dlco dlcopred  using  mtxlabs.txt
```

Note the definition of variable rdate as a "string" variable. We need to do this because the randomization date is in the mm/dd/yyyy format. It is easier to do arithmetic with dates if they are in "Julian" dates, which are defined as the number of days from some reference (each computer program tends to use its own reference date). Stata has a facility to change string representation of dates to Julian dates. For instance, the following code will generate a new variable rndmzdt by converting the string variable rdate to a Julian date (which is, of course, just an integer):

- **g rndmzdt = date(rdate,"mdy")**

Because a Julian date is not very interpretable at first reading by a human, it is probably wise to keep the string variable around, as well.

The questions can be answered using the Stata commands (other commands would also work)

- **tabstat** …, **stat(n mean sd min p25 med p75 max iqr r) col(stat) format**
- **hist** …, **bin(20)**
- **means …**

Note that I added the statistics "iqr" for interquartile range and "r" for range. You will have to use "means" to get the geometric mean, though it could also be obtained by generating a new variable that is the log transformed lab values, taking the mean of that new variable, and then exponentiating the result (you would do this last step with "display" or a hand calculator).

You many want to create a new variable which dichotomizes lab values at the requested levels. There are many ways to do this. One way is as follows:

- **g**enerate *bilihigh = 1*
- **replace** *bilihigh* = 0 if *bili < 2.5*

If you use this method, you will also need to make sure that missing data is handled appropriately. In this case, we can set all cases with missing data for *bili* to also have missing for *bilihigh* (a period by itself is used as the code for missing data):

- **replace** *bilihigh* = . if *bili* == .

Similar variables could be created to indicate low albumin (less than 2.5) or high PTINR (greater than 1.3).

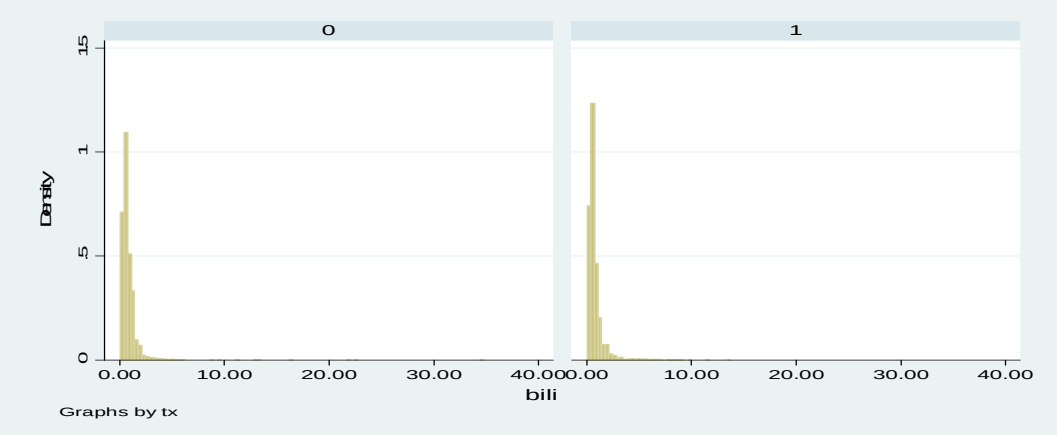
Using the three laboratory values of bilirubin, albumin, and PTINR, generate the following descriptive statistics for each treatment group in the sample.

- 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)
- 25^{th}, 75^{th} percentiles
- Interquartile range (the difference between 25^{th} and 75^{th} percentiles)
- Proportion of cases with "advanced" liver disease, which we will define as a bilirubin that is 2.5 or greater, an albumin that is 2.5 or less, or a PTINR greater than 1.3.
- Proportion of cases with "severe" liver disease, which we will define as a bilirubin that is 5.0 or greater, an albumin that is 2.0 or less, or a PTINR greater than 2.0

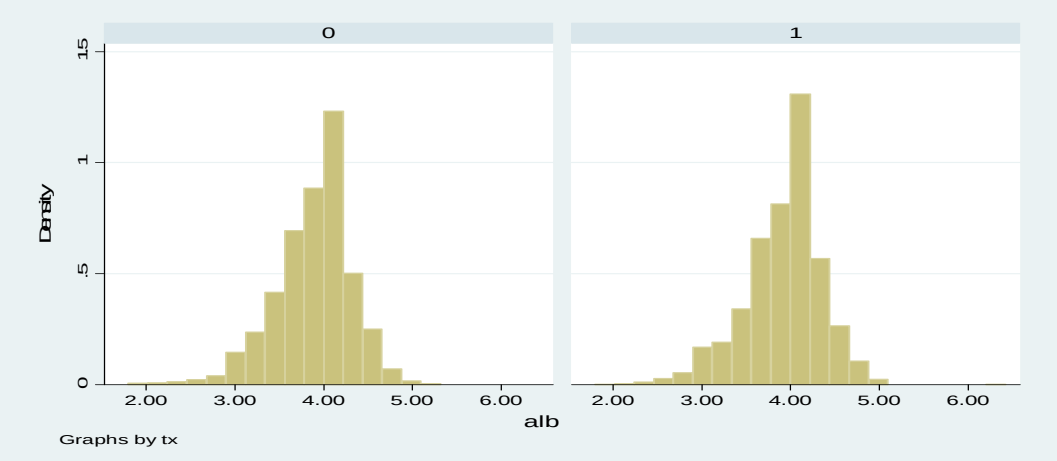
Answer:

Histograms for bilirubin, albumin, and PTINR by treatment group are displayed below:

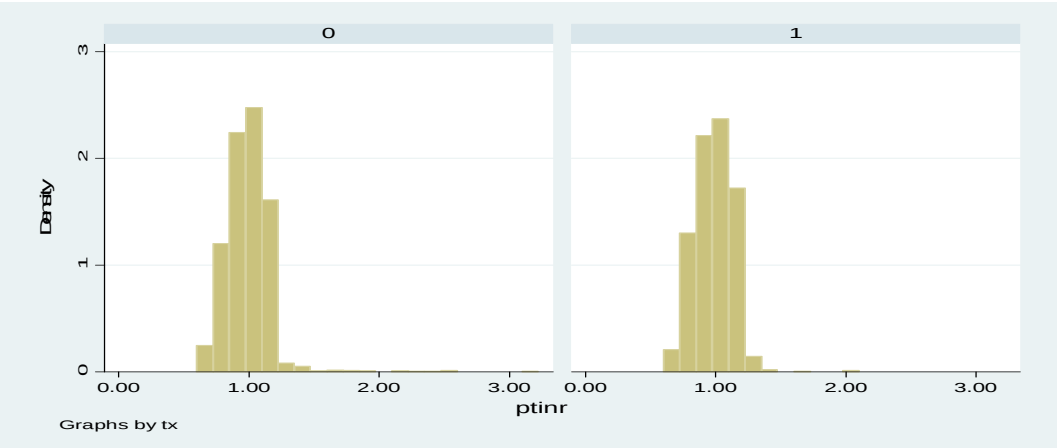
Bilirubin: **Placebo** **Methotrexate**



Albumin: **Placebo** **Methotrexate**



PTINR: **Placebo** **Methotrexate**



Descriptive statistics by treatment arm and for both treatment arms combined. Note that "Advanced Disease" is defined differently for each laboratory test: bilirubin 2.5 or higher, albumin 2.5 or lower, PTINR greater than 1.3. Similarly "Severe Disease" is defined by bilirubin 5.0 or higher, albumin 2.0 or lower, PTINR greater than 2.0.

	Msng	N	Mean	SD	Geom Mean	Mode	Min	25th %ile	Mdn	75th %ile	Max	Range	IQR	Prop Adv	Prop Sev
							Placebo Arm								
Bilirubin	12	4797	0.83	1.04	0.65	0.50	0.07	0.50	0.60	1.00	34.40	34.33	0.50	0.0254	0.0052
Albumin	355	4454	3.88	0.43	3.85	4.00	1.80	3.60	3.90	4.20	5.20	3.40	0.60	0.0085	0.0013
PTINR	2987	1822	0.97	0.17	0.96	1.00	0.60	0.90	1.00	1.00	3.10	2.50	0.10	0.0170	0.0044
							Methotrexate Arm								
Bilirubin	11	4984	0.80	0.81	0.63	0.50	0.10	0.40	0.60	0.80	13.50	13.40	0.40	0.0317	0.0094
Albumin	366	4629	3.91	0.44	3.88	4.00	1.90	3.70	4.00	4.20	6.20	4.30	0.50	0.0063	0.0004
PTINR	3102	1893	0.97	0.14	0.96	1.00	0.60	0.90	1.00	1.00	2.10	1.50	0.10	0.0048	0.0005
							Both Treatment Arms Combined								
Bilirubin	23	9781	0.82	0.93	0.64	0.50	0.07	0.50	0.60	0.90	34.40	34.33	0.40	0.0286	0.0074
Albumin	721	9083	3.89	0.44	3.87	4.00	1.80	3.60	3.90	4.20	6.20	4.40	0.60	0.0074	0.0009
PTINR	6089	3715	0.97	0.16	0.96	1.00	0.60	0.90	1.00	1.00	3.10	2.50	0.10	0.0108	0.0024

a. For each laboratory test, how would you answer the question regarding whether measurements made on patients in the placebo group tend to be worse than those made on patients made on patients in the methotrexate group?

Answer:

First, a comment on the missing data in this study. As it turns out, bilirubin was to be measured most often: biweekly at first, then monthly, then quarterly. Albumin was to be measured monthly, then quarterly. PTINR was only to be measured every 6 months. So by design we would expect different patterns of missing data. For ease of presenting the data, I made rows correspond to weeks and just supplied NA when the measurements would not be indicated. This mechanism explains the overwhelming majority of missing data recorded in the data set.

I note however that there is also missing data that you would have to explore by means other than looking for the "NA"s. Some patients do not have measurements for the "week" that they should have. I did not expect you to look at this. But if you did, you would find that

- **265 patients had week 1 data for all labs.**
- **Only 225 had week 2 data for bilirubin, even though the protocol called for all of them to have such a measurement. But if that timeframe passed due to a missed appointment, we would miss that measurement, and there would be no row in the dataset for that missing data.**
- **We then have week 4 measurements on 255 patients. Thus there would be times that patients missed individual appointments, leading to "interval missing data"—we did have some measurements before and after the missing data, and we can imagine that we might be able to interpolate to "impute" the missing data.**
- **Other times, we are missing data for all scheduled visits after a specific visit. This arises from a variety of mechanisms.**
  - **When patients died or had a liver transplant, we would not be interested in making the measurements.**
  - **And if the patient withdrew consent or were lost to follow-up, we were unable to obtain the measurements even though we were still interested in them.**

We would of course worry that the latter mechanisms for missing data (liver transplant, death, withdrawn consent, loss to follow-up) were "nonignorable" sources of missing data—the missing data might have tended to very different values than the measurements we did have. (The interval missing data might also present a problem if they were missing clinic visits due to illness due to, say, transient exacerbations of liver disease that would not have been captured by interpolating between the measurements we did have—this is not really so much of an issue in this disease).

And there is another mechanism for us to be missing later measurements for some patients: The study ended while the patients were still being followed. We accrued patients over a 4 year period, so there is a variable length of time we would have followed patients. This mechanism for missing data is usually regarded as missing at random (MAR), where missingness depends only on the date of accrual. However, time trends in the types of patients accrued to the study might mean that "imputing" the missing data from other patients is not entirely representative. (Early on in the study, we likely recruited "prevalent" cases, i.e., cases who had PBC for a long period of time. After all those cases were exhausted at each center, we would have had to recruit "incident" cases, i.e., cases who newly met the eligibility criteria.)

A major question then arises about how to treat missing data in descriptive statistics. Not surprisingly, this will depend upon the purpose of your descriptive statistics.

- **A major starting place is to describe how many cases are missing data. This should always be done. And this might include describing the patterns of missing data. For instance, we might describe patterns of missing bilirubin within strata defined by patient sex, age, etc. Key for any clinical trial is to describe it by treatment arm, because we always fear that adverse event profiles cause more patients to withdraw consent. (We should always continue to follow subjects even if they stop taking study drug.)**
- **With continuous data, we do not usually report any statistics that use "imputed" data.**
- **With binary or categorical data, when we report frequencies, we sometimes do different things:**
    - **Most often, we report the frequency of each category among cases without missing data. Hence, the numerator will be cases known to be in a given category, and the denominator will be cases with nonmissing data.**
        - **For purposes of describing the sample, this is now a "conditional frequency", i.e., frequency of the category among those with known values. We will have also given an idea of what proportion of the entire sample has nonmissing data. Imagining a setting in which 100 subjects have 60 missing sex, 15 known males, and 25 known females, wording used to make clear the conditional nature of our descriptive statistics might be: "60% of subjects have missing data for sex. Among the subjects whose sex is known, 37.5% are male and 62.5% are female."**
        - **These descriptive statistics will present information about the entire sample only to the extent that the missing data is "missing completely at random" (MCAR).**
        - **Similarly, these descriptive statistics provide inference about the population from which the population is drawn only to the extent that the missing data is MCAR.**
    - **Sometimes, we treat missing data as its own category alongside the others. Hence for each category, the numerator will be the cases known to be in that category, and the denominator will be the total number of cases.**
        - **For purposes of describing the sample, this treats all categories (including the category of "missing data") the same.**
        - **For purposes of making inference about the population, estimating the proportion of subjects for whom you would be missing data in your study is not usually of scientific value, though if the problem were to be a difficulty of measurement (e.g., inadequate cell count in bone marrow aspirations), that might be relevant for predictive models**

- **For binary data, when frequencies are computed using the three categories of "missing", "nonevent", or "event", the frequencies computed for "event" are the same that would be used if we "imputed" all missing data to be nonevents. While there are times that such might be reasonable for a specific purpose, I do not think that such an interpretation is always appropriate.**
- **In light of the above, when using "missingness" as a category, I would never just report the percentage of, say, males by itself. Instead I use the words "known to be" when describing percentages in this setting. For the same example given above, I might say "15% of subjects are known to be male, 25% of subjects are known to be female, and 60% have sex unknown."**

Now to really answer the question. In the spirit of describing the distribution of "measurements" (as opposed to something related to patients as the unit of interest), I can just ignore the cases for which no measurement was made:

With respect to bilirubin measurements, the placebo group has higher values for the mean, geometric mean, 25th percentile, 75th percentile, and maximum. On the other hand, the methotrexate arm has higher values for the minimum, proportion with "advanced" disease, and proportion with "severe" disease. The modes and medians are equal for the two treatment arms.

With respect to albumin measurements, the methotrexate group has higher values for the mean, geometric mean, minimum, 25th percentile, median, and maximum. On the other hand, the placebo arm has higher values for the proportion with "advanced" disease and proportion with "severe" disease. The modes and 75th percentiles are equal for the two treatment arms.

With respect to PTINR measurements, the placebo group has higher values for the maximum, proportion with "advanced" disease, and proportion with "severe" disease. The means, geometric means, modes, minimums, 25th percentiles, medians, and 75th percentiles are equal for the two treatment arms.

Scientifically, I do not think any of the differences are clinically important. Later we will consider whether the observed differences could merely represent random sampling error. The major points I would make are:

- **In this study, bilirubin, albumin, and PTINR all represent indicators of (at least) subclinical disease. While these measures do tend to be correlated with one another, they are not exactly the same, and thus it is possible to come up with different answers about our scientific question ("Which treatment arm does worse?") depending upon which scientific measurement we choose. We will later discuss that it would be totally inappropriate to wait to see the data to decide which such "clinical endpoint" we use as our measure of treatment effect.**
- **We can obtain different answers depending upon which summary measure we choose. The two treatment arms had identical modes for all three laboratory measures. The other summary statistics were sometimes higher for one arm, and sometimes for the other as we consider the three laboratory measures. Again, it will be totally inappropriate to wait to see the data to see which measure we choose.**
- **As illustrated in the next problem, the degree to which the various summary measures can be influenced by "outliers" is quite different. Later in the course we will also discuss the precision with which we can estimate these different measures: How variable would be the estimates across repetitions of the exact same experiment.**

b. Suppose you were instead interested in answering the question of whether after treatment patients in the placebo group tend to have worse liver disease than patients in the methotrexate group? Discuss the difficulties in answering such a question with these data and the descriptive statistics

you produced above. *(You do not have to answer this question yet, just identify the issues. Note the very careful wording I choose when I talk about "measurements made on patients" in part a versus just referring to "patients" in part b.)*

Answer:

The above descriptive statistics were calculated based on all measurements we obtained in the study. This presents a couple different problems:

- **We have varying numbers of measurements on different patients. Hence in the above descriptive statistics we are counting some patients' data more often than others. This might seem to allow some patients to influence our results than others. In particular, if patients die we will have fewer measurements on them than on the patients who live. Now, if patients' bilirubin levels increase just prior to their death from PBC (this does appear to be the case), then we will have only a few measurements on those patients swamped by the many measurements made on the surviving patients. We are likely more interested in how patients fare, rather than just some technical question about how the measured bilirubin measurements might be distributed. Thus we might prefer getting statistics per patient rather than per measurement. (Note, however, that a clinical laboratory might be interested in knowing what percentage of bilirubin measurements might need special attention due to the extremely high values, in which case the descriptive statistics we presented here would be exactly what they want.)**
- **On each patient, we have measurements before randomization, post randomization while they are taking the study drug, and (often) post randomization after they quit taking study drug for whatever reason. Scientifically we probably want to separate out those different time periods. This is the subject of Homework #3.**

When addressing the statistics per person, missing data is much more problematic than it would be when only addressing the measurements that we actually have. In particular, we need to consider the impact that having more measurements for some subjects than others might have on the "extreme values" of the distribution. If we suppose that two people are identical in every way, except we have more measurements on one of them, the person with more measurements is expected to have a lower minimum and a higher maximum. Thus the patient with more measurements is more likely to be diagnosed as "advanced" or "severe" disease.

Of course, a subject who truly has more severe progression might have more measurements made in order to allow the treating physician to better treat the patient. And getting more measurements of, say, PTINR because bilirubin is higher might lead to

- **"indication bias" on PTINR, in which the greater number of measurements are made precisely because we expect them to be higher (this might predominate when PTINR and bilirubin are highly correlated), or**
- **"ascertainment bias" on PTINR, in which a treatment arm that has higher bilirubin leads to more measurements for PTINR, and thus a greater chance that random high measurements are spuriously considered to be progression of disease (this might predominate when PTINR and bilirubin are less highly correlated).**

3. Suppose you are an unethical researcher who wants to prove that MTX provides a substantial benefit, where "substantial" is thought (by you) to be a tendency for the bilirubin level to be half as high in the MTX arm as it is in the placebo arm.

a. Alter one measurement (tell which case you use by row number and tell how you change that bilirubin measurement) in such a way that would have the mean bilirubin for measurements made on patients from the MTX arm less than half the mean bilirubin measurements made on patients from the placebo arm.

Answer:

By dramatically increasing the bilirubin level for the lowest observed bilirubin in the placebo group from 0.07 to 3,862, the arithmetic means would be 1.64 mg/dl in the placebo group and 0.80 mg/dl in the methotrexate group. (I doubt that a referee would be fooled by this approach.)

b. Alter one measurement (tell which case you use by row number and tell how you change that bilirubin measurement) in such a way that would have the geometric mean bilirubin for measurements made on patients from the MTX arm less than half the geometric mean bilirubin measurements made on patients from the placebo arm.

Answer:

By even more dramatically increasing the bilirubin level for the lowest observed bilirubin in the placebo group from 0.07 to 4.83×10^{1447}, the geometric means would be 1.31 mg/dl in the placebo group and 0.63 mg/dl in the methotrexate group. (I am pretty sure that this bilirubin level in outside the range of biological plausibility.)

c. If possible, alter one measurement (tell which case you use by row number and tell how you change that bilirubin measurement) in such a way that would have the median bilirubin for measurements made on patients from the MTX arm less than half the median bilirubin measurements made on patients from the placebo arm. If it is not possible, explain why not.

Answer:

There are 641 measurements in the placebo group that are equal to the median of 0.6, and 729 measurements in the methotrexate group that are equal to its median of 0.6. We cannot alter the median in either group by changing a single measurement.

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

Answer:

Due to the large sample size, there had to be a very large change to a single measurement in order to cause much of a change in the mean. Nevertheless, it was far easier to change the mean that it was to change the geometric mean, and a single measurement has very limited influence on the median. This follows the same pattern of the sensitivity of these three summary measures to outliers.

Ramifications of this scientifically: If you are interested in the effect some risk factor might have on outlying values, then using the mean is better than the geometric mean, which is better than the median. If you are uninterested in effects that are limited to the frequency of outliers, using the median is better than the geometric mean (which can only be used with positive data), which is better than the mean.

Statistical properties will also often be considered. In the absence of "heavy tails", we generally estimate the mean more precisely than the geometric mean, which we estimate more precisely than the median. In the presence of outliers, we have greatly reduced precision for estimating the mean. If taking the logarithm of the data removes outliers (this is often, but not always, the case), then using the geometric mean will typically be estimated with more precision than the median, providing the data is never zero or negative.

In order to do this problem, you can consider using the data editor to modify a single case (I don't usually recommend this, but in this case it is the fastest way). You may alternatively want to create a variable listing the case number, have the data sorted by the value of bili, list the values in a few rows, replace the values in a single row, and examine the arithmetic and geometric means:

- **g case= *1/_N*** *(_N is a special Stata variable storing the number of cases)*
- **sort** *bili*
- **list** *bili case* **in** *1/10 (will list the cases in the first 10 rows (after any sorting))*
- **replace** *bili= bili + 0.5* **in** *1 (will increase the bilirubin of the case in the first row of the dataset (as currently sorted) by 0.5)*
- **replace** *bili= bili + 0.5* **if** *case==10 (will increase the bilirubin of the case with variable case equal to 10 by 0.5)*
- **means** *bili (will provide arithmetic, geometric, and harmonic means)*