

Biost 517: Applied Biostatistics I

Emerson, Fall 2009

Homework #3 Key

October 30, 2009

A file containing the annotated Stata commands I used to solve this homework is available on the class web pages.

*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.)*

This homework assignment 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")**

(Note that Stata v10 and 11 do require upper case for “MDY”, Stata v9 wanted lower case—go figure.) 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.

Questions for Biost 514 and Biost 517:

1. Generate appropriate descriptive statistics on all relevant variables for all measurements in the dataset according to treatment assignment.
 - 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:

The descriptive statistics are given in the table below.

| | N msng | Mean | Min | 25th %ile | Mdn | 75th %ile | Max |
|-------------------------------|--------|--------------|------|-----------|------|-----------|-------|
| Placebo (n= 4,809) | | | | | | | |
| Bilirubin (mg/dl) | 12 | 0.83 (1.04) | 0.07 | 0.50 | 0.60 | 1.00 | 34.40 |
| Albumin (g/dl) | 355 | 3.88 (0.43) | 1.80 | 3.60 | 3.90 | 4.20 | 5.20 |
| PTINR | 2987 | 0.97 (0.17) | 0.60 | 0.90 | 1.00 | 1.00 | 3.10 |
| FVC (l) | 3917 | 3.17 (0.73) | 1.40 | 2.69 | 3.04 | 3.55 | 6.29 |
| FVC % predicted | 3917 | 97.5 (14.9) | 52 | 88 | 97 | 108 | 150 |
| DLCO (ml / min / mmHg) | 3925 | 18.9 (5.1) | 7.1 | 15.9 | 18.3 | 21.3 | 49.3 |
| DLCO % predicted | 3925 | 81.8 (17.4) | 26 | 70 | 81 | 92 | 139 |
| Methotrexate (n=4,995) | | | | | | | |
| Bilirubin (mg/dl) | 11 | 0.80 (0.81) | 0.10 | 0.40 | 0.60 | 0.80 | 13.50 |
| Albumin (g/dl) | 366 | 3.91 (0.44) | 1.90 | 3.70 | 4.00 | 4.20 | 6.20 |
| PTINR | 3102 | 0.97 (0.14) | 0.60 | 0.90 | 1.00 | 1.00 | 2.10 |
| FVC (l) | 4078 | 3.32 (0.72) | 1.66 | 2.83 | 3.27 | 3.68 | 6.48 |
| FVC % predicted | 4078 | 101.2 (14.6) | 56 | 93 | 103 | 111 | 150 |
| DLCO (ml / min / mmHg) | 4082 | 19.6 (5.0) | 7.0 | 16.2 | 19.3 | 22.4 | 39.8 |
| DLCO % predicted | 4082 | 82.7 (17.3) | 25 | 72 | 82 | 92 | 159 |
| All patients (n=9804) | | | | | | | |
| Bilirubin (mg/dl) | 23 | 0.82 (0.93) | 0.07 | 0.50 | 0.60 | 0.90 | 34.40 |
| Albumin (g/dl) | 721 | 3.89 (0.44) | 1.80 | 3.60 | 3.90 | 4.20 | 6.20 |
| PTINR | 6089 | 0.97 (0.16) | 0.60 | 0.90 | 1.00 | 1.00 | 3.10 |
| FVC (l) | 7995 | 3.25 (0.73) | 1.40 | 2.74 | 3.16 | 3.64 | 6.48 |
| FVC % predicted | 7995 | 99.4 (14.9) | 52 | 89 | 100 | 110 | 150 |
| DLCO (ml / min / mmHg) | 8007 | 19.3 (5.0) | 7.0 | 16.0 | 18.8 | 21.9 | 49.3 |
| DLCO % predicted | 8007 | 82.2 (17.4) | 25 | 71 | 82 | 92 | 159 |

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. FVC and DLCO were to be measured annually. 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.

Now to really answer the question:

The bilirubin, albumin, and PTINR measurements represent severity of liver disease in this data set.

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

With respect to albumin measurements, the methotrexate group has higher values for the mean, minimum, 25th percentile, median, and maximum. The 75th percentiles are equal for the two treatment arms.

With respect to PTINR measurements, the placebo group has higher values for the maximum. The means, minimums, 25th percentiles, medians, and 75th percentiles are equal for the two treatment arms.

The FVC and DLCO measurements were obtained to examine potential lung toxicity of the treatment. The FVC, FVC percent predicted, and DLCO percent predicted descriptive statistics were lower for the placebo group than the methotrexate group; the DLCO measurements showed a mixed pattern.

Scientifically, I do not think any of the differences for the liver or lung measurements 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, though that data is not shown here), 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 the pre-randomization and post-randomization measurements. We may, for descriptive and exploratory purposes, also want to consider what happens when patients stop taking their assigned treatment. These types of analyses are the subject of the remaining questions.**
2. In problem #1, you generated descriptive statistics using all measurements in the dataset. However, multiple measurements were made on each subject. This problem guides you through the process of using Stata to determine how many repeat measurements are made on each individual.

The data file contains repeated measurements on each individual. When our interest is on how patients fare, we often combine such repeated measurements into a single summary. For instance, we might consider taking the average of the measurements, the maximum or minimum of the measurements, or only the last measurement. Stata provides a command “egen” that will allow us to easily abstract such summaries by patient.

For instance, suppose we want the mean bilirubin for each patient. We can obtain a variable *mbili* that will contain that by:

▪ `egen mnbili = mean(bili), by(ptid)`

Each row will now have a value for variable *mnbili* that is equal to the mean of all the bilirubin values for that patient. If you wanted to have instead the mean of bilirubin measurements made after randomization (so after week 1) you could use:

▪ `egen mnbili = mean(bili) if week > 1, by(ptid)`

After this command, you would have a variable that had missing values for any rows corresponding to week 1, and for all other rows, the value for variable *mnbili* would be equal to the mean of all bilirubin values made after week 1 for that patient.

In this and the following problems you will need to use “egen” repeatedly in order to be able to perform analyses on a per patient rather than per measurement basis.

- a. Use “egen” to generate a variable *nbili* counting the number of non-missing bilirubin measurements made for each individual, and provide suitable descriptive statistics for this variable using all cases in the datafile. The following Stata code will generate the variable:

```
egen nbili= count(bili) if bili!=., by(ptid)
```

How many measurements in the datafile correspond to a patient having a maximum of 6 nonmissing bilirubin measurements? How many patients does this represent? You might consider either or both of the following Stata commands:

```
table nbili
list ptid bili nbili if nbili==6
```

Answer:

There are 18 measurements in the dataset which correspond to the 3 patients having 6 measurements each.

- b. As can be seen in part a, doing descriptive statistics on the summarized variable is still complicated due to the number of repeated measurements on each individual. If we want to find out the distribution of *nbili* across patients (rather than rows in the file), we will need to restrict our analysis to one row for each patient. In this clinical trial, you might think that every subject should have had a week 1 measurement. We can check that by considering the minimum value of *week* for each individual. Generate a variable *minweek* containing the earliest week for which a subject has a row in the data set, and provide summary statistics to show that each subject has a week 1 measurement. The following Stata code can be used to generate *minweek*:

```
egen minweek=min(week), by(ptid)
```

Answer:

Descriptive statistics for variable *minweek*.

| | N | Mean | SD | Min | 25 th %ile | Mdn | 75 th %ile | Max |
|---------------|------|------|------|-----|-----------------------|-----|-----------------------|-----|
| Combined Arms | 9804 | 1.00 | 0.00 | 1 | 1 | 1 | 1 | 1 |

There were no missing observations for *minweek*. The mean value was 1.0, and the standard deviation was 0. If the standard deviation is 0, that must mean that all measurements are exactly equal to the mean. Similarly, the minimum value was 1, and the maximum value was 1. This also signifies that all values were equal to 1,

- c. Now, since we know that every individual has a row corresponding to week 1, when we desire statistics on each patient, we could obtain summary statistics just for rows corresponding to $week=1$. Describe the distribution of the number of measurements made on each subject. Provide descriptive statistics that allow us to compare the number of measurements per patient by treatment group. What might be the scientific importance of any differences between treatment groups? What might be the statistical ramifications of any differences? Are there differences that concern you?

Answer:

Descriptive statistics for the number of observations available for each patient.

| | N | Mean | SD | Min | 25 th %ile | Mdn | 75 th %ile | Max |
|---------------|-----|-------|-------|-----|-----------------------|-----|-----------------------|-----|
| Placebo | 133 | 36.07 | 10.64 | 3 | 31 | 37 | 44 | 52 |
| Methotrexate | 132 | 37.76 | 9.89 | 6 | 33 | 40 | 45 | 52 |
| Combined Arms | 265 | 36.91 | 10.29 | 3 | 32 | 38 | 45 | 52 |

Scientifically, if there were different numbers of observations on patients across treatment groups, this might signify higher rates of liver transplantation or deaths. After either of these events, measurements of bilirubin are not of interest to us. Similarly, different numbers of measurements might signify differential study dropout across treatment arms. Patients might drop out of the study because of adverse events or because they were feeling so well that they embarked on extended vacations. Either of these might be cause to worry that the missing data was informative about the treatment's effect on the patients' health.

Statistically, the number of measurements affect the precision with which we estimate trends in the bilirubin. Furthermore, if our interest is in the minimum or maximum bilirubin measured over time, the sample size has great impact on the aspect of the distribution being estimated (the highest of four measurements tends toward the 80th percentile, the highest of 49 measurements tends toward the 98th percentile).

In examining the above distributions, patients in the methotrexate group average two more measurements than patients in the placebo group. This is probably not enough of a difference to indicate substantial scientific issues (though it is of course possible that people are dropping out of the placebo group due to illness and out of the methotrexate group to go on vacation, or vice versa). Similarly, a difference in 2 measurements when the average is 40 does not likely cause a big problem statistically.

- d. (There is nothing to answer in this part, it is purely informational.) An alternative approach to find a unique row for each patient is to use the "tag" function in "egen", which will tag a unique row for each ptid. The following Stata code can be used to generate variable *tag*, and then obtain descriptive statistics for *nbili* on a per patient basis:
- ```
egen tag=tag(ptid)
tabstats nbili if tag, stat(n mean ... max)
```
3. Generate a variable *mbili* reflecting the average of all bilirubin measurements made for each individual (both before and after randomization).
- Provide summary statistics for both *bili* and *mbili* for the two treatment groups using all available data in the data set. What scientific question could be addressed using these descriptive statistics?

**Answer:**

**Descriptive statistics on all bilirubin measurements made on subjects in the clinical trial.**

|               | N    | Mean | SD   | Min  | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max   |
|---------------|------|------|------|------|-----------------------|------|-----------------------|-------|
| Placebo       | 4797 | 0.83 | 1.04 | 0.07 | 0.50                  | 0.60 | 1.00                  | 34.40 |
| Methotrexate  | 4984 | 0.80 | 0.81 | 0.10 | 0.40                  | 0.60 | 0.80                  | 13.50 |
| Combined Arms | 9781 | 0.82 | 0.93 | 0.07 | 0.50                  | 0.60 | 0.90                  | 34.40 |

**This analysis could be appropriate if the scientific question were about the typical sorts of bilirubin measurements that would be made in such a clinical trial. For instance, if high levels of bilirubin required special processing to avoid saturation of the measurement process, a clinical laboratory might want to know about the distribution of all measurements, and the lab would likely care little about patient specific statistics (except as it might affect inference). I can probably even make up a scenario in which the lab would care about whether the measurements were made on subjects receiving placebo or MTX. In any case, I am not impressed by any differences in the distribution of those measurements, with the exception of the maximum value.**

**Descriptive statistics on the patient specific mean bilirubin measurements made on subjects in the clinical trial, when patients are represented in proportion to the number of measurements made on them.**

|               | N    | Mean | SD   | Min  | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max  |
|---------------|------|------|------|------|-----------------------|------|-----------------------|------|
| Placebo       | 4809 | 0.83 | 0.64 | 0.32 | 0.48                  | 0.64 | 0.94                  | 5.04 |
| Methotrexate  | 4995 | 0.80 | 0.62 | 0.28 | 0.48                  | 0.59 | 0.82                  | 4.69 |
| Combined Arms | 9804 | 0.82 | 0.63 | 0.28 | 0.48                  | 0.61 | 0.89                  | 5.04 |

**I have a bit harder time trying to explain why I might ever be interested in these descriptive statistics. Why might I want patient specific means, but then have multiple representations of those means in my analysis. In this analysis I am claiming I am more interested in the patient specific mean, and I am also saying that the patients who had more measurements are in some way more important. While there are some settings that might mimic this sort of an analysis in some small way, I think this analysis has no scientific value. For what it is worth, we do note that while the mean is similar to that obtained in the immediately preceding table, the standard deviation is less in this latter table and the minima and maxima are less extreme. Both of these come from taking the mean of multiple observations.**

- b. Provide summary statistics for *mbili* for the two treatment groups when each patient is represented only once. What scientific question could be addressed using these descriptive statistics?

**Answer:**

**Descriptive statistics on the patient specific mean bilirubin measurements made on subjects in the clinical trial, when each patient is represented only once.**

|               | N   | Mean | SD   | Min  | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max  |
|---------------|-----|------|------|------|-----------------------|------|-----------------------|------|
| Placebo       | 133 | 0.87 | 0.67 | 0.32 | 0.49                  | 0.65 | 0.98                  | 5.04 |
| Methotrexate  | 132 | 0.83 | 0.65 | 0.28 | 0.48                  | 0.59 | 0.88                  | 4.69 |
| Combined Arms | 265 | 0.85 | 0.66 | 0.28 | 0.48                  | 0.63 | 0.96                  | 5.04 |

By using each patient only once, we now are describing the behavior of patients, rather than measurements. However, as we are combining measurements made both before and after randomization, these statistics are probably of interest only if we absolutely knew that treatment with methotrexate had no effect whatsoever (no effect on bilirubin or anything that affected the number of observations we might have). And if we knew all that, why would we want statistics broken down by treatment group.

At least we do see that there are not huge differences in the distribution reflected in these summary statistics. That is, I am not impressed by differences of 0.04 mg/dl in the mean bilirubin, nor even by a difference of 0.35 in the maximum bilirubin when the measurements are as large as 4.69.

4. In problem 3, you took the mean of all bilirubin measurements for an individual—both before and after randomization. The following code will create a variable *mtrtbili* which will be the mean of bilirubin measurements made post randomization. (Note the need to ensure that the first row for each patient, or the “tagged” case if you use that approach, will not have a missing value for *mtrtbili*.):

```
egen grbg=mean(bili) if week>1, by(ptid)
egen mtrtbili=mean(grbg), by(ptid)
```

- a. Provide descriptive statistics which compare the treatment groups with respect to the patient specific mean bilirubin post randomization. Based on these statistics, do you worry about any outliers in the data? Explain.

**Answer:**

**Descriptive statistics on the patient specific mean bilirubin measurements made on subjects in the clinical trial post randomization. The distributions seem remarkably similar between the two treatment groups. There does seem to be evidence suggestive of outlying values:**

- The mean is a bit larger than the median,
- the SD is much larger than one-half to one-third the mean on these positive measurements,
- the median is not midway between the minimum and maximum, and
- the minimum is about half a SD below the median, while the maximum is 6 SD above the median.

|               | N   | Mean | SD   | Min  | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max  |
|---------------|-----|------|------|------|-----------------------|------|-----------------------|------|
| Placebo       | 133 | 0.88 | 0.69 | 0.32 | 0.49                  | 0.65 | 1.00                  | 5.16 |
| Methotrexate  | 132 | 0.83 | 0.66 | 0.28 | 0.48                  | 0.59 | 0.88                  | 4.79 |
| Combined Arms | 265 | 0.86 | 0.68 | 0.28 | 0.48                  | 0.63 | 0.96                  | 5.16 |

- b. Provide descriptive statistics which compare the treatment groups with respect to the difference between the patient specific mean bilirubin post randomization and the patient's bilirubin at randomization (week 1). (Note that for the case representing week 1, the difference  $mtrtbili - bili$  is the value we are interested in.)

**Answer:**

**Descriptive statistics on the difference between the patient specific mean bilirubin measurements made on subjects post randomization and the patient's baseline bilirubin measurement made prior to randomization. Again, I do not see impressive differences in the distributions across treatment groups. It is of interest to note that the mean difference is greater than 0 for both groups. This is to be expected, as we are following a disease that tends to progress over time. This is why it is important to have a control group. I did not ask you to comment on the possibility of outliers, but I think we do have some suggestion of outliers:**

- **The mean is a bit larger than the median,**
- **the median is not midway between the minimum and maximum, and**
- **the minimum is about 1.5 SD below the median, while the maximum is 6 SD above the median.**

**(Note that I could not compare the mean and SD, because these measurements can be both positive and negative, and I am unsure of the differences that are biologically impossible.)**

|                      | N   | Mean | SD   | Min   | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max  |
|----------------------|-----|------|------|-------|-----------------------|------|-----------------------|------|
| <b>Placebo</b>       | 133 | 0.16 | 0.57 | -0.65 | -0.09                 | 0.04 | 0.23                  | 3.96 |
| <b>Methotrexate</b>  | 132 | 0.16 | 0.44 | -0.67 | -0.03                 | 0.09 | 0.22                  | 2.99 |
| <b>Combined Arms</b> | 265 | 0.16 | 0.51 | -0.67 | -0.06                 | 0.06 | 0.22                  | 3.96 |

- c. Create a new variable *mdrgbili* representing the mean bilirubin for each patient while taking study drug (methotrexate or placebo), and repeat parts (a) and (b) for this measure of treatment outcome.

**Answer:**

**Descriptive statistics on the patient specific mean bilirubin measurements made on subjects in the clinical trial while taking the study drug (methotrexate or placebo), when each patient is represented only once.**

|                      | N   | Mean | SD   | Min  | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max  |
|----------------------|-----|------|------|------|-----------------------|------|-----------------------|------|
| <b>Placebo</b>       | 133 | 0.81 | 0.53 | 0.21 | 0.48                  | 0.64 | 0.96                  | 3.28 |
| <b>Methotrexate</b>  | 132 | 0.77 | 0.54 | 0.27 | 0.48                  | 0.57 | 0.82                  | 3.67 |
| <b>Combined Arms</b> | 265 | 0.79 | 0.54 | 0.21 | 0.48                  | 0.61 | 0.93                  | 3.67 |

**Descriptive statistics on the difference between the patient specific mean bilirubin measurements made on subjects in the clinical trial while taking the study drug (methotrexate or placebo) and the baseline bilirubin measurement for each patient prior to randomization.**

|                      | N   | Mean | SD   | Min   | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max  |
|----------------------|-----|------|------|-------|-----------------------|------|-----------------------|------|
| <b>Placebo</b>       | 133 | 0.10 | 0.40 | -0.65 | -0.09                 | 0.03 | 0.20                  | 2.52 |
| <b>Methotrexate</b>  | 132 | 0.09 | 0.32 | -0.65 | -0.05                 | 0.06 | 0.17                  | 1.87 |
| <b>Combined Arms</b> | 265 | 0.10 | 0.36 | -0.65 | -0.08                 | 0.05 | 0.17                  | 2.52 |

**In both cases, I find that the distribution of measurements is quite similar across treatment groups. Arguably, even if you were to tell me the difference was beyond what could be explained by sampling error, I would not think the difference was clinically important.**

- d. Which of these analyses are scientifically useful in assessing the effect of methotrexate on bilirubin levels? Why? What are their relative advantages and disadvantages?

**Answer:**

**We need to consider the two variants we explored: Post randomization vs On study drug and Absolute measurement vs Change in measurement.**

**The scientific difference between post-randomization and on-study drug revolves around the question of why subjects might stop taking their drug: progression of disease, incidence of adverse events, they got tired of the clinical trial, or the study ended. Many researchers focus only on the “tired of the clinical trial” and thus propose a “per protocol” analysis of the data. They imagine they are telling us what would happen if you could force everyone to take their drug. Personally, I always worry about the progression of disease and adverse events scenarios. In progression of disease you may have identified people for whom the treatment does not work and forcing them to take the drug would not make them like the other people. In the case of adverse events, forcing a patient who gets a headache to continue on the drug might make them progress to a cerebral hemorrhage (okay, a little dramatic hyperbole, but the point is there). In any case, we probably have higher compliance with the treatment on the clinical trial than we would in real life (an unsubstantiated claim, but an accurate reflection of my belief). Hence, clinical trial standards call for the “intent to treat” analysis based on post randomization summaries regardless of compliance with the protocol. (I will note that in this data set, we did take some measurements after stopping the study for all subjects. I did analyses which both included and excluded the measurements made after study end. This was because there might have been some residual effect of the drug even after the patients stopped taking it.)**

**I should note that when analyzing adverse event data, there is a greater tendency to do the “per protocol” analysis, but even then we include measurements made up to 30 days after stopping the study drug.**

**Now, what about the advantage of comparing the measurements post randomization versus the change in measurements from baseline. A surprising result (one which we will come back to full force in Biost 518) is that in the setting of a randomized clinical trial you can often do worse by analyzing the difference in measurements than just analyzing the follow-up values. This is a very hard thing to convince statistical novices of. But it is true. It all depends on how highly correlated the baseline values are with the follow-up measurements within each treatment group. In this case, there is a reasonably high correlation of 0.66 between *mtrtbili* and *bili* week 1. When the correlation is greater than 0.5, the difference is more precise than using the follow-up measurements by themselves. (The best approach to “adjusting for baseline values” is to adjust for baseline as a predictor in a regression model, so that is why we discuss it in Biost 518.)**

**So if I knew all this in advance: Because bilirubin is a measure of efficacy of the treatment, I would definitely use post randomization instead of on study drug, and I would use the difference from baseline instead of just the follow-up.**

**(I obviously did not expect you to know all this, only to think about it.)**

5. Now suppose we consider a treatment outcome based on the maximum bilirubin for each patient, instead of the mean. The following code will create a variable *mxttbili* which will be the maximum of bilirubin measurements made post randomization. (Note the need to ensure that the first row for each patient, or the “tagged” case if you use that approach, will not have a missing value for *mxttbili*):

```
egen grbg=max(bili) if week>1, by(ptid)
egen mxttbili=mean(grbg), by(ptid)
```

- a. Provide descriptive statistics which compare the treatment groups with respect to the patient specific maximum bilirubin post randomization. Based on these statistics, do you worry about any outliers in the data? Explain.

**Answer:**

**Descriptive statistics on the patient specific maximum bilirubin measurements made on subjects in the clinical trial post randomization. The distributions seem remarkably similar between the two treatment groups, with the possible exception of the means and maxima. The difference in means of 0.3 is of questionable clinical importance, and it is undoubtedly influenced greatly by the obvious outlier: There does seem to be even more evidence suggestive of outlying values than there was when using patient specific means:**

- The mean is a bit larger than the median,
- the SD is much larger than the mean on these positive measurements,
- the median is not midway between the minimum and maximum, and
- the minimum is about one-fourth a SD below the median, while the maximum is 11 SD above the median.

|               | N   | Mean | SD   | Min  | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max   |
|---------------|-----|------|------|------|-----------------------|------|-----------------------|-------|
| Placebo       | 133 | 2.01 | 3.80 | 0.50 | 0.80                  | 1.10 | 1.70                  | 34.40 |
| Methotrexate  | 132 | 1.70 | 1.87 | 0.30 | 0.80                  | 1.00 | 1.70                  | 13.50 |
| Combined Arms | 265 | 1.85 | 3.00 | 0.30 | 0.80                  | 1.00 | 1.70                  | 34.40 |

- b. Provide descriptive statistics which compare the treatment groups with respect to the difference between the patient specific maximum bilirubin post randomization and the patient’s bilirubin at randomization (week 1). (Note that for the case representing week 1, the difference  $mxttbili - bili$  is the value we are interested in.)

**Answer:**

**Descriptive statistics on the difference between the patient specific maximum bilirubin measurements made on subjects post randomization and the patient’s baseline bilirubin measurement made prior to randomization. Again, I do not see impressive differences in the distributions across treatment groups except in the statistics too heavily influenced by the outliers.**

|               | N   | Mean | SD   | Min   | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max   |
|---------------|-----|------|------|-------|-----------------------|------|-----------------------|-------|
| Placebo       | 133 | 1.29 | 3.73 | -0.30 | 0.20                  | 0.40 | 0.80                  | 33.20 |
| Methotrexate  | 132 | 1.02 | 1.65 | -0.20 | 0.30                  | 0.50 | 0.85                  | 11.70 |
| Combined Arms | 265 | 1.16 | 2.88 | -0.30 | 0.20                  | 0.50 | 0.80                  | 33.20 |

- c. Create a new variable *mxdrgbili* representing the maximum bilirubin for each patient while taking study drug (methotrexate or placebo), and repeat parts (a) and (b) for this measure of treatment outcome.

**Answer:**

**Descriptive statistics on the patient specific maximum bilirubin measurements made on subjects in the clinical trial while taking the study drug (methotrexate or placebo), when each patient is represented only once.**

|               | N   | Mean | SD   | Min  | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max   |
|---------------|-----|------|------|------|-----------------------|------|-----------------------|-------|
| Placebo       | 133 | 1.54 | 1.94 | 0.30 | 0.80                  | 1.00 | 1.50                  | 16.40 |
| Methotrexate  | 132 | 1.43 | 1.29 | 0.30 | 0.80                  | 0.90 | 1.60                  | 8.60  |
| Combined Arms | 265 | 1.48 | 1.65 | 0.30 | 0.80                  | 1.00 | 1.50                  | 16.40 |

**Descriptive statistics on the difference between the patient specific maximum bilirubin measurements made on subjects in the clinical trial while taking the study drug (methotrexate or placebo) and the baseline bilirubin measurement for each patient prior to randomization.**

|               | N   | Mean | SD   | Min   | 25 <sup>th</sup> %ile | Mdn  | 75 <sup>th</sup> %ile | Max   |
|---------------|-----|------|------|-------|-----------------------|------|-----------------------|-------|
| Placebo       | 133 | 0.82 | 1.86 | -0.30 | 0.20                  | 0.40 | 0.70                  | 15.70 |
| Methotrexate  | 132 | 0.76 | 1.08 | -0.20 | 0.30                  | 0.40 | 0.70                  | 7.00  |
| Combined Arms | 265 | 0.79 | 1.52 | -0.30 | 0.20                  | 0.40 | 0.70                  | 15.70 |

**In both cases, I find that the distribution of measurements is quite similar across treatment groups, allowing for the tendency to outliers. I do note that the outliers are not as extreme, which might be explained by the patient with most progressive disease dropping off the study drug prior to their last available measurements.**

- d. Which of these analyses are scientifically useful in assessing the effect of methotrexate on bilirubin levels? Why? What are their relative advantages and disadvantages?

**Answer:**

**All of the same issues arise here that were discussed in my answer to problem 3(d). I do note that the correlation between the patient specific maximum and the baseline value is only 0.33 in this dataset. If that were to be the true correlation, taking the difference from baseline is the WRONG thing to do. Because that correlation is less than 0.5, we would actually lose precision in estimating the effect of treatment. (Again, multiple regression will come to the rescue next quarter.)**

- e. How additional problem might be posed by using the maximum rather than the mean as was used in problem 3?

**Answer:**

**When using the maximum, we have to worry about the statistical issues related to sample size when using extrema (minima or maxima). Even when there is no true difference in distributions, a group that has more measurements per patient will also tend to have more extreme patient specific minima and maxima. (In this particular trial, the distribution of**

**number of measurements was similar between the groups, so this is probably not such a big issue.)**