Biost 517: Applied Biostatistics I
Emerson, Fall 2007

Homework #3 Key
October 27, 2007

Written problems: To be handed in at the beginning of class on Wednesday, October 17, 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.)*

The following problems make use of the polyamine data in the DFMO clinical trial (I would suggest you use the DFMO-long dataset dfmolong.txt). All of the following variable names refer to the definitions in that file. The data can be input into Stata using the command

```
infile ptid time female age dose put spd spm using dfmolong.txt
```

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 spermine for each patient. We can obtain a variable *mnspm* that will contain that by:

- `egen mnspm = mean(spm), by(ptid)`

Each row will now have a value for variable *mnspm* that is equal to the mean of all the spermine values for that patient. If you wanted to have instead the mean of spermine measurements made after randomization (so after time 0) you could use:

- `egen mnspm = mean(spm) if time > 0, by(ptid)`

After this command, you would have a variable that had missing values for any rows corresponding to month 0, and for all other rows, the value for variable *mnspm* would be equal to the mean of all spermine values made after time 0 for that patient.

In 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.

1. Consider first a naïve approach to analyzing this data. Provide descriptive statistics for putrescine, spermidine, and spermine values by dose group while disregarding the fact that multiple measurements might be made on each subject.

Ans: The following table presents descriptive statistics for each of the dose groups as well as for the entire sample. *(Note that I presented the descriptive statistics in order to facilitate comparison of each measurement across dose groups. This would seem logical, when we consider each polyamine individually. If my major interest had been to describe how DFMO affects the relative concentration of the three polyamines, I would have presented all three polyamines for each dose group.)*

Table: Descriptive statistics for all samples by dose group irrespective of time on study

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Putrescine (nmol /mg protein)								
0.000	117	1.09	1.29	0.00	0.48	0.72	1.13	9.14
0.075	109	0.84	0.78	0.00	0.38	0.60	0.96	4.28
0.200	90	0.71	0.65	0.00	0.27	0.51	0.93	3.21
0.400	91	0.72	0.89	0.00	0.12	0.53	0.87	5.48
All	407	0.85	0.96	0.00	0.34	0.60	0.96	9.14
Spermidine (nmol /mg protein)								
0.000	117	3.16	1.35	1.01	2.17	2.84	3.95	7.05
0.075	109	3.00	1.17	0.00	2.24	2.82	3.63	7.02
0.200	90	2.92	1.36	0.29	1.92	2.62	3.82	7.84
0.400	91	2.80	1.49	0.00	1.86	2.39	3.28	7.60
All	407	2.98	1.34	0.00	2.04	2.68	3.64	7.84
Spermine (nmol /mg protein)								
0.000	117	7.17	3.86	1.46	4.88	6.47	8.69	35.55
0.075	109	7.76	3.96	0.00	5.59	7.43	9.19	37.67
0.200	90	7.43	4.60	1.93	4.82	7.13	8.89	41.68
0.400	91	7.27	4.09	0.00	4.85	6.56	8.73	34.04
All	407	7.41	4.10	0.00	5.01	6.86	8.89	41.68

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.
    a. Use "egen" to generate variables *nput, nspd, nspm* counting, respectively, the number of non-missing putrescine, spermidine, and spermine measurements made for each individual, and provide suitable descriptive statistics for this variable using all cases in the datafile.

Ans: The following table presents the counts for each dose group. The annotated Stata file presents analyses illustrating that measurements were available for putrescine, spermidine, and spermine for each sample. It should be noted that this data is misleading, because a subject with four measurements would be counted four times as having four measurements, while a subject with three measurements would be counted three times as having three measurments, and so on.

	Number of Samples				
Dose	**1**	**2**	**3**	**4**	**Total**
0.000	1	4	12	100	117
0.075	1	4		104	109
0.200	1	4	9	76	90
0.400	3	8	12	68	91
Total	6	20	33	348	407

    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 *nput, nspd, nspm* 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 month 0 measurement. We can check that by considering the minimum value of *time* for each individual. Generate a variable *mintime* containing the earliest time for which a subject has a row in the data set, and provide summary statistics to show that each subject has a time 0 measurement. The following Stata code can be used to generate *mintime*:

```
egen mintime=min(time), by(ptid)
```

Ans: The following table presents the number of cases corresponding to a patient with the minimum time as shown. Note that this agrees with the total number of cases for each dose group as shown above.

	Minimum Time
Dose	**0**
0.000	117
0.075	109
0.200	90
0.400	91
Total	407

c. Now, since we know that every individual has a row corresponding to time 0, when we desire statistics on each patient, we could obtain summary statistics just for rows corresponding to *time==0*. 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?

Ans: The following table presents the counts per patient for each dose group. There is some variation among doses with respect to the patients who were missing measurements at some follow-up times. The highest dose group has the highest rate of missing measurements. This might make us worry about toxicity leading to patient drop out. We also see some drop out in the placebo group. Had this been a treatment trial, we might worry that patients were dropping out because they were not receiving a benefit of the treatment. In either case, we need to worry that such missing data might be "nonignorable". That is, we need to worry that the missing measurements would have been substantially different than those for the patients who continued on the study.

	Number of Samples				
Dose	**1**	**2**	**3**	**4**	**Total**
0.000	1	2	4	25	32
0.075	1	2		26	29
0.200	1	2	3	19	25
0.400	3	4	4	17	28
Total	6	10	11	87	114

3. Generate variables *mnput, mnspd, mnspm* reflecting the average of all polyamine measurements made for each individual (both before and after randomization).

a. Provide summary statistics for both *mnput, mnspd, mnspm* for the treatment groups using all available data in the data set. What scientific question could be addressed using these descriptive statistics?

Ans: The following table presents descriptive statistics for the patient specific mean polyamine values for each of the dose groups as well as for the entire sample. Subjects with more measurements are represented more heavily in this analysis, because the mean value for each patient was repeated as many times as that patient had biopsies. Conceivably, this could represent the distribution of measurements the laboratory would have to be prepared to report (e.g., informing the lab the range of patient specific means that would actually be measured), it is unlikely that this is of very much interest to the cancer prevention researchers.

It is of interest to note that the means in this table agree with the means in problem 1. The SD and the minima and maxima are less extreme in this table, however, because we have reduced the variability of the measurements by taking patient specific means. This is something that holds in general: Means of several measurements are less variable than were the original measurements.

Table: Descriptive statistics for all samples by dose group irrespective of time on study

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Putrescine (nmol /mg protein)								
0.000	117	1.09	0.66	0.34	0.53	0.99	1.25	3.24
0.075	109	0.84	0.39	0.11	0.54	0.73	1.23	1.75
0.200	90	0.71	0.32	0.12	0.44	0.69	0.95	1.28
0.400	91	0.72	0.52	0.04	0.36	0.53	0.94	1.98
All	407	0.85	0.52	0.04	0.47	0.75	1.12	3.24
Spermidine (nmol /mg protein)								
0.000	117	3.16	0.88	1.87	2.62	2.93	3.75	5.26
0.075	109	3.00	0.70	1.95	2.52	2.98	3.47	6.62
0.200	90	2.92	0.91	1.59	2.43	2.73	3.18	6.86
0.400	91	2.80	0.78	1.54	2.11	2.66	3.19	5.21
All	407	2.98	0.83	1.54	2.42	2.92	3.34	6.86
Spermine (nmol /mg protein)								
0.000	117	7.17	2.30	4.68	5.81	6.79	7.81	16.43
0.075	109	7.76	2.04	5.42	6.28	7.63	8.56	15.81
0.200	90	7.43	2.28	5.77	5.96	6.90	7.97	16.56
0.400	91	7.27	1.91	4.17	5.90	7.06	8.44	13.70
All	407	7.41	2.15	4.17	5.91	7.01	8.34	16.56

b. Provide summary statistics for *mnput, mnspd, mnspm* for the treatment groups when each patient is represented only once. What scientific question could be addressed using these descriptive statistics?

Ans: The following table presents descriptive statistics for the mean polyamine value for each patient in a way that treats all patients equally. There are still problems with this analysis, because in collapsing across all times, we are not considering the role that DFMO treatment might have on the polyamine measurements. We are also not considering the possibility that patients with missing values for some biopsies might represent a very

different subpopulation (this latter problem is not addressed by any of the analyses in this homework).

Table: Descriptive statistics for patient specific mean polyamine levels by dose group irrespective of time on study

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Putrescine (nmol /mg protein)								
0.000	32	1.06	0.65	0.34	0.59	0.99	1.19	3.24
0.075	29	0.80	0.41	0.11	0.50	0.72	1.13	1.75
0.200	25	0.70	0.32	0.12	0.44	0.67	0.94	1.28
0.400	28	0.71	0.54	0.04	0.34	0.52	0.96	1.98
All	114	0.83	0.52	0.04	0.44	0.72	1.10	3.24
Spermidine (nmol /mg protein)								
0.000	32	3.14	0.89	1.87	2.57	2.91	3.61	5.26
0.075	29	3.09	0.92	1.95	2.52	2.98	3.63	6.62
0.200	25	3.01	1.05	1.59	2.49	2.84	3.18	6.86
0.400	28	2.76	0.83	1.54	2.10	2.59	3.18	5.21
All	114	3.01	0.92	1.54	2.41	2.90	3.34	6.86
Spermine (nmol /mg protein)								
0.000	32	7.09	2.26	4.68	5.81	6.76	7.71	16.43
0.075	29	7.79	2.02	5.42	6.54	7.63	8.56	15.81
0.200	25	7.44	2.25	5.77	6.11	6.90	7.97	16.56
0.400	28	7.20	1.89	4.17	5.80	7.19	8.39	13.70
All	114	7.37	2.10	4.17	5.91	7.00	8.34	16.56

4. In problem 3, you took the mean of all polyamine measurements for an individual—both before and after randomization. The following code will create a variable *mtrtspm* which will be the mean of spermidine 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 *mtrtspm*.):

```
egen grbg=mean(spm) if time>0, by(ptid)
egen mtrtspm=mean(grbg), by(ptid)
```

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

<u>Ans:</u> The following table presents descriptive statistics for the mean polyamine value post randomization for each patient in a way that treats all patients equally. There are still problems with this analysis, because in collapsing across all times, we are not considering the role that DFMO treatment might have on the polyamine measurements over time: Perhaps there is a steadily increasing effect over time, or perhaps the effect of DFMO wears off soon after treatment is stopped. These problems might not create such a problem if we had equal numbers of measurements on each subject, but because we are lacking more measurements on the highest dose group, that group has proportionately more measurements made at months 6 and 12 (while on DFMO) than at month 15 (after DFMO stopped). We are also not considering the possibility that patients with missing values for some biopsies might represent a very different subpopulation (this latter problem is not addressed by any of the analyses in this homework).

From the table, we see that the standard deviations of these positive valued measurements do tend to be relatively large compared to the mean. The mean is also not the midpoint of the range (the maximum tends to be a little further from the mean than is the minimum). So there does seem to be a little skewness, though I am not struck by any evidence for very extreme outliers.

Table: Descriptive statistics for patient specific mean polyamine levels by dose group post randomization

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Putrescine (nmol /mg protein)								
0.000	31	1.23	0.87	0.14	0.59	1.06	1.34	4.26
0.075	28	0.87	0.46	0.00	0.51	0.88	1.20	1.97
0.200	24	0.72	0.38	0.16	0.49	0.60	1.06	1.64
0.400	25	0.73	0.59	0.12	0.32	0.54	1.04	2.40
All	108	0.91	0.65	0.00	0.48	0.81	1.19	4.26
Spermidine (nmol /mg protein)								
0.000	31	3.14	0.88	1.95	2.53	2.96	3.92	5.24
0.075	28	2.80	0.54	1.79	2.48	2.80	3.05	3.93
0.200	24	2.89	1.29	1.27	2.28	2.63	3.09	7.84
0.400	25	2.42	0.79	1.28	1.90	2.07	2.81	4.61
All	108	2.83	0.93	1.27	2.17	2.70	3.14	7.84
Spermine (nmol /mg protein)								
0.000	31	6.76	1.85	3.98	5.47	6.55	8.08	11.75
0.075	28	7.56	1.57	4.41	6.38	7.66	8.90	10.19
0.200	24	6.95	1.79	4.74	5.56	6.28	8.71	11.52
0.400	25	7.09	2.04	2.83	5.97	7.05	8.37	12.25
All	108	7.08	1.82	2.83	5.77	6.89	8.35	12.25

b. Provide descriptive statistics which compare the treatment groups with respect to the difference between the patient specific mean polyamines post randomization and the patient's polyamines at randomization (time 0). (Note that for the case representing time 0, the difference *mtrtspm – spm* is the value we are interested in for spermidine.)

<u>Ans:</u> The analysis presented below has all the failings of the one in part a, though it does have the advantage of considering the change in measurement for each patient. When measurements within a patient are highly correlated over time, such an analysis might control for initial differences among patients. We will later find that in randomized clinical trials, this is not the best way to analyze the data, however, because if the measurements are not highly correlated, we can actually lose precision when comparing across dose groups.

Table: Descriptive statistics for patient specific mean change in polyamine levels by dose group.

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Putrescine (nmol /mg protein)								
0.000	31	0.56	0.98	-0.99	0.16	0.40	0.75	4.08
0.075	28	0.21	0.57	-1.70	-0.13	0.23	0.65	0.91
0.200	24	0.11	0.53	-1.37	-0.20	0.16	0.35	1.43
0.400	25	0.07	0.63	-1.40	-0.18	0.04	0.40	1.66

All	108	0.26	0.73	-1.70	-0.11	0.19	0.50	4.08
Spermidine (nmol /mg protein)								
0.000	31	-0.14	1.31	-3.06	-0.84	0.01	0.85	1.96
0.075	28	-0.56	1.41	-4.51	-1.48	-0.31	0.55	1.21
0.200	24	-0.48	1.47	-3.88	-1.19	-0.37	0.44	1.98
0.400	25	-1.28	2.15	-5.35	-2.92	-0.82	0.16	2.41
All	108	-0.59	1.63	-5.35	-1.43	-0.22	0.53	2.41
Spermine (nmol /mg protein)								
0.000	31	-1.51	5.32	-25.49	-2.01	-1.07	0.56	6.47
0.075	28	-0.85	5.95	-29.14	-1.19	-0.23	1.45	4.53
0.200	24	-2.20	7.23	-33.50	-2.83	-1.08	0.36	6.27
0.400	25	-1.19	6.18	-27.13	-1.50	-0.81	1.42	7.62
All	108	-1.42	6.07	-33.50	-1.94	-0.89	1.24	7.62

c. Create new variable *mdrgput, mdrgspd, mdrgspm* representing the mean polyamines for each patient while taking study drug, and repeat parts (a) and (b) for this measure of treatment outcome.

Ans: The following table presents descriptive statistics for the mean polyamine value for each patient while actively taking DFMO in a way that treats all patients equally. Although we are not considering the possibility that longer treatment with DFMO might lead to different polyamine values, we are at least not confusing our measurements with those made while not taking DFMO. We are still not considering the possibility that patients with missing values for some biopsies might represent a very different subpopulation (this latter problem is not addressed by any of the analyses in this homework).

Table: Descriptive statistics for patient specific mean polyamine levels by dose group during the period of DFMO treatment

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Putrescine (nmol /mg protein)								
0.000	31	1.09	0.94	0.03	0.59	0.80	1.44	5.47
0.075	28	0.75	0.54	0.00	0.38	0.58	0.95	2.39
0.200	23	0.63	0.47	0.12	0.30	0.47	1.05	1.72
0.400	25	0.59	0.75	0.00	0.08	0.33	0.57	3.02
All	107	0.78	0.74	0.00	0.33	0.54	1.20	5.47
Spermidine (nmol /mg protein)								
0.000	31	3.39	1.18	1.78	2.53	3.20	4.08	6.91
0.075	28	2.75	0.73	1.56	2.34	2.67	3.27	4.43
0.200	23	2.75	1.49	0.83	1.67	2.53	3.23	7.84
0.400	25	2.34	0.95	1.07	1.85	2.07	2.59	4.70
All	107	2.84	1.16	0.83	2.00	2.62	3.41	7.84
Spermine (nmol /mg protein)								
0.000	31	6.89	2.24	2.32	5.14	6.79	8.33	11.59
0.075	28	7.94	1.90	3.87	6.71	7.92	9.07	11.77
0.200	23	7.22	2.50	3.41	4.94	7.02	8.09	12.04
0.400	25	7.14	2.37	2.71	6.08	7.26	8.73	12.25
All	107	7.29	2.25	2.32	5.95	7.29	8.73	12.25

In the above analysis, we do see somewhat of a trend to lower putrescine and spermidine measurements with higher dose. We can see that this does in fact represent a decrease from baseline (at least in the higher dose groups) in the following table, which presents the mean change in polyamine levels while being treated with DFMO. Spermine does not show such a consistent picture by dose: All groups seem to have decreased spermine levels.

Table: Descriptive statistics for patient specific mean change in polyamine levels by dose group while treated with DFMO (or placebo).

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Putrescine (nmol /mg protein)								
0.000	31	0.43	1.06	-1.11	0.08	0.26	0.59	5.28
0.075	28	0.08	0.67	-1.97	-0.33	0.00	0.55	1.59
0.200	23	-0.01	0.59	-1.62	-0.28	-0.14	0.28	1.51
0.400	25	-0.06	0.78	-1.40	-0.57	-0.18	0.19	2.28
All	107	0.13	0.83	-1.97	-0.28	0.06	0.43	5.28
Spermidine (nmol /mg protein)								
0.000	31	0.11	1.43	-2.89	-1.03	0.04	1.43	2.42
0.075	28	-0.61	1.33	-3.65	-1.52	-0.48	0.26	1.37
0.200	23	-0.60	1.41	-4.02	-1.25	-0.42	0.33	1.98
0.400	25	-1.36	2.26	-5.61	-3.10	-1.31	0.15	2.50
All	107	-0.58	1.70	-5.61	-1.59	-0.37	0.51	2.50
Spermine (nmol /mg protein)								
0.000	31	-1.38	5.37	-25.49	-2.02	-0.40	0.64	6.47
0.075	28	-0.47	5.93	-28.59	-1.09	0.24	1.78	4.76
0.200	23	-1.98	7.83	-34.30	-3.38	-0.94	1.26	9.50
0.400	25	-1.14	6.50	-27.80	-1.22	-0.10	1.44	7.62
All	107	-1.21	6.30	-34.30	-2.15	-0.10	1.26	9.50

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

<u>Ans:</u> As noted in the above answers, it would make the most scientific sense to count each patient equally and to take into account the times that the patients were actually receiving DFMO. Descriptively, it is beneficial to report the change in polyamine values, but as also noted above, we will find that we gain the most precision in a randomized clinical trial when we adjust for baseline values in a regression model (see Biost 518), rather than just taking the difference. In all of the analyses, we need to wonder why subjects in the highest dose group dropped out: Could it be because their polyamine levels dropped to some harmfully low level? Or is it merely a sign of DFMO toxicity that affects some other physiologic system? Or is it merely random chance that the patients who dropped out more often were at the highest dose.

5. Now suppose we consider a treatment outcome based on the minimum putrescine measurement for each patient, instead of the mean. The following code will create a variable *mindrgput* which will be the minimum of putrescine measurements made post randomization while on study drug. (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 *mindrgput*):

```
egen grbg=min(put) if time>0 & time<15, by(ptid)
egen mindrgput=mean(grbg), by(ptid)
```

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

Ans: The following table presents descriptive statistics for the minimum putrescine measurements while on treatment, as well as the change from baseline for those measurements. Note that the data for the minimum values are certainly skewed (SD large compared to the mean for these positive measurements), and the maximum value in most dose groups is markedly higher than the 75th percentile in several cases. These might be outliers, though they are not too extreme.

Table: Descriptive statistics for patient specific minimum putrescine levels by dose group while taking study drug. Also presented is the maximum decrease from baseline.

Dose	N	Mean	St Dev	Min	25th %ile	Median	75th %ile	Max
Minimum Putrescine (nmol /mg protein)								
0.000	31	0.67	0.40	0.00	0.43	0.61	0.81	1.80
0.075	28	0.39	0.23	0.00	0.26	0.38	0.56	0.82
0.200	23	0.33	0.34	0.00	0.15	0.26	0.42	1.43
0.400	25	0.28	0.44	0.00	0.00	0.02	0.38	1.73
All	107	0.43	0.39	0.00	0.17	0.38	0.60	1.80
Change from Baseline to Minimum Putrescine (nmol /mg protein)								
0.000	31	0.01	0.60	-1.19	-0.21	-0.06	0.29	1.62
0.075	28	-0.27	0.55	-2.31	-0.51	-0.24	0.07	0.71
0.200	23	-0.30	0.47	-1.78	-0.57	-0.36	0.02	0.70
0.400	25	-0.37	0.47	-1.40	-0.63	-0.38	-0.13	0.56
All	107	-0.22	0.55	-2.31	-0.51	-0.21	0.08	1.62

b. Provide descriptive statistics which compare the treatment groups with respect to the difference between the patient specific minimum putrescine on drug post randomization and the patient's putrescine at randomization (time 0). (Note that for the case representing time 0, the difference *mindrgput – put* is the value we are interested in.)

Ans: See above table.

c. What additional problem might be posed by using the minimum rather than the mean as was used in problem 4?

Ans: Looking at extreme values (minima or maxima) is heavily influenced by sample size. Hence, even if the missing data were ignorable, we would have to worry that the minimum of two measurements would logically tend to be less than the minimum of one measurement. Hence, use of the minimum with "unbalanced" data (unequal sample sizes in each group) is problematic.