Biost 517: Applied Biostatistics I
Emerson, Fall 2009

Homework #1 Key
Octoberer 30, 2009

Written problems: To be handed in at the beginning of class on Wednesday, October 7, 2009 (See the end of this handout for the Data Analysis problem to be discussed in Discussion Section October 5, 7, 9.)

> *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.)*

Questions for Biost 514 and Biost 517:

The class web pages contains a description of a dataset regarding the association between inflammation and cardiovascular disease (inflamm.doc and inflamm.txt). For this homework, we are only interested in the 5 variables measuring patient **age**, sex (**male**), body mass index **(bmi),** systolic blood pressure (**systBP**), and the ratio of blood pressures measured in the ankle to those measured in the arm (the “ankle – arm index”, **aai**), Where relevant, provide descriptive statistics for each of the variables in the entire sample, as well as within groups defined by sex. The descriptive statistics should provide information on the number of missing observations, the mean, the standard deviation, the minimum, 25th percentile, median, 50th percentile, and the maximum, where such statistics are of scientific interest.

Comment on what the results of you analysis might say about the differences between men and women in the sample.

The class web pages contain a file of annotated Stata code I used to solve this homework. I copied tables produced by Stata into Excel. I then used the “Text to Columns” feature under the Data pull-down menu. I formatted the tables in Excel, and then cut-and-pasted the table into this Word document. The Excel file that I used is posted on the web page.

There are 5000 observations in the dataset, of which

The data set was comprised of data on 2,096 males (41.9%) and 2,904 females (58.1%).

The following table presents relevant descriptive statistics on age, body mass index, systolic blood pressure, and ankle : arm index for the entire sample, as well as within groups defined by sex.

	N msng	Mean	Min	25th %ile	Mdn	75th %ile	Max
Females (n=2,904)							
Age (y)	0	73 (5.5)	65	68	71	76	100
Body mass index (kg/m^2)	9	27 (5.3)	14.7	23.2	26.1	29.6	58.8
Systolic Blood Pressure (m Hg)	7	137 (22.3)	77	122	135	151	235
Ankle : arm index	91	1.05 (0.15)	0.298	0.994	1.072	1.142	1.596
Males (n=2,096)							
Age (y)	0	73 (5.7)	65	69	72	77	95
Body mass index (kg/m^2)	4	26 (3.8)	15.6	23.9	26.1	28.5	46.2
Systolic Blood Pressure (m Hg)	3	136 (21.3)	79	121	133	148	219
Ankle : arm index	30	1.08 (0.20)	0.278	1.007	1.107	1.195	2.385
All patients (n=5,000)							
Age (y)	0	73 (5.6)	65	68	72	76	100
Body mass index (kg/m^2)	13	27 (4.7)	14.7	23.5	26.1	29.2	58.8
Systolic Blood Pressure (m Hg)	10	137 (21.9)	77	121	134	150	235
Ankle : arm index	121	1.06 (0.17)	0.278	1.000	1.085	1.167	2.385

From this table we see that ages range from 65 to 100 years. The men and women are relatively similar with respect to their age, BMI, SBP and ankle : arm index distributions. There do appear to be some subjects with relatively large values for BMI (at least one woman with 58.8 kg / m^2 and at least one man with 46.2 kg / m^2) and at least one man with a large ankle : arm index (2.385). These large values, while not impossible, could represent data entry errors.

Additional question for Biost 514

In the following presentation of methods I suggest that the sample mean of a binary variable can be used to compute the proportion. If you use the commands given below, it will also compute the sample standard deviation of the binary variables. Why is this boring?

Notes on using Stata to answer this homework

In solving this problem, you are encouraged to use Stata or some other statistical package. We can help you with Stata (or with R or S-Plus in office hours). The following Stata commands may be of use.

1. If using the HSLIC microcomputer lab, I highly recommend that you use a USB drive or a floppy drive to save your work. The instructions that follow presume that you will be using a USB drive, and that that drive has been designated E: by Windows. (Macintosh or Unix users will have to modify commands for their operating systems.)
2. Copy the file inflamm.txt onto your USB drive (I presume you will just use the root directory of this drive).
3. Start Stata, and change the working directory to your USB drive by typing the following command into the Commands window:

```
cd e:\
```

4. Read the data into Stata by typing the following command into the Commands window and then pressing ENTER:

   ```
   infile id site age male bkrace smoker estrogen prevdis
   diab2 bmi systBP aai cholest crp fib ttodth death
   cvddeath using inflamm.txt
   ```

   (The file I gave you has white space delimited data. Had it been commas or tabs, you could have used the insheet command.)
5. The first line of the file contained the variable names, which Stata could not read as a number. That case has thus been entered as missing data (denoted as '.' by Stata). We might as well drop that case by typing the following command into the Commands window and then pressing ENTER:

   ```
   drop in 1
   ```
6. To produce nice formatting of data, you might specify the format you want the results to be presented in. Some authors recommend only showing 3 significant digits at any time. This is hard to accomplish exactly, but we can do something toward that goal:
   a. For instance, age tends to be a number between about 65 and 100 (in these patients), so nicely formatted output would have one digit behind the decimal place. You can tell Stata that you always want bili results printed with that level of precision by specifying in the Commands window: `format age %9.1f`
   b. Similarly, the ankle arm index tends to be between 0 and 2, so for this variable we might only want 2 digits behind the decimal point: `format aai %9.2f`
7. The indicator of sex is a binary variable. The only summary statistics that really are of much use are the frequencies (i.e., the number of cases who are female and male) or the proportion of the sample in each category. When considering the proportion, I often get lazy and just use the Stata functions for the mean, because the mean of a 0-1 variable is the proportion of the sample that has a value of 1. (How does it behave for a 1-2 variable?) For that reason, I choose to format this variable to give 3 significant digits by specifying 3 digits behind the decimal point: `format male %9.3f`
8. Now, having gotten the dataset in shape, I am ready to do analyses. But because I might have to stop in the middle, and because I don't want to have to do all of this again, I choose to save the dataset as a Stata data file: `save inflamm` This will create a file named inflamm.dta in my default directory (defined using `cd` above). In a future Stata session, I will be able to access this file by first specifying the default directory, and then entering the command `use inflamm.`
9. So now for the descriptive statistics. See what the following commands do:
   a. `summarize`
   b. `summ`
   c. `summ age`
   d. `summ age, detail`
   e. `summ age systBP bmi`
   f. `tabstat age male systBP bmi aai, stat(n mean sd min p25 p50 p75 max)`
   g. `tabstat age male bmi systBP aai, stat(n mean sd min p25 p50 p75 max) col(stat)`
   h. `bysort male: tabstat age bmi systBP aai, stat(n mean sd min p25 p50 p75 max) col(stat)`
   i. `tabstat age bmi systBP aai, stat(n mean sd min p25 med p75 max) col(stat) by(male)`

10. Personally, I find it easiest to cut and paste output that I want to keep into a Word document. Sometimes, I first put it in Excel in order to get formatting the way I want. Other times, I just enter by hand the relevant numbers gleaned from the output.
11. When you are done with Stata, you can just close the window. It may ask if you want to save the dataset.

DATA ANALYSIS

To be discussed in discussion section October 5, 7, and 9..

We will discuss the general approach to an analysis of the scientific question posed in the documentation for the data set on smoking and lung function in children (fev.doc and fev.txt on the class web pages). Based on your reading of the documentation, be prepared to discuss the information you would want to address the scientific question.