Biost 517: Applied Biostatistics I
Emerson, Fall 2007

Homework #1 Key
October 13, 2007

Written problems: To be handed in at the beginning of class on Wednesday, October 3, 2007 (See the end of this handout for the Data Analysis problem to be discussed in Discussion Section October 3, 5, 8.)

> *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 class web pages contain a description of a dataset regarding the screening of patients for inclusion in a clinical trial of methotrexate in the treatment of primary biliary cirrhosis (pbcscreen.doc and pbcscreen.txt). 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 how the results of your descriptive analyses relate to the scientific question posed in the description of the data.

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 535 observations in the dataset, and upon inspecting the subject identification numbers (*ptid*), I found that there were 535 unique values suggesting that each observation in this dataset was made on a different subject. Measurements were available on subject sex, age (in years), serum albumin (g/dl), serum alkaline phosphatase (U/l), serum alanine transaminase (ALT) (U/l), serum aspartate transaminase (AST) (U/l), serum bilirubin (mg/dl), and serum cholesterol (mg/dl). height (in inches), sex, self-reported current smoking status (yes/no), and 1 second forced expiratory volume (FEV) (l/sec). Some cases were missing data for the serum measurements.

The data set was comprised of data on 34 males (6.36%) and 501 females (93.6%).

The following table presents relevant descriptive statistics on age and serum measurements for the entire sample, as well as within groups defined by sex.

	N msng	Mean (SD)	Min	25th %ile	Mdn	75th %ile	Max
Males (n=34)							
Age (y)	**0**	**53 (8.1)**	**37**	**46**	**52**	**60**	**66**
Albumin (g/dl)	**3**	**4.01 (0.37)**	**3**	**3.8**	**4.1**	**4.3**	**4.6**
Alkaline Phosphatase (U/l)	**0**	**334 (179.0)**	**107**	**203**	**303**	**439**	**744**
ALT (U/l)	**0**	**117 (106.0)**	**18**	**50**	**87**	**147**	**550**
AST (U/l)	**0**	**100 (98.0)**	**23**	**50**	**70**	**100**	**550**
Bilirubin (mg/dl)	**3**	**1.4 (2.0)**	**0.2**	**0.6**	**0.8**	**1.1**	**10.2**
Cholesterol (mg/dl)	**7**	**263 (95.0)**	**138**	**194**	**237**	**301**	**513**
Females (n=501)							
Age (y)	**0**	**52 (9.7)**	**1**	**46**	**51**	**59**	**80**
Albumin (g/dl)	**27**	**3.96 (0.45)**	**1.8**	**3.8**	**4**	**4.3**	**5.2**
Alkaline Phosphatase (U/l)	**4**	**375 (329.0)**	**60**	**174**	**281**	**458**	**3741**
ALT (U/l)	**4**	**110 (364.0)**	**8**	**42**	**62**	**108**	**5550**
AST (U/l)	**2**	**95 (252.0)**	**12**	**45**	**69**	**103**	**5550**
Bilirubin (mg/dl)	**21**	**1.1 (2.1)**	**0.1**	**0.5**	**0.7**	**1.1**	**35.2**
Cholesterol (mg/dl)	**106**	**249 (70.0)**	**79**	**205**	**235**	**283**	**716**
All patients (n=535)							
Age (y)	**0**	**52 (9.6)**	**1**	**46**	**51**	**59**	**80**
Albumin (g/dl)	**30**	**3.97 (0.44)**	**1.8**	**3.8**	**4**	**4.3**	**5.2**
Alkaline Phosphatase (U/l)	**4**	**372 (321.0)**	**60**	**174**	**282**	**458**	**3741**
ALT (U/l)	**4**	**111 (353.0)**	**8**	**42**	**64**	**111**	**5550**
AST (U/l)	**2**	**95 (245.0)**	**12**	**46**	**69**	**102**	**5550**
Bilirubin (mg/dl)	**24**	**1.1 (2.1)**	**0.1**	**0.5**	**0.7**	**1.1**	**35.2**
Cholesterol (mg/dl)	**113**	**250 (72.0)**	**79**	**205**	**236**	**284**	**716**

From this table we see that ages range from 1 to 80 years, with the young age of at least one patient extremely unusual (could this be a data entry error?).. The fact that the vast majority of patients are women is not surprising given what is known about this auto-immune disease. The extremely low values of albumin and the high values of alkaline phosphatase, ALT, AST, and bilirubin are not unexpected for a population that might include some patient with severe PBC. The large standard deviations relative to the means of alkaline phosphatase, ALT, AST and bilirubin suggests that the largest values might appear to be “outliers” in the sample. The maximal observed values of ALT and AST, however, are perhaps suggestive of much more severe disease than are the alkaline phosphatase and bilirubin measurements. Again, this might be suggestive of a data entry error.