Biost 517: Applied Biostatistics I
Emerson, Fall 2005

Homework #1 Key
October 5, 2005

Written problems: To be handed in at the beginning of class on Wednesday, October 5, 2005. (See the end of this handout for the Data Analysis problem to be discussed in Discussion Section October 5-10.)

> *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 contains a description of a dataset regarding the association between lung function and smoking in children (fev.doc and fev.txt). Where relevant, provide descriptive statistics for each of the variables in the entire sample, as well as within groups defined by smoking status. 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.

ANSWER

The class web pages contain a file of annotated Stata code I used to solve this homework.

The variable *seqnbr* is not of any scientific or statistical interest and is thus not commented on further.

There are 654 observations in the dataset, and upon inspecting the subject identification numbers (*subjid*), I found that there were 654 unique values suggesting that each observation in this dataset was made on a different subject. Measurements were available on subject age (in years), height (in inches), sex, self-reported current smoking status (yes/no), and 1 second forced expiratory volume (FEV) (l/sec). There are no cases with missing data on any variable.

The data set was comprised of data on 318 females (48.6%) and 336 males (51.4%). The vast majority of subjects (589 or 90.1%) reported that they were currently nonsmokers, while only 9.9% (n=65) reported being a current smoker. Smoking was slightly more prevalent among females (39 smokers of 318 females, or 12.3%) than it was among males (26 smokers of 336 males, or 7.74%).

The following table presents relevant descriptive statistics for the entire sample, as well as within groups defined by self-reported current smoking status.

	Mean	Std Dev	Min	25th Pctile	Median	75th Pctile	Max
(All Subjects: N=654)							
Age (y)	9.9	3.0	3.0	8.0	10.0	12.0	19.0
Height (in)	61.1	5.7	46.0	57.0	61.5	65.5	74.0
FEV (l/sec)	2.64	0.87	0.79	1.98	2.55	3.12	5.79
(Nonsmokers: N=589)							
Age (y)	9.5	2.7	3.0	8.0	9.0	11.0	19.0
Height (in)	60.6	5.7	46.0	57.0	61.0	64.5	74.0
FEV (l/sec)	2.57	0.85	0.79	1.92	2.46	3.05	5.79
(Smokers: N=65)							
Age (y)	13.5	2.3	9.0	12.0	13.0	15.0	19.0
Height (in)	66.0	3.2	58.0	63.5	66.0	68.0	72.0
FEV (l/sec)	3.28	0.75	1.69	2.8	3.17	3.75	4.87

From this table we see that ages range from 3 to 19 years—an age range that is a little surprising given the major interest in the effects of smoking. This is further highlighted when we examine the descriptive statistics within groups defined by smoking status: Nonsmokers encompass the entire range of ages, while no smoker is less than 9 years old. As might therefore be expected, the summary statistics for FEV and height are also higher among smokers than nonsmokers. Because height is related to both age and FEV, and because age is related to smoking, for greatest scientific relevance, I would think that we will need to make some adjustment for the age discrepancy seen between the smokers and nonsmokers.