Biost 517: Applied Biostatistics I
Emerson, Fall 2006

Homework #1 Key
October 22, 2006

Written problems: To be handed in at the beginning of class on Wednesday, October 4, 2006. (See the end of this handout for the Data Analysis problem to be discussed in Discussion Section October 4, 6, 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.)*

The class web pages contains a description of a dataset regarding the association between lung function and smoking in elderly adults (adultfev.doc and adultfev.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.

There are 735 observations in the dataset, and upon inspecting the subject identification numbers (*id*), 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). Ten cases were missing data for FEV, but no cases were missing data on any other variable.

The data set was comprised of data on 369 females (50.2%) and 366 males (49.8%). The vast majority of subjects (636 or 86.5%) reported that they were currently nonsmokers, while only 13.5% (n=99) reported being a current smoker. Smoking was slightly more prevalent among females (57 smokers of 369 females, or 15.4%) than it was among males (42 smokers of 366 males, or 11.5%).

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

	N msng	Mean	Std Dev	Min	25th Pctile	Median	75th Pctile	Max
(Nonsmokers: N=636)								
Age (y)	0	74.8	5.5	65.0	71.0	74.0	78.0	99.0
FEV (l/sec)	7	2.25	0.69	0.41	1.80	2.21	2.70	4.47
Height (in)	0	65.3	3.8	54.5	62.5	65.5	68.5	74.5
(Smokers: N=99)								
Age (y)	0	73.1	4.6	67.0	70.0	72.0	75.0	89.0
FEV (l/sec)	3	1.89	0.59	0.57	1.53	1.89	2.22	3.84
Height (in)	0	64.9	4.1	55.5	62.0	64.5	67.5	75.0
(All Subjects: N=735)								
Age (y)	0	74.6	5.5	65.0	71.0	74.0	78.0	99.0
FEV (l/sec)	10	2.21	0.69	0.41	1.75	2.16	2.65	4.47
Height (in)	0	65.3	3.8	54.5	62.0	65.5	68.5	75.0

From this table we see that ages range from 65 to 99 years, with heights seemingly appropriate for this elderly population. The summary statistics for FEV suggest a tendency for the smokers to have lower FEV than nonsmokers: The mean and 25^{th}, 50^{th}, and 75^{th} percentiles are lower in the smokers than the nonsmokers. (The minimum and maximum are both more extreme in the nonsmokers, but as the sample size is markedly larger for the smokers, it is difficult to compare the sample extrema in a scientifically meaningful way.)