

```
#### Biost 517: Applied Biostatistics I
#### Emerson, Fall 2006
```

```
#### Homework #1 Key
#### Annotated Stata Log File
#### October 22, 2006
```

```
#### NOTE: I most definitely did not want you to hand in such
#### output as this. I do this to aid you in understanding
#### how I got the answers for the Key.
```

```
#### Comments edited into the log file produced by Stata are
#### on the lines that start with the four '#' signs and are
#### printed in italics.
```

```
#### The Stata commands are put in bold face.
```

```
#### Stata output is displayed in regular typeface in blue.
```

```
#### Reading in the data from the textfile
```

```
. infile id age fev height male smoke using adultfev.txt
```

```
'id' cannot be read as a number for id[1]
'age' cannot be read as a number for age[1]
'fev' cannot be read as a number for fev[1]
'height' cannot be read as a number for height[1]
'male' cannot be read as a number for male[1]
'smoke' cannot be read as a number for smoke[1]
'NA' cannot be read as a number for fev[11]
'NA' cannot be read as a number for fev[150]
'NA' cannot be read as a number for fev[197]
'NA' cannot be read as a number for fev[204]
'NA' cannot be read as a number for fev[230]
'NA' cannot be read as a number for fev[250]
'NA' cannot be read as a number for fev[273]
'NA' cannot be read as a number for fev[434]
'NA' cannot be read as a number for fev[669]
'NA' cannot be read as a number for fev[683]
(736 observations read)
```

```
#### Drop the first case, because it was just the column headings
```

```
. drop in 1
```

```
(1 observation deleted)
```

```
#### Declare the format to provide approximately 3 significant digits in print out
. format age height %9.1f
. format fev %9.2f
```

```
#### Save the data file so I don't have to do all of the above again
. save adultfev
file adultfev.dta saved
```

```
#### Checking to see if all subject ID numbers are unique.
#### I do this using the "by subjid:" prefix with the "egen"
#### command which will generate a new variable containing a
#### constant equal to the count of nonmissing data. Then
#### when I do a table of that new constant, I find that
#### there are 654 cases with the value 1. Had there been
#### a duplicate subject ID number, I might have found, say,
#### 652 cases with a value of 1 and 2 cases with a value of 2.
. egen idcnt= count(id), by(id)
. table idcnt
```

```
-----
      idcnt |          Freq.
-----+-----
           1 |             735
-----
```

```
#### Descriptive statistics for the entire sample in the format I like.
#### Note the fact that I specified the statistics that I wanted, I
#### specified that the statistics were to be in columns, and I specified
#### that I wanted Stata to use the formats that I had pre-specified for
#### the variables.
```

```
. tabstat age height fev, stat(n mean sd min p25 p50 p75 max) col(stat) format
```

```
-----+-----
      variable |          N      mean      sd      min      p25      p50      p75      max
-----+-----
           age |       735.0       74.6       5.5       65.0       71.0       74.0       78.0       99.0
           fev |       725.00       2.21       0.69       0.41       1.75       2.16       2.65       4.47
           height |       735.0       65.3       3.8       54.5       62.0       65.5       68.5       75.0
-----
```

Now doing the same within groups defined by smoking status. Note
 #### that I had to sort the data first. I could have avoided that had
 #### I used the command "bysort" instead of "by".

```
. sort smoke
. by smoke: tabstat age height fev, stat(n mean sd min p25 p50 p75 max) col(stat) format
```

Summary for variables: age fev height
 by categories of: smoke

smoke	N	mean	sd	min	p25	p50	p75	max
0	636.0	74.8	5.5	65.0	71.0	74.0	78.0	99.0
	629.00	2.25	0.69	0.41	1.80	2.21	2.70	4.47
	636.0	65.3	3.8	54.5	62.5	65.5	68.5	74.5
1	99.0	73.1	4.6	67.0	70.0	72.0	75.0	89.0
	96.00	1.89	0.59	0.57	1.53	1.89	2.22	3.84
	99.0	64.9	4.1	55.5	62.0	64.5	67.5	75.0
Total	735.0	74.6	5.5	65.0	71.0	74.0	78.0	99.0
	725.00	2.21	0.69	0.41	1.75	2.16	2.65	4.47
	735.0	65.3	3.8	54.5	62.0	65.5	68.5	75.0

Crosstabulation of smoking status and sex. I asked to get the
 #### row, column, and cell percentages as well as the counts.

. tabulate smoke female, row column

```
+-----+
| Key   |
+-----+
|       |
| frequency |
| row percentage |
| column percentage |
| cell percentage  |
+-----+
```

male	smoke		Total
	0	1	
0	312	57	369
	84.55	15.45	100.00
	49.06	57.58	50.20
	42.45	7.76	50.20
1	324	42	366
	88.52	11.48	100.00
	50.94	42.42	49.80
	44.08	5.71	49.80
Total	636	99	735
	86.53	13.47	100.00
	100.00	100.00	100.00
	86.53	13.47	100.00