

Biost 517: Applied Biostatistics I
Emerson, Fall 2005

Homework #1 Key
Annotated Stata Log File
October 5, 2005

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 seqnbr subjid age fev height sex smoke using fev.txt
'seqnbr' cannot be read as a number for seqnbr[1]
'subjid' cannot be read as a number for subjid[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]
'sex' cannot be read as a number for sex[1]
'smoke' cannot be read as a number for smoke[1]
(655 observations read)
```

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

```
. drop in 1
(1 observation deleted)
```

Create indicator variables in the format that I prefer

```
. gen female = sex - 1
. gen smoker = 2 - smoke
. drop sex smoke
```

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 fev

file fev.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.

. sort subjid

. by subjid: egen idcnt= count(subjid)

. table idcnt

idcnt	Freq.
1	654

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	654.0	9.9	3.0	3.0	8.0	10.0	12.0	19.0
height	654.0	61.1	5.7	46.0	57.0	61.5	65.5	74.0
fev	654.00	2.64	0.87	0.79	1.98	2.55	3.12	5.79

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
```

```
-----
-> smoker = 0
```

variable	N	mean	sd	min	p25	p50	p75	max
age	589.0	9.5	2.7	3.0	8.0	9.0	11.0	19.0
height	589.0	60.6	5.7	46.0	57.0	61.0	64.5	74.0
fev	589.00	2.57	0.85	0.79	1.92	2.46	3.05	5.79

```
-----
-> smoker = 1
```

variable	N	mean	sd	min	p25	p50	p75	max
age	65.0	13.5	2.3	9.0	12.0	13.0	15.0	19.0
height	65.0	66.0	3.2	58.0	63.5	66.0	68.0	72.0
fev	65.00	3.28	0.75	1.69	2.80	3.17	3.75	4.87

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

. tabulate smoke female, row column

+-----+			
Key			
+-----+			
frequency			
row percentage			
column percentage			
+-----+			
smoker	female		Total
	0	1	

0	310	279	589
	52.63	47.37	100.00
	92.26	87.74	90.06

1	26	39	65
	40.00	60.00	100.00
	7.74	12.26	9.94

Total	336	318	654
	51.38	48.62	100.00
	100.00	100.00	100.00