***Stata tips #2***

***Epidemiologic Methods***

***Adapted from notes of previous TAs***

This week:

1. Generating categorical variables from continuous variables (using if/and/or options)
2. Review labeling variables
3. The tabstat and bysort command
4. Hypothesis testing (categorical and continuous variables)
5. Table1

*Note: For this Stata session we will be using the Framingham dataset*

# **Categorical variables from continuous variables**

Sometimes it makes more sense to break up a continuous variable into categories for data exploration and modeling. For example, if we want to categorize one’s cigarette smoking behavior as light (1-2/day), medium (3-5/day), and heavy (5+), we can do the following:

gen smkbhv=.

replace smkbhv=1 if cigpday>=1 & cigpday<=2 replace smkbhv=2 if cigpday>=3 & cigpday<=5 replace smkbhv=3 if cigpday>5 & cigpday!=.

An alternative to the first line is:

gen smkbhv=1 if cigpday=1 | cigpday<=2

A few notes:

* “&”=and, “|” =or
* Warning: Stata treats a missing value as positive infinity, so if you only type “replace smkbhv=3 if cigpday>5” without the “!=.”, you will erroneously get an extra 1754 individuals categorized as heavy smokers who actually having missing data for cigpday!
* Always use the “tabulate” function after generating new variables to make sure that you did it correctly.

. tab smkbhv

smkbhv | Freq. Percent Cum.

+

1| 260 5.25 5.25

2 | 429 8.67 13.92

3 | 4,261 86.08 100.00 -

+

Total | 4,950 100.00

To make sure we don’t forget later what “1”, “2”, “3” represent, we can create a variable label that appears on the upper right hand under the variables window. This is similar to a legend:

label var smkbhv "1=light(1-2cig), 2=medium(3-5cig), 3=heavy(5+cig)“

Now when we tabulate smkbhv, it’s labeled for us:

1=light(1-2 cig), 2=medium(3-5cig), 3=heavy(5+cig)“

|  | Freq. | Percent | Cum. |
| --- | --- | --- | --- |
| 1 | 260 | 5.25 | 5.25 |
| 2 | 429 | 8.67 | 13.92 |
| 3 | 4,261 | 86.08 | 100.00 |
| Total | 4,950 | 100.00 |  |

What if I actually want the numbers “1”, “2”, “3” in the variable list to be defined as “light”, “medium” and “heavy” when I use the “tabulate” command? It’s a two-step command:

label define smkbhv\_label 1 “light” 2 “medium” 3 “heavy” label values smkbhv smkbhv\_label

# **The tabstat command:**

We can use the ‘tabstat’ command to tabulate summary statistics of any continuous variable by categorical variables.

tabstat[originalVariable], by([categoryVariable]) statistics(n mean min max)

. tabstat age, by(sex) stat(n mean min max)

Summary for variables: age

by categories of: sex (sex)

sex | N mean min max

+

1 | 5022 54.49363 33 80

2 | 6605 55.02029 32 81

+

Total | 11627 54.79281 32 81



# **The bysort command:**

While exploring the data, you may want to look at one variable stratified by another variable. For example, what is the smoking behavior by sex? With a chi-square value?

Bysort sex: tab smkbhv

But this is the long form of the Framingham data, and simply tabulating over all the data is probably not very meaningful—let’s restrict the tabulation to period 1 only:

bysort sex: tab smkbhv if period==1

# **Hypothesis testing**

Now that you have learned how to create the descriptive aspects of a Table 1, the next step is to learn how to perform hypothesis testing in Stata to complete your table. Tests of statistical significance show up in most Table 1’s of observational studies (i.e. baseline characteristics) in the medical literature, and two of the most common tests performed are:

* Chi-square tests: compares frequency distributions of categorical variables
* t-tests: mean comparisons of continuous variables

Today we will cover how to perform hypothesis testing with these two types of variables. **You will need to perform chi-square tests for HW3, and t-tests for future homeworks**.

# **Categorical variables Chi-square tests:**

Let’s say we were interested in exploring the relationship between sex and smoking status. In Stata, type:

# **. tab sex cursmoke**

| current cig smoker

| y/n

sex | 0 1 | Total

+ +

1 | 2,428 2,594 | 5,022

2 | 4,170 2,435 | 6,605

+ +

Total | 6,598 5,029 | 11,627

The table above is a 2x2 contingency table, and the frequency of smokers and nonsmokers by sex (1=men, 2=women) are shown above. For example, there are a total of 5022 total instances of sex coded as a “1” and the subject is currently smoking in 2594 of these instances. Note that in the command above we are ignoring the longitudinal aspects of the data (i.e. the fact that people are measured in up to 3 different periods in this data). If we were interested only in period 1, the “if” command could be used here:

# **. tab sex cursmoke if period==1**

| current cig smoker

| y/n

sex | 0 1 | Total

+ +

1 | 769 1,175 | 1,944

2 | 1,484 1,006 | 2,490

+ +

Total | 2,253 2,181 | 4,434

Now lets take a look at some of the options available for the tabulate command. The first option that we will look at is the “row” option. In Stata type:

# **. tab sex cursmoke if period==1, row**

+ +

| Key |

| |

| frequency |

| row percentage |

+ +

| current cig smoker

| y/n

sex | 0 1 | Total

+ +

0 | 769 1,175 | 1,944

| 39.56 60.44 | 100.00

+ +

1 | 1,484 1,006 | 2,490

| 59.60 40.40 | 100.00

+ +

Total | 2,253 2,181 | 4,434

| 50.81 49.19 | 100.00

As shown above, the row option provides us with the row percentages in our 2x2 table. This is in contrast to the “column” option, which provides us with column percentages.

# **. tab sex cursmoke if period==1, column**

+ +

| Key |

| |

| frequency |

| column percentage |

+ +

| current cig smoker

| y/n

sex | 0 1 | Total

+ +

1 | 769 1,175 | 1,944

| 34.13 53.87 | 43.84

+ +

2 | 1,484 1,006 | 2,490

| 65.87 46.13 | 56.16

+ +

Total | 2,253 2,181 | 4,434

| 100.00 100.00 | 100.00

By utilizing these options, you can quickly and easily calculate the necessary information for your Table 1 (where categorical variables are often displayed as number (percent) for each category).

Let’s say we are interested in determining whether the differences in smoking status in men versus women is statistically significant. The most commonly used test for this, especially in large samples, is the chi-squared test.

# **. tab cursmoke sex if period==1, chi**

current |

cig smoker | sex

y/n | 1 2 | Total

+ +

0 | 769 1,484 | 2,253

1 | 1,175 1,006 | 2,181

+ +

Total | 1,944 2,490 | 4,434

Pearson chi2(1) = 175.4299 Pr = 0.000

You can see at the bottom of the output that the degrees of freedom (chi2(1), a chisquared value (175), and a P-value (0.000) are all provided. We will not go in to depth regarding the degrees of freedom (although from your stats classes you may remember that here it denotes the number of independent pieces of information in the table above; if I gave you the value of one of the four center squares you could use the marginal values to fill in the rest of the table). For our purposes, the P-value is what we are interested in. You should know that a P-value cannot be equal to zero here, and that in Stata a P-value of 0.000 is actually <0.0005, so reporting P<0.001 here is what many investigators would do.

You should also note that the order in which you type the row and column variables does not matter for the chi-squared test. I usually run the test so that the table in Stata looks like the table in my paper in terms of column/row orientation.

Finally, it is important to emphasize that the numbers and percent in each group are often as important as the P-values because of the relationship between the P-value and sample size, so be aware of this in your interpretations.

# **Continuous variables**

Now we will touch on how to compare the means between two groups using t-tests in Stata. Let’s compare the mean systolic blood pressure between men and women. First, to “eyeball” the data and get the mean and standard deviation for our table 1, we will use the “summary” command:

# **. sum sysbp if period==1 & sex==1**

| Variable |  + | Obs | Mean | Std. Dev. | Min | Max |
| --- | --- | --- | --- | --- | --- |
| sysbp | | 1,944 | 131.7369 | 19.44124 | 83.5 | 235 |

**. sum sysbp if period==1 & sex==2**

| Variable |  + | Obs | Mean | Std. Dev. | Min | Max |
| --- | --- | --- | --- | --- | --- |
| sysbp | | 2,490 | 133.8219 | 24.46263 | 83.5 | 295 |

We could have generated the same results using “bysort sex: sum sysbp if period==1,” which we went over in Session 2. Also, remember the “detail” option for the summary command as we discussed previously to get more detailed information such as skewness, median and inter-quartile range, and some of the highest and lowest values in the dataset.

You can see that the mean systolic blood pressure is slightly higher in women than in men, but is this difference statistically significant? To compare the two means using ttests in Stata, you can use the ttest command:

# **. ttest sysbp if period==1, by (sex)**

Two-sample t test with equal variances



Group | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

+

1 | 1944 131.7369 .4409362 19.44124 130.8721 132.6016

2 | 2490 133.8219 .490234 24.46263 132.8606 134.7832

+

combined | 4434 132.9078 .3367198 22.4216 132.2476 133.5679

+

diff | -2.085005 .6779578 -3.414141 -.755869 ---



diff = mean(1) - mean(2) t = -3.0754 Ho: diff = 0 degrees of freedom = 4432

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0 Pr(T < t) = 0.0011 Pr(|T| > |t|) = 0.0021 Pr(T > t) = 0.9989

Testing that the means between the two groups are equal is the same as testing that the difference between the means of the two groups is equal to zero. That’s why you see the last row in the table as the difference between the two means, the term under the table “Ho: diff=0” (null hypothesis that the difference equals zero), and in the middle at the bottom Ha: diff !=0 (alternative two-sided hypothesis that the difference is not equal to zero). For a two-sided test (which we should use here and is almost always used in practice), the P-value we are interested in P=0.0021, which we should report as P=0.002 (or maybe as P<0.01).

FYI, you can also perform a one-sample t-test (testing that the mean is equal to a specific number):

**ttest sysbp==120 if period==1 & sex==1**