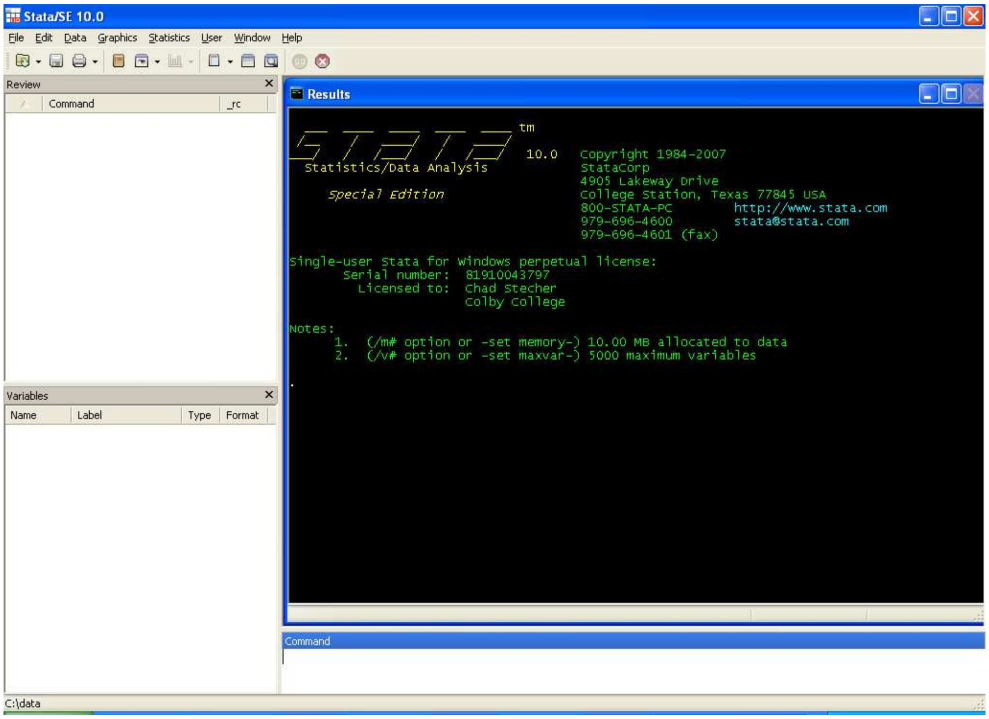
# MA397B – Applied Regression Modeling
# Introduction to Stata

## Goals

The objectives of this exercise are to review some basic concepts of hypothesis testing, and to become familiar with the use of Stata 10. Stata is a useful and relatively straightforward software package for analyzing data available for both PCs and Macs. All data for this handout are available from the course webpage: http://www.colby.edu/personal/l/lobrien/ma397.html.

## Overview of Stata

Stata 10 is available on the keyserver for both PCs and Macs. You may run it off the keyserver from the machines in Olin 323, or you may run it from the computers in the Davis classroom in Miller. Its location varies depending on the computer you're using, but it's generally under **Start > Statitsics > Intercooled Stata 10**. When you start the program you should see something similar to the following.



```
                                   tm
                                  10.0    Copyright 1984-2007
  Statistics/Data Analysis                StataCorp
                                          4905 Lakeway Drive
     Special Edition                      College Station, Texas 77845 USA
                                          800-STATA-PC        http://www.stata.com
                                          979-696-4600        stata@stata.com
                                          979-696-4601 (fax)

Single-user Stata for Windows perpetual license:
       Serial number:  81910043797
         Licensed to:  Chad Stecher
                       Colby College

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables

.
```

Note that Stata has four basic windows:

1. The **Stata Results** window will display all the commands issued to Stata, as well as all the output resulting from these commands (with the exception of graphs).

2. The **Stata Command** window is used to type commands directly into Stata. Most of what you will be doing will be done through the drop down menus.
3. The **Review** windows keeps a running history of all the commands that you have issued Stata during your session. To recall a command, simply click on it in this window.
4. The **Variables** window shows you all the variables that you currently have in your dataset.

If you should ever "lose" a window, click on **Prefs > Default Windowing** and all four will return. There are other windows in Stata, but the only other one you're likely to encounter is the **Graph** window. This window pops up anytime you generate a graph. You can right-lick on a graph for printing and copying options. Graphs can be copied in Stata and pasted into any Microsoft Office application (such as Word).

Oftentimes you want to keep a running history of all your commands and the output from those commands. You can do this by using a *log* file. To begin a log file click on the 📒 button and choose where you would like to save the file (use the .log extension rather than the .scml extension). This file is a text file and can be opened in Notepad or any word processing package.

**Review of Descriptive Statistics**

**Numerical Summaries**

Download the *hbp.dta* file from the course webpage. The variable "baseline" contains the baseline blood pressures, "final" contains the final blood pressure, and "med" contains the indictor of medication group. We are interested in summarizing the baseline pressures in this exercise. To calculate the numerical summary measures choose **Statistics** > **Summaries, tables & tests** > **Summaries and Descriptive Statistics** > **Summary Statistics.** Enter the name of he variable you want to summarize (baseline) in the variables box and hit enter.

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    baseline |        33    152.0303    33.20061        100        230
```

Alternately, you can ask Stata to give you summary statistics split by levels of a grouping variable. To do this, click on the by/if/in tab and click on the "Repeat command for groups defined by" box and enter the grouping variable (med) in the box.

```
-------------------------------------------------------------------------
-> med = Nifedipine

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    baseline |        18    150.8333    34.05575        106        230


-------------------------------------------------------------------------
```

```
-> med = Propranolol

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    baseline |        15    153.4667    33.27347        100        230
```

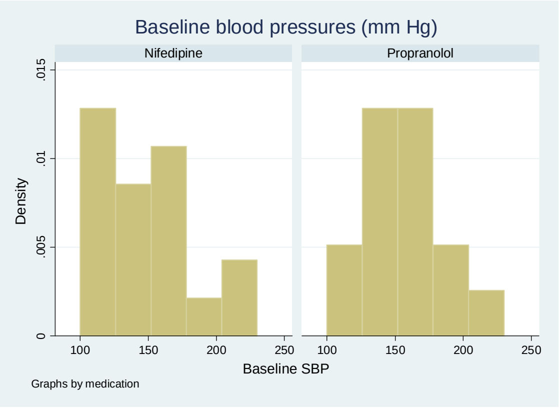
Note that we haven't been given medians or quartiles. To get those click on the "Display additional statistics" button in the dialog box. For example, for the entire data set the median is 150.

```
. summarize baseline, detail

                        Baseline SBP
-------------------------------------------------------------
      Percentiles      Smallest
 1%          100            100
 5%          100            100
10%          110            106       Obs                  33
25%          128            110       Sum of Wgt.          33

50%          150                      Mean           152.0303
                        Largest       Std. Dev.      33.20061
75%          170            190
90%          190            210       Variance        1102.28
95%          230            230       Skewness       .5481966
99%          230            230       Kurtosis       3.107986
```

## Graphical Summaries

Stata can also generate a wide variety of graphs. For this exercise we'll only examine histograms, but you can see many of the other graphics options by exploring the menu options.

To generate a histogram of baseline pressures click on **Graphics > Histogram** and enter the variable name in the variable box (baseline in this case). You can also split the histogram across levels of a grouping variable by clicking on the "By" tab and entering the grouping variable in the box after clicking "Draw subgraphs for unique values of variables" box.



### One-sample t-test

We can use a one-sample t-test to compare the final blood pressure in the combined group to see if it differs from 120. To do this in Stata, select **Statistics > Summaries, tables & tests > Classical tests of hypotheses > One sample mean comparison test.** You should enter the name of the variable you want to test in the *variable* box (final in this case), and enter the hypothesized null value of 120 in the hypothesized mean box.

The output is below. Notice that there are three p-values at the bottom of the output. The middle p-value (0.0000) is for a **two-sided** test. You should use this if you are interested in detecting any difference from the null value (i.e., you do not have reason to believe that one will always be bigger than the other). The p-value on the left (1.000) means the mean is smaller than the null value. The p-value on the right (0.0000) is for the one-sided test that the mean is greater than the null value.

```
. ttest final == 120

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   final |      33    143.6364    4.718961    27.10837    134.0242    153.2486
------------------------------------------------------------------------------
    mean = mean(final)                                            t =   5.0088
Ho: mean = 120                                   degrees of freedom =       32

   Ha: mean < 120               Ha: mean != 120               Ha: mean > 120
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

### Two-sample t-test

We can use a two-sample t-test to compare the baseline blood pressures in the two medication groups. To do this in Stata, select **Statistics > Summaries, tables & tests > Classical tests of hypotheses > Two group mean comparison test**. You should enter the name of the variable you want to test in the *variable* box (baseline in this case), and enter the grouping variable in the other box (med in this case).

The output is below. Notice that there are three p-values at the bottom of the output. The middle p-value (0.0010) is for a **two-sided** test. You should use this if you are interested in detecting any difference in the means (i.e., you do not have reason to believe that one will always be bigger than the other). The p-value on the left (0.0005) means the nifedipine mean is smaller than the propanolol mean (you can tell this from the line above that says the null is defined as the nifedipine mean minus the propanolol mean). The p-value on the right (0.9995) is for the one-sided test that the nifedipine mean is greater than the propanolol mean.

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Nifedipi |      18    150.8333    8.027017    34.05575    133.8978    167.7689
Proprano |      15    153.4667    8.591173    33.27347    135.0404    171.8929
---------+--------------------------------------------------------------------
combined |      33    152.0303    5.779484    33.20061    140.2579    163.8027
---------+--------------------------------------------------------------------
    diff |           -2.633333    11.78327               -26.66546     21.3988
------------------------------------------------------------------------------
    diff = mean(Nifedipi) - mean(Proprano)                        t =  -0.2235
Ho: diff = 0                                     degrees of freedom =       31

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.4123         Pr(|T| > |t|) = 0.8246          Pr(T > t) = 0.5877
```

**Test of Variance Equality**

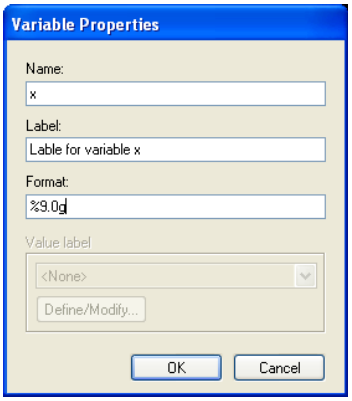
Note that we assumed that the variance was the same in both the nifedipine and propanolol groups when doing the t-test above. Stata can test the null hypothesis that the variances are equal by going to **Statistics > Summaries, tables, and tests > Classical tests of hypotheses > Two group variance comparison test.** Enter the name of the variable in the *vatiable* box (baseline in this case) and the grouping variable in the *group* box (med in this case). The output below shows that we do not have enough evidence to say that the variances are not equal.

```
. sdtest baseline, by(med)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Nifedipi |      18    150.8333    8.027017    34.05575    133.8978    167.7689
Proprano |      15    153.4667    8.591173    33.27347    135.0404    171.8929
---------+--------------------------------------------------------------------
combined |      33    152.0303    5.779484    33.20061    140.2579    163.8027
------------------------------------------------------------------------------
    ratio = sd(Nifedipi) / sd(Proprano)                           f =   1.0476
Ho: ratio = 1                                    degrees of freedom =   17, 14

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.5291         2*Pr(F > f) = 0.9419           Pr(F > f) = 0.4709
```

## <u>Some Notes on Getting Data into Stata</u>

***To Get an Excel Speadsheet into Stata:*** Reading data from an Excel spreadsheet is straightforward. Make sure that the variables are in the columns with each variable name at the top of each column. Highlight the data – including the first row of variable names – and copy to the clipboard. Open the data editor in Stata by clicking on the ▦ button. You will see an empty spreadsheet (if not then you already have a Stata file open). Right-click on the upper left cell and select “paste.” You can click on the “x” in the upper-left corner to exit. The data will be preserved in the temporary memory, but will not be saved until you choose **File > Save As** from the menu and enter the name under which you wish to save it.

***Entering Data in Stata by Hand:*** Open the data editor in Stata by clicking on the ▦ button. Click on the upper-left cell of the empty spreadsheet and begin entering the data by putting each variable in its own column (each row is a separate observation). Stata names the variables *var1, var2, var3,* etc. by default. To give the variables more meaningful names, double-click on the column you want to change and "Variable Properties" box will appear. You can also enter a longer, more descriptive, label for the variable in the box.



Variable Properties

Name: x

Label: Lable for variable x

Format: %9.0g

Value label

<None>

Note that Stata also decides what format the variable should be in. This can take on many different forms and it is usually best to let it default.

***Changing Data Values Once They are Entered:*** If you want to change any numbers that are in your data set, open the editor and click on the cell you want to change. Simply enter the new number and press enter. When you close the editor window, Stata will ask if you want to preserve the changes. If you click "accept changes" your change(s) will be applied. At this point, only the data in the local memory have been changed. If you want to save the data set to the disk then you must select **File > Save As**. If you click "discard changes" no changes will be made.

***Labeling Categorical Variables:*** If you have one or more categorical variables (variables that take only a finite number of distinct categories) you may want to label the values in them. For example, if your dataset includes a variable called "treetype" that can take on two values (1 for deciduous and 2 for coniferous), then you will see a series of 1s and 2s in the data set. To give these numbers meaningful labels you can select **Data > Labels > Label values > Define or modify label values**. Click on "Define" to define a new label and give the label a name. A new box will pop up asking for you to give the numerical value of one of the categories. It will also ask you to give the label you wish to assign to that numerical value. Keep entering labels until you have exhausted all the categories, then click "cancel." To actually assign the labels you just created to a variable, select **Data > Labels > Label values > Assign value labels to variable**. Select the label name you just created and the variable you wish to apply them to and hit enter**.**