SOC 3811/5811:
BASIC SOCIAL STATISTICS

Introduction to STATA

Vocabulary

Reality → Data → Knowledge

Data file

The set of numeric values for each variable and for each observation

For nominal and ordinal variables, categories must be assigned (ultimately arbitrary) numeric values

Documentation (or **Codebook**)

A description of how units were sampled from the population, how variables are defined and coded, how measurements were made for each variable, how the data file is organized, etc.

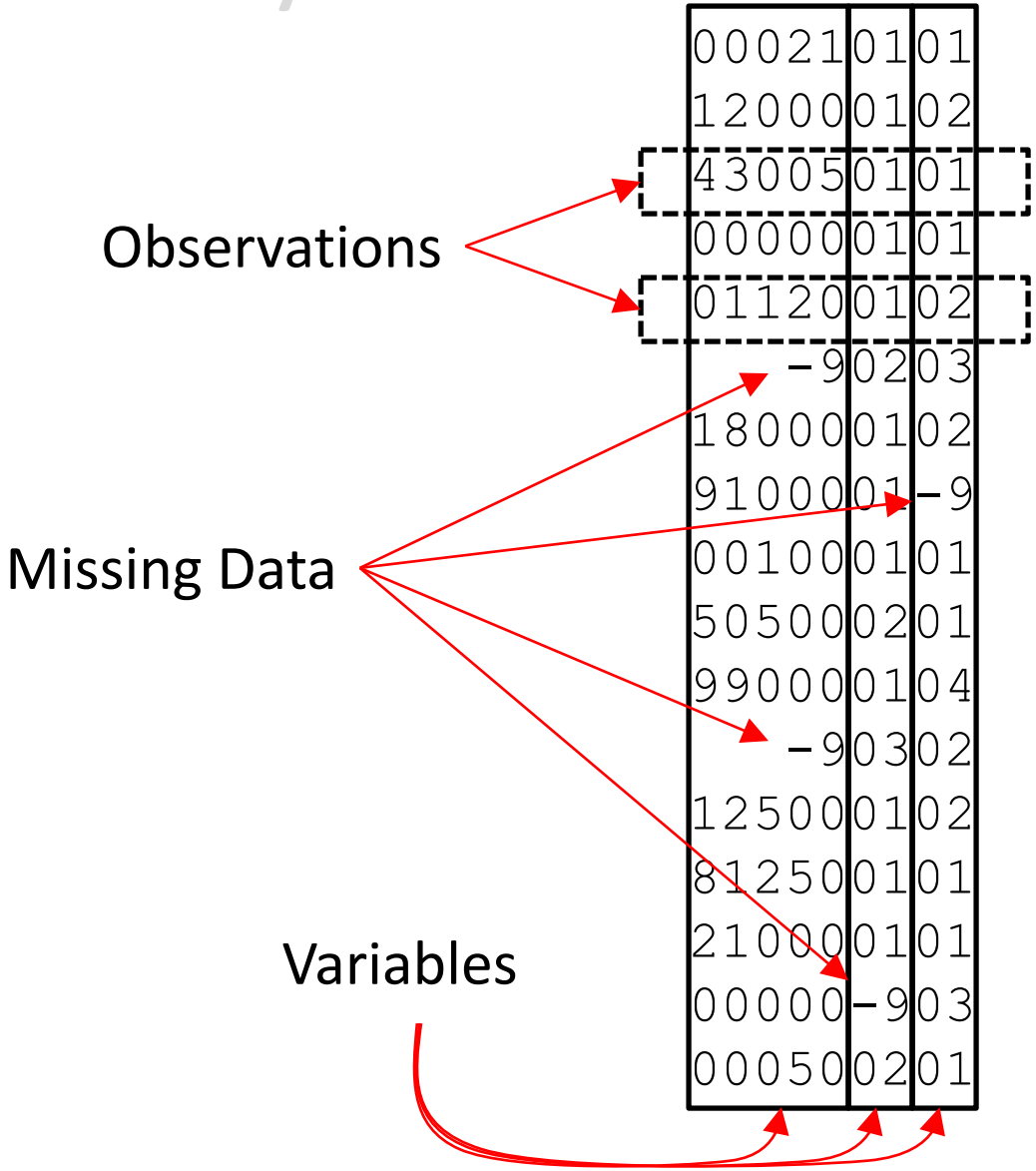
Missing data

A situation in which a numeric value is not available (for whatever reason) for a particular variable for a particular case

Reality → **Data** → Knowledge

```
000210101
120000102
430050101
000000101
011200102
   -90203
180000102
910000102
001000101
505000201
990000104
   -90302
125000102
812500101
210000101
00000-903
000500201
```

Reality → **Data** → Knowledge



Reality → Data → Knowledge

Some Questions for
the Documentation:

1. What do rows and columns represent?
2. How were individuals selected?
   (Sampling procedure; Population covered)
3. How were variables measured?
   (Continuous? Categorical? Ordinal?
   What do values mean? Name of variables?)
4. Missing data?
   (How generated? What do values mean?)
5. How many observations should there be?
6. Measures valid? Reliable?
7. More…

```
000210101
120000102
430050101
000000101
011200102
   -90203
180000102
910000102
001000101
505000201
990000104
   -90302
125000102
812500101
210000101
00000-903
000500201
```

Reality → Data → Knowledge

Some Questions for
the Documentation:

1. What do rows and columns represent?
2. How were individuals selected?
   (Sampling procedure; Population covered)
3. How were variables measured?
   (Continuous? Categorical? Ordinal?
   What do values mean? Name of variables?)
4. Missing data?
   (How generated? What do values mean?)
5. How many observations should there be?
6. Measures valid? Reliable?
7. More…

Female	Democrat	2
Female	Democrat	4
Female	Democrat	2
Male	Democrat	3
Male	Independent	5
Female	Independent	4
Female	Independent	2
Male	Independent	2
Male	Independent	5
Female	Independent	4
Transgender or a different identity	Independent	1
Female	Independent	2
Female	Independent	2
Female	Independent	3
Female	Independent	3
Male	Independent	4
Male	Independent	5
Male	Independent	3
Female	Independent	4
Female	Independent	4
Female	Independent	1
Female	Independent	1
Male	Independent	6
Female	Independent	1
Female	Independent	1
Female	Independent	2
Male	Republican	4
Female	Republican	5
Female	Republican	4
Female	Republican	4
Female	Republican	5
Female	Republican	5

Reality → Data → Knowledge

Statistical Software

Software that allows the researcher to read and manipulate a data file and to extract information from the data

Syntax (or Code)

Computer programming commands that instruct statistical software as to how to manipulate a data file and what statistical information to extract from the data

Output

Information produced as the result of applying syntax to a data file

Contents

Introduction to Stata

Sociology Department, 3/5/8-811

Tom VanHeuvelen

Department of Sociology

University of Minnesota

tvanheuv@umn.edu

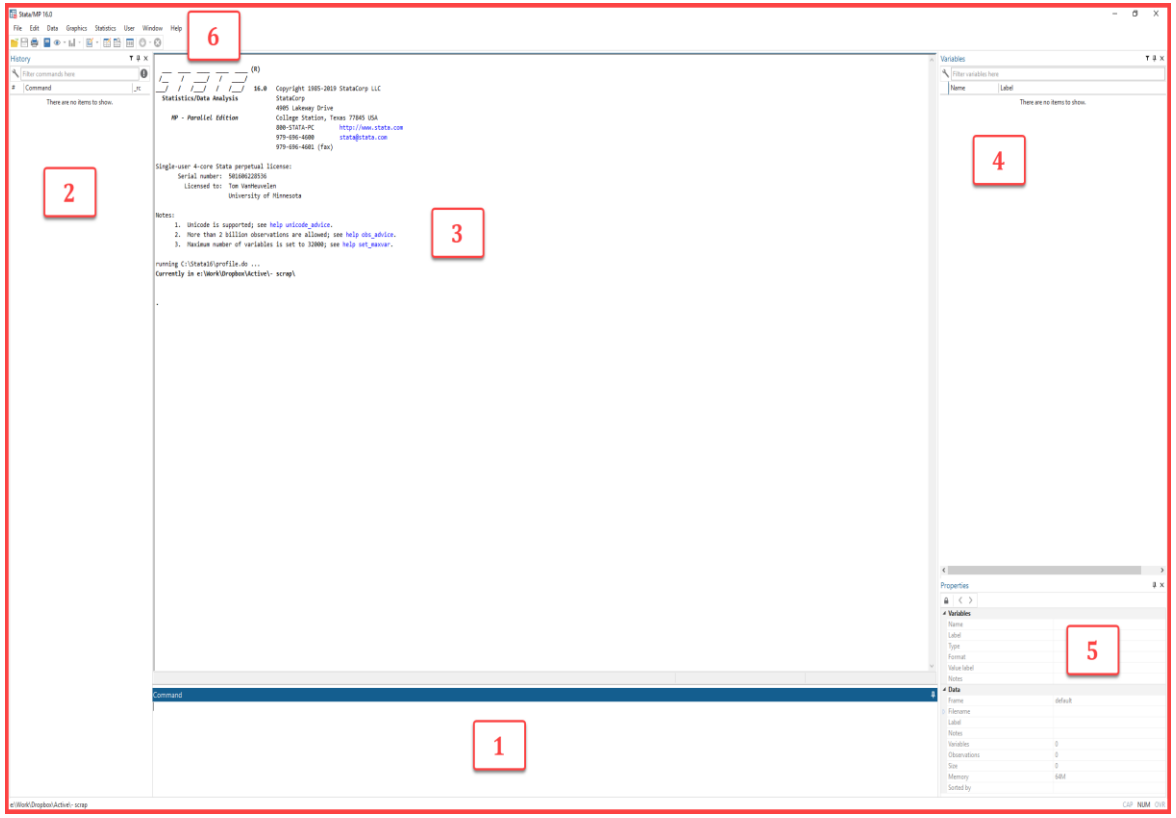
Overview of Topics

1. How to access Stata
2. Opening Stata for the first time
3. Locating and loading data
4. Look at, and summarize, your data

How to access Stata

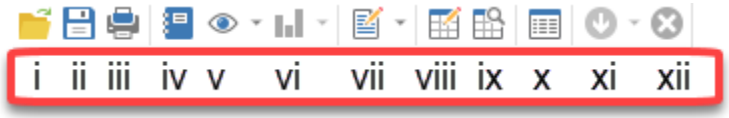
1. This may prove to be more challenging than in previous years. You need to access Stata without the luxury of a lab with a computer with Stata easily installed.
2. You have three main options to access Stata:
    a. AppsToGo
        i. Install Citrix Workspace
        ii. Link your Google Drive
    b. WTS server (graduate students)
    c. Purchase student copy
        i. https://www.stata.com/order/new/edu/gradplans/student-pricing
        ii. 6 month version can be purchased for $48
        iii. Perpetual copy can be purchased for $225
3. Instructions for accessing Stata via [a] and [b] are provide in the Access Instructions in this class' shared Google Drive.

Opening Stata for the first time



1. Command Window
2. Review Window
3. Results Window
4. Variable Window
5. Properties Window
6. Toolbar

The Toolbar



i. Open a dataset.
ii. Save the dataset you're working on.
iii. Print any of the files you have open: the dataset you're working on, do-file you have open, etc.
iv. Begin/Close/Suspend/Resume a Log (see next section).
v. Open the Viewer (you'll use this mainly to get help).
vi. Bring a graph to the front (you'll be able to choose from whatever graphs you have open).
vii. Open the do-file editor.
viii. Open the data editor. Here, you can edit the dataset. Use rarely.
ix. Browse the dataset. No editing capabilities.
x. Open the Variables Manager. Here, you can view and edit your variables' information.
xi. Prompts Stata to continue displaying output when the command fills the window. This has the same effect as entering a space into the Command Window.
xii. Stops the current command(s) from being estimated. Helpful when a model won't converge or you notice a mistake in your code.

Command Window

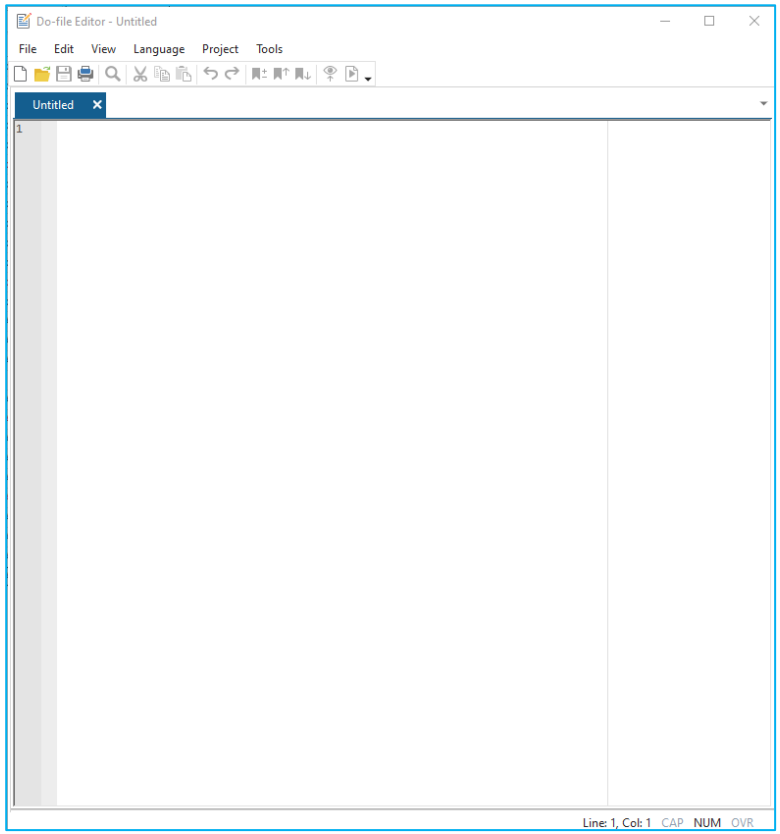
1. You can enter commands directly into the command window.
2. Benefits: quick, easy
3. Drawbacks: not robust, not systematic, not reproducible
    a. Sometimes, the slow way is the fast way.

```
. display "Greetings 5811 / 8811!!"
Greetings 5811 / 8811!!

. display 2+2
4

. display 5+1
6

. display 25*4 - 2
98
```

Command

```
display 1+1
```

Do-File Editor



1. Do-files allow you to save your work, and rerun your programs with ease. Use them EVERY TIME you wish to use Stata for a task.

2. I provided you a default do-file to use for every project. Let's take a look at it now.

a. Either click the do-file editor icon on the toolbar, or type `doedit` in the command window.

Looking at / Loading your data

1. You use Stata to assess and analyze data. So the first thing you need to do is locate your data!

2. All your materials for a project—data, do-files, documentation, etc.—should be kept in a single location on your computer that is dedicated to that project. So you may wish to create a folder titled “S5811” in Google Drive to house your work in this course.

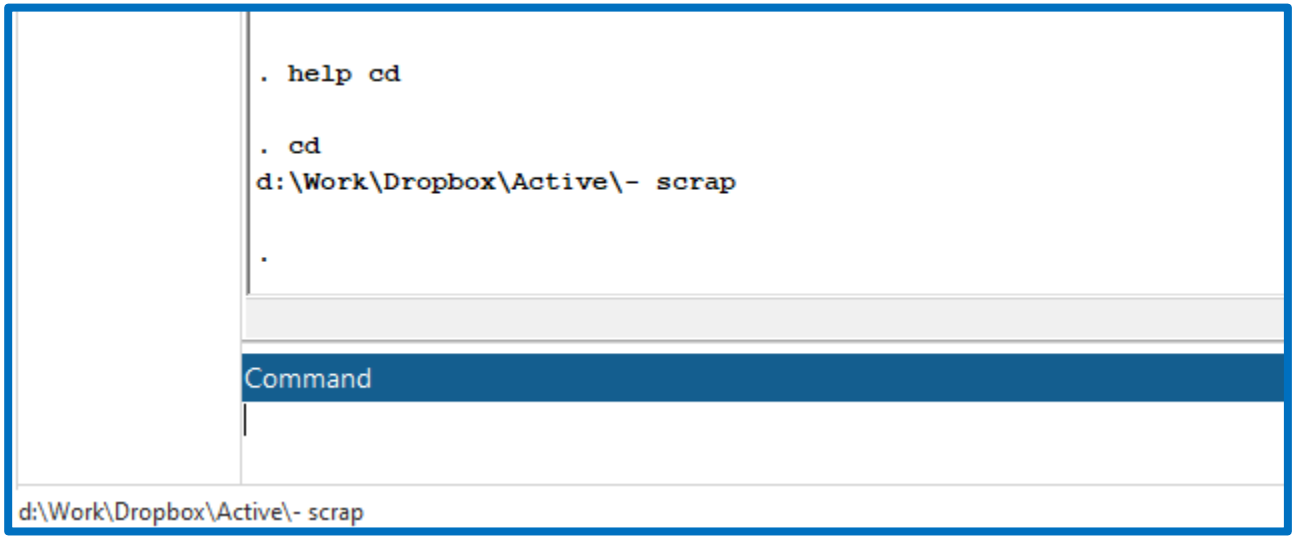
3. Stata operates using a **working directory**, or the place where Stata will look for files and the place that Stata will save new files it creates.
    a. The best way to operate is to have a specific working directory that is only used for a specific project. When you open Stata to work on this project, you set the working directory to this location.

Where you work, your working directory

1. Bottom left corner shows your current working directory.

2. You can change you working directory by typing `cd <path on your computer>`

3. You can also use a point-and-click menu from the Toolbar: File/Change Working Directory



Load your data

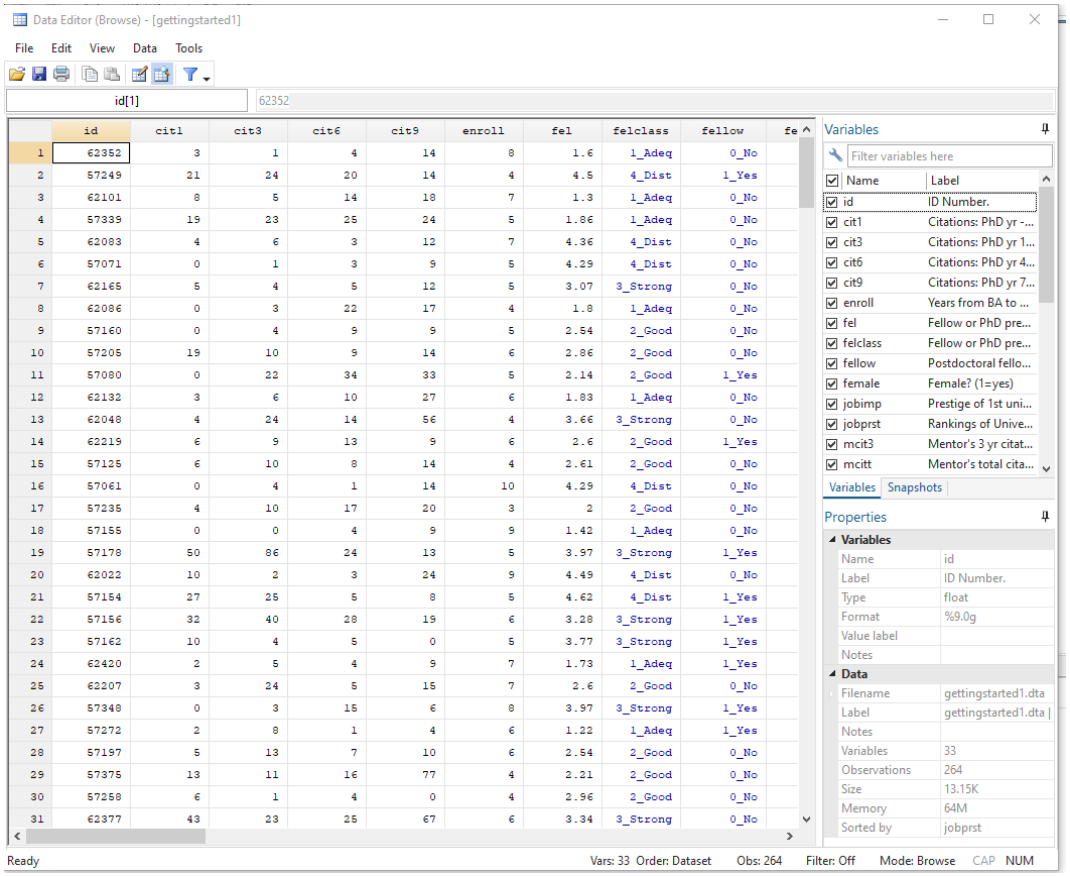
1. Now that you've set your working directory to where you house your project, let's open up our dataset.

2. In the do-file for this course, highlight line `use s8811-science4`. Either click "Execute do" in the top-right, or press control+D.

    a. You could similarly type `use s8811-science4` in the working directory.

Looking at your data-spreadsheet

If you type `browse` in the command window, you can look at your data in a spreadsheet format. The columns are your different variables, and the rows are observations in your data.



	id	cit1	cit3	cit6	cit9	enroll	fel	felclass	fellow	fe
1	62352	3	1	4	14	8	1.6	1_Adeq	0_No	
2	57249	21	24	20	14	4	4.5	4_Dist	1_Yes	
3	62101	8	5	14	18	7	1.3	1_Adeq	0_No	
4	57339	19	23	25	24	5	1.86	1_Adeq	0_No	
5	62083	4	6	3	12	7	4.36	4_Dist	0_No	
6	57071	0	1	3	9	5	4.29	4_Dist	0_No	
7	62165	5	4	5	12	5	3.07	3_Strong	0_No	
8	62086	0	3	22	17	4	1.8	1_Adeq	0_No	
9	57160	0	4	9	9	5	2.54	2_Good	0_No	
10	57205	19	10	9	14	6	2.86	2_Good	0_No	
11	57080	0	22	34	33	5	2.14	2_Good	1_Yes	
12	62132	3	6	10	27	6	1.83	1_Adeq	0_No	
13	62048	4	24	14	56	4	3.66	3_Strong	0_No	
14	62219	6	9	13	9	6	2.6	2_Good	1_Yes	
15	57125	6	10	8	14	4	2.61	2_Good	0_No	
16	57061	0	4	1	14	10	4.29	4_Dist	0_No	
17	57235	4	10	17	20	3	2	2_Good	0_No	
18	57155	0	0	4	9	9	1.42	1_Adeq	0_No	
19	57178	50	86	24	13	5	3.97	3_Strong	1_Yes	
20	62022	10	2	3	24	9	4.49	4_Dist	0_No	
21	57154	27	25	5	8	5	4.62	4_Dist	1_Yes	
22	57156	32	40	28	19	6	3.28	3_Strong	1_Yes	
23	57162	10	4	5	0	5	3.77	3_Strong	1_Yes	
24	62420	2	5	4	9	7	1.73	1_Adeq	1_Yes	
25	62207	3	24	5	15	7	2.6	2_Good	0_No	
26	57348	0	3	15	6	8	3.97	3_Strong	1_Yes	
27	57272	2	8	1	4	6	1.22	1_Adeq	1_Yes	
28	57197	5	13	7	10	6	2.54	2_Good	0_No	
29	57375	13	11	16	77	4	2.21	2_Good	0_No	
30	57258	6	1	4	0	4	2.96	2_Good	0_No	
31	62377	43	23	25	67	6	3.34	3_Strong	0_No	

Looking at your data-summarize

Often, you'll want to see summary statistics for your variables (e.g., means, minimum and maximum values). The **`summarize`** and **`codebook`**, compact commands are useful for this.

```
. summarize job phd pub1 female

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         job |        163    2.967117     .880396       1.01       4.69
         phd |        308    3.177987    1.012738          1       4.77
        pub1 |        308    2.545455    3.092685          0         24
      female |        308    .3474026    .4769198          0          1
```

Listing observations

Listing observations in your dataset is another way to explore the data.

```
. sort pubtot

. list id pubtot female phdclass felclass enrol in 1/5
```

	id	pubtot	female	phdclass	felclass	enroll
1.	57132	0	1_Yes	4_Dist	4_Dist	5
2.	57050	0	1_Yes	2_Good	2_Good	7
3.	57246	0	1_Yes	2_Good	2_Good	8
4.	57192	0	0_No	2_Good	2_Good	9
5.	57087	0	0_No	1_Adeq	1_Adeq	4

Variable distributions

For categorical variables, use the **`tabulate`** command. This command will allow you to tabulate one variable on its own, or cross-tabulate it with another:

```
. tabulate female, miss

    Female?
   (1=yes) |      Freq.     Percent        Cum.
-----------+-----------------------------------
      0_No |        201       65.26       65.26
     1_Yes |        107       34.74      100.00
-----------+-----------------------------------
     Total |        308      100.00

. tab phdclass female, m

  Prestige |
  class of |
     Ph.D. |     Female? (1=yes)
     dept. |       0_No      1_Yes |     Total
-----------+----------------------+----------
    1_Adeq |        33         12 |        45
    2_Good |        71         32 |       103
  3_Strong |        54         13 |        67
    4_Dist |        43         50 |        93
-----------+----------------------+----------
     Total |       201        107 |       308
```

Data management

Creating new variables

In each example, notice that the commands begin with **`gen [newvar] =`** . The command **`gen`** is simply shorthand for generate (*gen*erate); you may use either.

```
. gen totcit = cit1 + cit3 + cit6 + cit9
. list cit1 cit3 cit6 cit9 totcit in 1/5
```

	cit1	cit3	cit6	cit9	totcit
1.	3	4	1	5	13
2.	0	0	3	9	12
3.	0	4	1	5	10
4.	3	13	0	5	21
5.	3	3	8	14	28

```
. gen phdcat = phd
```

```
. recode phdcat (.=.) (1/1.99=1) (2/2.99=2) (3/3.99=3) (4/5=4)
(phdcat: 297 changes made)
```

Variable names and labels

When you generate new variables from existing ones, the variable and value labels do not transfer. You'll want to make sure you attach labels to the variable; otherwise analysis will be confusing.

Labeling the variables we've created:

```
. gen workres3 = workres
(6 missing values generated)

. label var totcit    "Total # of citations"
```

Stata assigns labels in two steps. In the first step, the command **`label define`** assigns labels to values. In the second step, the command **`label values`** is used to associate defined labels with one or more variables. Typically, you'll only apply value labels to categorical variables, although it is sometimes useful to indicate the meaning of high and low values of continuous variables.

```
. label define phdcat 1 "Adeq" 2 "Good" 3 "Strong" 4 "Dist"

. label value phdcat phdcat

. tab phdcat
```

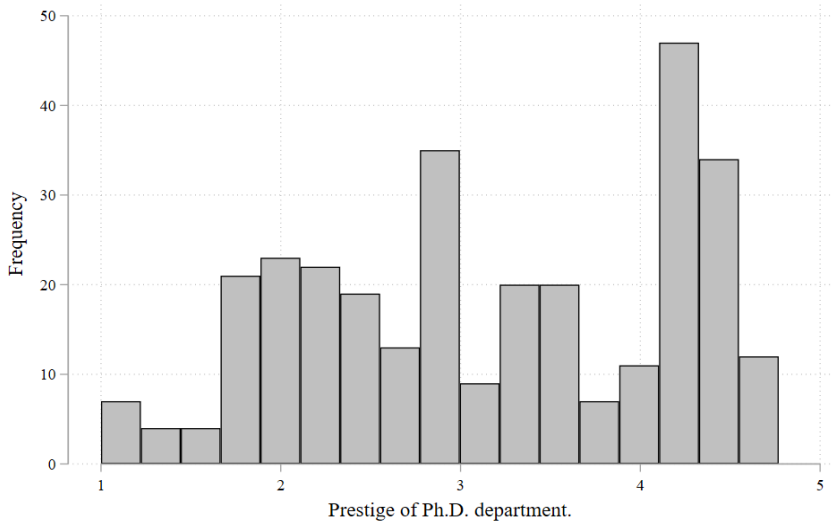
Phd Prestige: categories	Freq.	Percent	Cum.
Adeq	45	14.61	14.61
Good	103	33.44	48.05
Strong	67	21.75	69.81
Dist	93	30.19	100.00
Total	308	100.00	

Visualizing your data

Histograms

You often wish to visualize distributions of a single variable, or associations between two variables. For visual representation of categorical or continuous variables, histograms are a good way to go. The command is very simple:

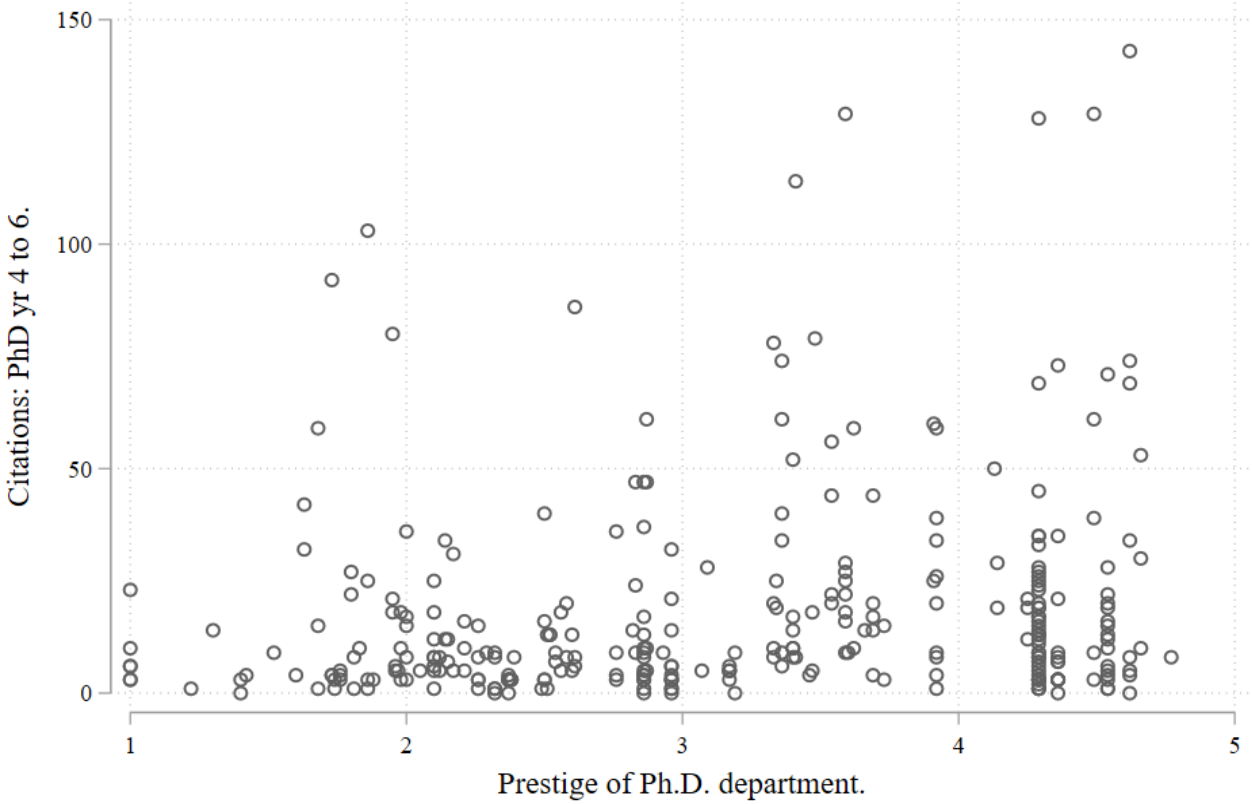
```
. histogram phd,freq
(bin=17, start=1, width=.2217647)

. graph export s8811-started_06.png, replace
(file s8811-started_06.png written in PNG format).
```



Scatterplots

1. In order to see the cross-distribution of two variables, you will need to use the scatter command:

```
. twoway scatter cit6 phd
```



A note on graphs

1. The options for graphs will become very complex in 3/5/8-811. As you become familiarized with creating graphs in Stata, I recommend two approaches:
    a. If you want to try out different options, it might be easier to use the point-and-click features of Stata for graphs. For example, selecting Graphics → Histogram brings up a dialog box you can use to customize your graph.
        i. Once you customize the graph the way you want it and submit the command, Stata will return the syntax for that command in the Results Window and produce the graph.
        ii. You can then copy the command syntax from the Results Window and paste it into your do-file.
    b. Templates: in 8811, you'll receive many example do-files that create graphs used in class and your assignments. One of the more effective methods of familiarizing yourself with Stata graphics is to use a template and build it from simple to complex, adding a single option to the command again and again.