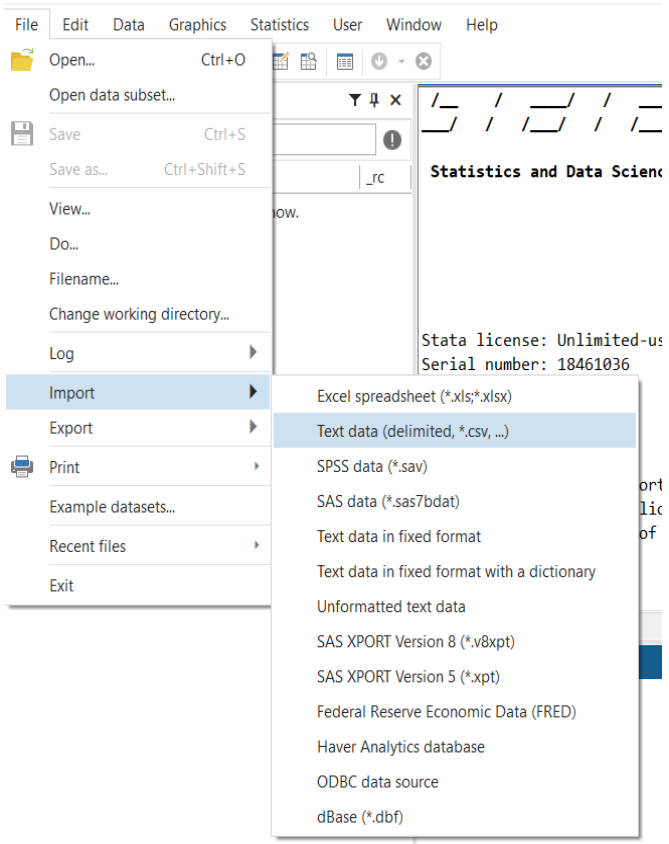
ECON 113 DI 02A/B

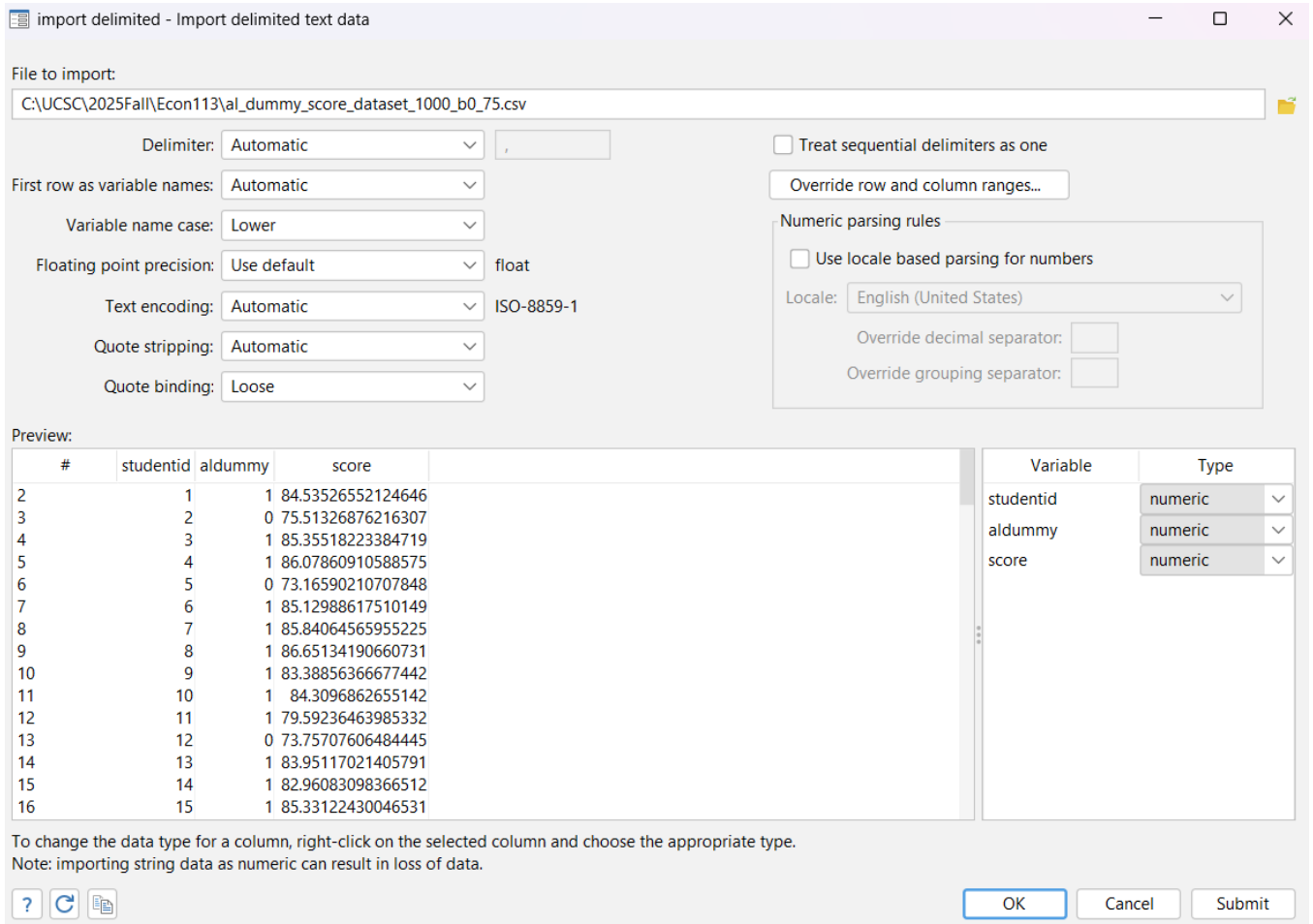
Dizhi Wang

OH: Thursday 8:30am – 10:30 am E2 403B

Stata

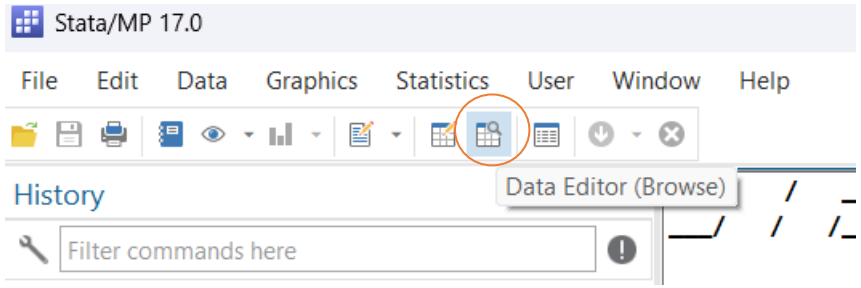
- Import data (csv)
  - Command:
    - *import delimited “path + file name",clear*
  - UI

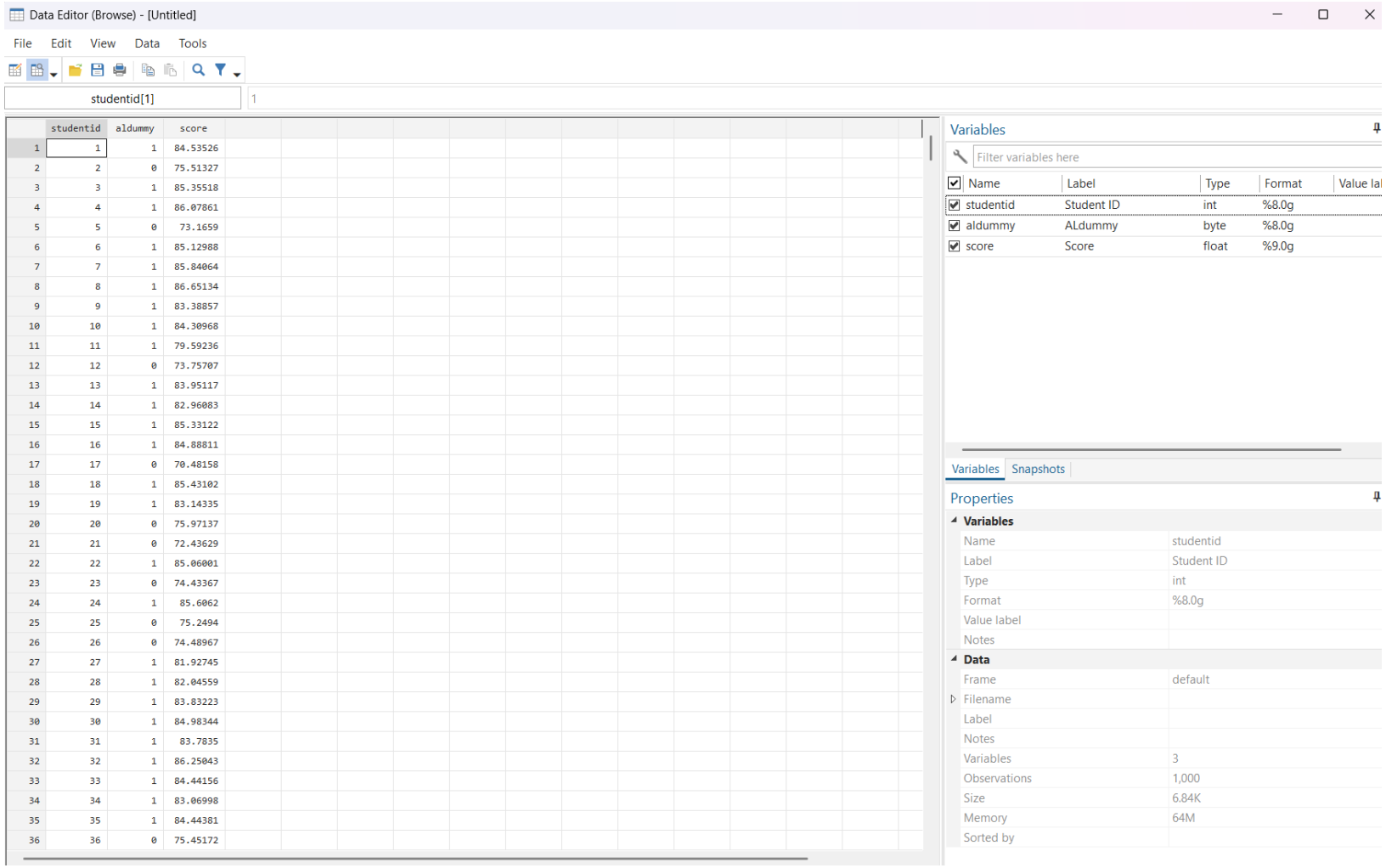




Stata

- Browse the data
  - Command:
    - *br*
    - Or *browse*
  - UI





Stata

- Describe the data
  - Command:
    - *desc*
    - Or *describe*

```
. desc

Contains data
 Observations:         1,000
    Variables:             3
```

Variable name	Storage type	Display format	Value label	Variable label
studentid	int	%8.0g		**Student ID**
aldummy	byte	%8.0g		**ALdummy**
score	float	%9.0g		**Score**

Stata

- Describe the data
  - Command:
    - *desc*
    - Or *describe*

```
. desc

Contains data
  Observations:         1,000
     Variables:             3
```

Variable name	Storage type	Display format	Value label	Variable label
studentid	int	%8.0g		**Student ID**
aldummy	byte	%8.0g		**ALdummy**
score	float	%9.0g		**Score**

- Summarize the data
  - Command:
    - *su*
    - Or summarize

```
. su
```

Variable	Obs	Mean	Std. dev.	Min	Max
studentid	1,000	500.5	288.8194	1	1000
aldummy	1,000	.507	.5002012	0	1
score	1,000	80.07408	5.32846	69.77561	90.6289

Stata

- Describe the data
  - Command:
    - *desc*
    - Or *describe*

```
. desc

Contains data
  Observations:         1,000
     Variables:             3
```

Variable name	Storage type	Display format	Value label	Variable label
studentid	int	%8.0g		**Student ID**
aldummy	byte	%8.0g		**ALdummy**
score	float	%9.0g		**Score**

- Summarize the data
  - Command:
    - *su*
    - Or summarize

```
. su
```

Variable	Obs	Mean	Std. dev.	Min	Max
studentid	1,000	500.5	288.8194	1	1000
aldummy	1,000	.507	.5002012	0	1
score	1,000	80.07408	5.32846	69.77561	90.6289

- Conditionally summarize the data
  - Command
    - *su if aldummy==1*

```
. su if aldummy == 1
```

Variable	Obs	Mean	Std. dev.	Min	Max
studentid	507	483.7357	294.7081	1	999
aldummy	507	1	0	1	1
score	507	84.95103	1.951967	79.59236	90.6289

Stata Question

- How can I see the average score for students who are NOT treated by active learning?
  - Hint: using conditional summarize

Stata Question

- How can I see the average score for students who are NOT treated by active learning?
  - Hint: using conditional summarize

```
. su if aldummy == 0
```

Variable	Obs	Mean	Std. dev.	Min	Max
studentid	493	517.7404	281.895	2	1000
aldummy	493	0	0	0	0
score	493	75.05864	2.004117	69.77561	81.9081

```
. su score if aldummy == 0
```

Variable	Obs	Mean	Std. dev.	Min	Max
score	493	75.05864	2.004117	69.77561	81.9081

  - You can summarize just one variable

Stata

- Tabulate a variable
  - Giving a more detailed information of the variable than summarize
  - Command:
    - tabulate or tab

```
. su aldummy
```

Variable	Obs	Mean	Std. dev.	Min	Max
aldummy	1,000	.507	.5002012	0	1

```
. tab aldummy
```

ALdummy	Freq.	Percent	Cum.
0	493	49.30	49.30
1	507	50.70	100.00
Total	1,000	100.00	

Question: How can I see the average score for students who are treated AND not treated by active learning using one command (hint: using tab)?

Stata

- Tabulate a variable
  - Giving a more detailed distribution of the variable than summarize
  - Command:
    - tabulate or tab

```
. su aldummy
```

Variable	Obs	Mean	Std. dev.	Min	Max
aldummy	1,000	.507	.5002012	0	1

```
. tab aldummy
```

ALdummy	Freq.	Percent	Cum.
0	493	49.30	49.30
1	507	50.70	100.00
Total	1,000	100.00	

Question: How can I see the average score for students who are treated AND not treated by active learning at the same time?

```
tab aldummy, su(score)
```

	Summary of Score		
ALdummy	Mean	Std. dev.	Freq.
0	75.058641	2.0041171	493
1	84.951027	1.9519671	507
Total	80.074081	5.3284595	1,000

Stata

- Count the number of observations:
  - Command:
    - *count*

```
. count
  1,000
```

Stata

- Count the number of observations:
  - Command:
    - *count*

```
. count
  1,000
```

  - Warning:
    - To count all observations just use *count*
    - You can’t count a variable

```
. count aldummy
varlist not allowed
r(101);
```

Stata

- Count the number of observations:
  - Command:
    - *count*

```
. count
  1,000
```

  - Warning:
    - To count all observations just use *count*
    - You can't count a variable

```
. count aldummy
varlist not allowed
r(101);
```

- Conditionally count:
  - Command:
    - *count if*
    - Question: How many students received the AL treatment (aldummy == 1) using count command?

Stata

- Count the number of observations:
  - Command:
    - *count*

```
. count
  1,000
```

  - Warning:
    - To count all observations just use *count*
    - You can't count a variable

```
. count aldummy
varlist not allowed
r(101);
```

- Conditionally count:
  - Command:
    - *count if*
    - Question: How many students accepted the AL treatment (aldummy == 1)?

```
. count if aldummy == 1
  507
```

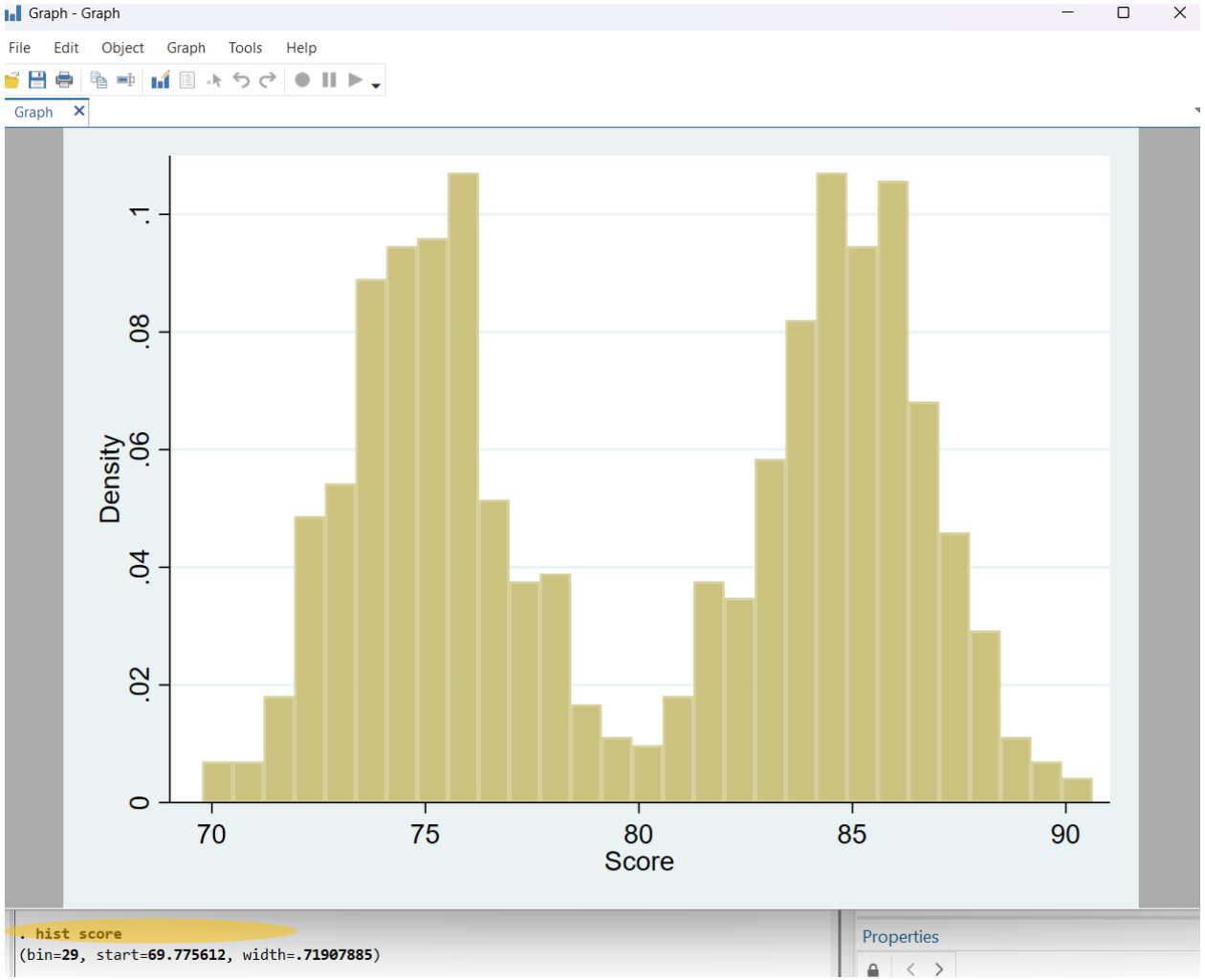
Histogram

- Histogram⇔ variables' distribution in sample
  - Command:
    - his
    - Or histogram

- Question: how to get the histogram for variable "score"

Histogram

- Histogram⇔ variables' distribution in sample
  - Command:
    - his
    - Or histogram



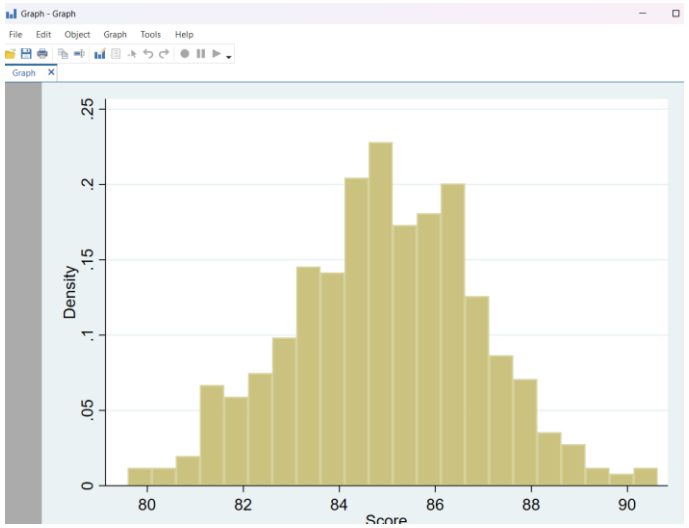
Histogram

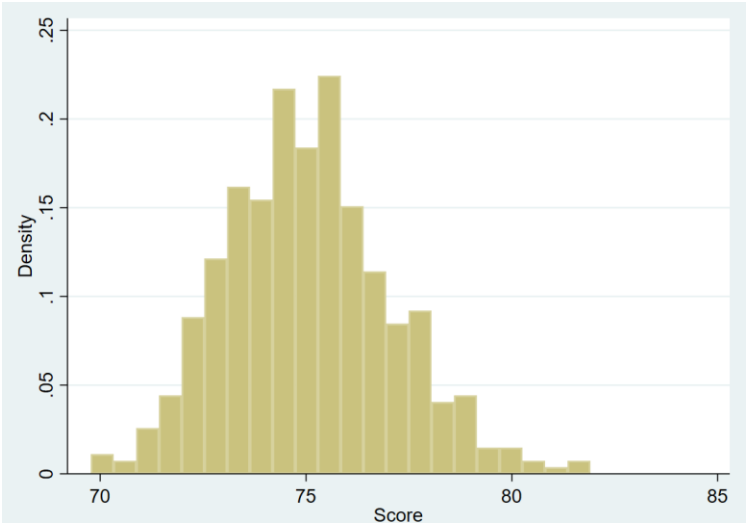
- Histogram⇔ variables' distribution in sample
  - Command:
    - his
    - Or histogram

- Conditional histogram:
  - Question:
    - What if I only want to see the histogram of score for students who got AL treatment?
    - What if I only want to see the histogram of score for students who DIDN'T get AL treatment?

Histogram

- Conditional histogram:
  - Question:
    - What if I only want to see the histogram of score for students who got AL treatment?
      - histogram score if aldummy == 1
    - What if I only want to see the histogram of score for students who DIDN'T get AL treatment?
      - histogram score if aldummy == 0





Population to Sample

- Population regression function (relationship)
  - $E[Y|X] = \beta_0 + \beta_1 X$
  - $\beta_1 = \frac{COV(X,Y)}{V(X)} = \frac{E[XY]-E[Y]E[X]}{E[X^2]-E[X]^2}$
- Sample regression function(relationship):
  - $\bar{y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{x}$
  - $\widehat{\beta_1} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2}$

Population to Sample

- Population regression function (relationship)
  - $E[Y|X] = \beta_0 + \beta_1 X$
  - $\beta_1 = \frac{COV(X,Y)}{V(X)} = \frac{E[XY]-E[Y]E[X]}{E[X^2]-E[X]^2}$
- Sample regression function(relationship):
  - $\bar{y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{x}$
  - $\widehat{\beta_1} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2}$
  - $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \sum_i \bar{x}\, y_i - \sum_i \bar{y}\, x_i - \sum_i \bar{x}\, \bar{y}$
  - $= \sum_i x_i y_i - \bar{x} \sum_i y_i - \bar{y} \sum_i x_i - \sum_i \bar{x}\, \bar{y}$

Population to Sample

- Population regression function (relationship)
  - $E[Y|X] = \beta_0 + \beta_1 X$
  - $\beta_1 = \frac{COV(X,Y)}{V(X)} = \frac{E[XY]-E[Y]E[X]}{E[X^2]-E[X]^2}$
- Sample regression function(relationship):
  - $\bar{y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{x}$
  - $\widehat{\beta_1} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2}$
  - $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \sum_i \bar{x}\, y_i - \sum_i \bar{y}\, x_i + \sum_i \bar{x}\, \bar{y}$
  - $= \sum_i x_i y_i - \bar{x} \sum_i y_i - \bar{y} \sum_i x_i + \sum_i \bar{x}\, \bar{y}$
  - Hint: $\sum_i \bar{x} = \sum_i x_i$
  - $= \sum_i x_i y_i - \sum_i \bar{x}\, \bar{y} - \sum_i \bar{y}\, \bar{x} + \sum_i \bar{x}\, \bar{y} = \sum_i x_i y_i - \sum_i \bar{x}\, \bar{y}$
  - Practice: do the same operations for the denominator

Regression using Stata

- Command:
  - reg
  - Or regression

```
. reg score aldummy
```

Source	SS	df	MS	Number of obs	=	1,000
				F(1, 998)	=	6252.75
Model	24460.0286	1	24460.0286	Prob > F	=	0.0000
Residual	3904.05974	998	3.91188351	R-squared	=	0.8624
				Adj R-squared	=	0.8622
Total	28364.0883	999	28.3924808	Root MSE	=	1.9778

score	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
aldummy	9.892386	.1251024	79.07	0.000	9.646892	10.13788
_cons	75.05864	.0890778	842.62	0.000	74.88384	75.23344

Regression≠Correlation

- Command:
  - corr
  - Or correlation

```
. corr aldummy score
(obs=1,000)

             |  aldummy    score
-------------+------------------
    aldummy  |   1.0000
      score  |   0.9286   1.0000


. corr aldummy score, cov
(obs=1,000)

             |  aldummy    score
-------------+------------------
    aldummy  |   .250201
      score  |   2.47509  28.3925
```

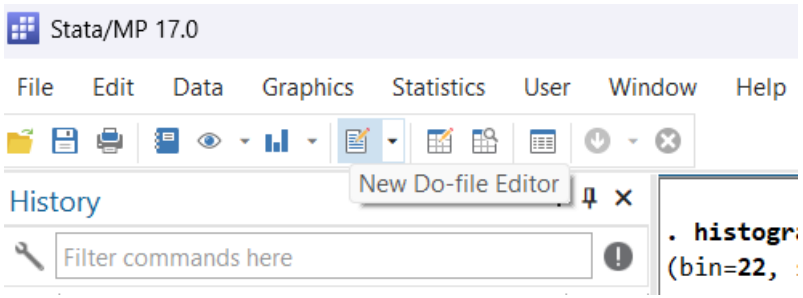
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

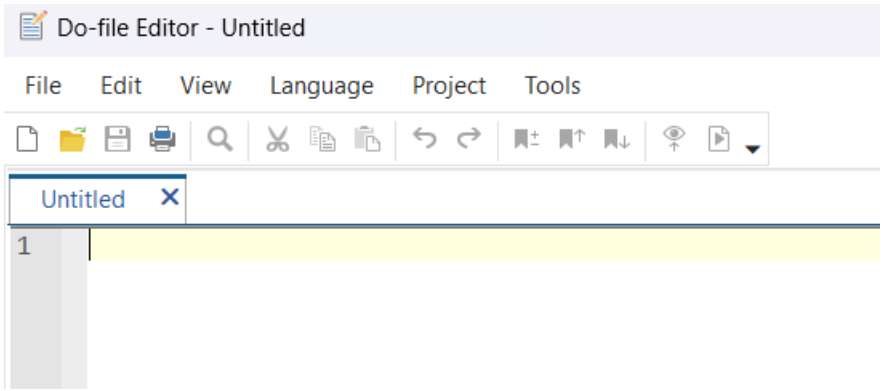
σ_X,σ_Y are the standard deviation of X and Y

$$\rho_{X,Y} \in [-1,1]$$

Do file

- Commands that is directly typed in the console will be lost
- Save them in the do file





- Copy and paste all the command we learned in the section to the do file
  - In the future you can directly write all of the command in the do file

Do file

ALScore.do

```
import delimited "C:\UCSC\2025Fall\Econ113\al_dummy_score_dataset_1000_b0_75.csv",clear

* Histogram for ALdummy = 0
histogram score if aldummy == 0, ///
    title("Histogram of Score (ALdummy = 0)") ///
    xtitle("Score") ytitle("Frequency") color(blue%60)

* Histogram for ALdummy = 1
histogram score if aldummy == 1, ///
    title("Histogram of Score (ALdummy = 1)") ///
    xtitle("Score") ytitle("Frequency") color(red%60)

* Histogram for all students
histogram score, ///
    title("Histogram of Score (All Students)") ///
    xtitle("Score") ytitle("Frequency") color(gs8)

* Count number of students with ALdummy = 1
count if aldummy == 1

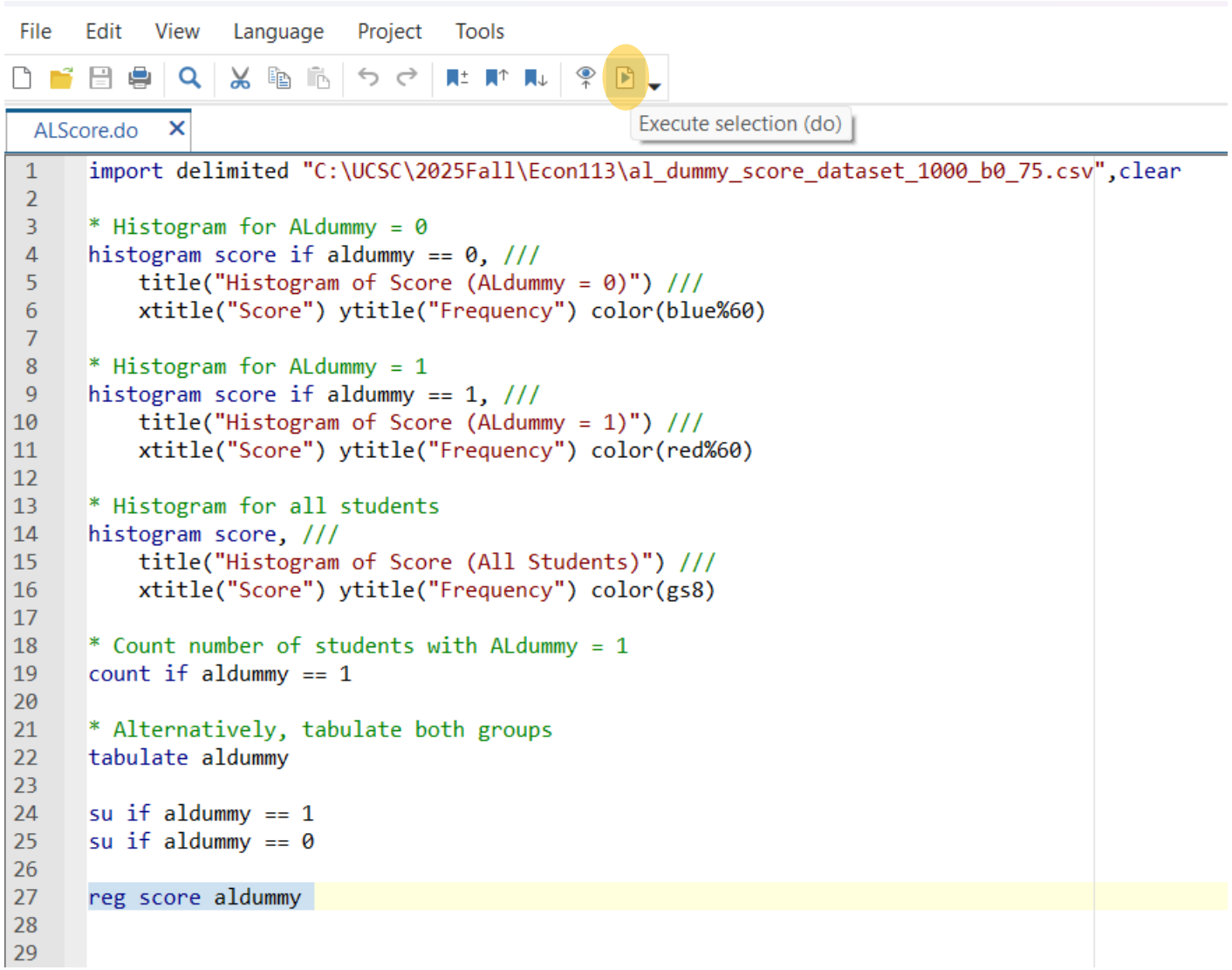
* Alternatively, tabulate both groups
tabulate aldummy

su if aldummy == 1
su if aldummy == 0

reg score aldummy


```

Execute the whole do file



Execute a section of the do file

- Select the section you want to execute
- Then press
  - Ctrl + D at the same time for windows
  - Command + shift + s at the same time for mac

```
import delimited "C:\UCSC\2025Fall\Econ113\al_dummy_score_dataset_1000_b0_75.csv",clear

* Histogram for ALdummy = 0
histogram score if aldummy == 0, ///
    title("Histogram of Score (ALdummy = 0)") ///
    xtitle("Score") ytitle("Frequency") color(blue%60)

* Histogram for ALdummy = 1
histogram score if aldummy == 1, ///
    title("Histogram of Score (ALdummy = 1)") ///
    xtitle("Score") ytitle("Frequency") color(red%60)

* Histogram for all students
histogram score, ///
    title("Histogram of Score (All Students)") ///
    xtitle("Score") ytitle("Frequency") color(gs8)

* Count number of students with ALdummy = 1
count if aldummy == 1

* Alternatively, tabulate both groups
tabulate aldummy

su if aldummy == 1
su if aldummy == 0

reg score aldummy
```