

Stata II

Econ 5/Poli 5D Lecture 6

## Announcements

- Second quiz due this Friday
- Homework 1 grades now available
- Second homework due in two weeks
- Any questions before we get started?

Today's Application: Intergenerational mobility and colleges in the United States

- Which colleges promote intergenerational mobility?

Today's Application: Intergenerational mobility and colleges in the United States

- Which colleges promote intergenerational mobility?

Today's Software: Stata

- Common functions used to manipulate data in Stata
- Making new variables in Stata
- Logical and relational operators in Stata
- Histograms and how to construct them

## Intergenerational Mobility by college

Higher education widely viewed as path for upward mobility and therefore often discussed as playing a large role in promoting intergenerational mobility

- Students from low-income backgrounds may move up the income distribution by going to college

However what if:

- Low-income students are not as likely to be admitted to colleges
- Low-income students do not particularly benefit from attending college

Today we will measure variation across colleges in the extent that they promote intergenerational mobility

## Definition of Intergenerational Mobility by college

$$\text{mobility rate} = \text{fraction of students from bottom quintile of earnings} \quad (1) \\ \times \text{fraction from bottom quintile that reach top quintile of earnings}$$

Intuition: If a college admits a large fraction of low-income students (the first term) and these students become high-earners (the second term), then this school promotes intergenerational income mobility.

So a college can promote intergenerational mobility by either (1) accepting a lot of students from low-income backgrounds or (2) being particularly effective in increasing incomes

## Opportunity Insights data

30 million college students from 1999-2013

Students linked to their colleges through (1) tax records and (2) Pell grant records from Department of labor

- Overall, data includes most individuals in the U.S. that attended college able to link most individuals to their college

Student income is measured in 2014

Parental income is average income when the student was between 15-19

We will use this data to generate two key variables

## Quintile of Parental income

Take all parents that have a child in the same year (e.g. 1985), also known as a cohort

Calculate the average income for this group when children are 15-19 (i.e. from 2000-2004)

If parents' income is in bottom 20 percent, then student is in the bottom quintile of parental earnings

This will be a key variable: we are interested in seeing how earnings of individuals from bottom quintile are impacted by college

## Quintile of Student income

Take all individuals that are the same age and calculate earnings in 2014

If an individual's income is in the top 20 percent out of everyone that is the same age, then this person is in the top quintile of earnings

This will be a key variable: we are interested in seeing if students from low-income backgrounds become high earners and how this varies across colleges

In [1]:

```
* Setup
cd "/Users/Brian/Dropbox/Grad School/Sixth Year/Econ:Poli 5/Lectures/Week 4"

* load data
use ./data/mrc.dta, replace
* describe dataset
describe
```

/Users/Brian/Dropbox/Grad School/Sixth Year/Econ:Poli 5/Lectures/Week 4

(Preferred Estimates of Access and Mobility Rates by College)

Contains data from ./data/mrc.dta

```
obs:          2,202          Preferred Estimates of Access
s
                                and Mobility Rates by College
vars:          7          21 Aug 2020 15:11
```

```
-----
-----
variable name      storage   display   value    variable label
                  type      format   label
-----
name               str141   %141s    Name of Institution / Super-
OPEID
state              str2    %9s      Cluster
on                State of constituent insti-
tution
par_median         double  %9.0g    with highest enrollment
k_median           float   %9.0g    Median parental income
Median kid income
```

par_q1	float	%9.0g	Fraction of parents in quintile 1
			(bottom quintile)
kq5_cond_parq1	float	%9.0g	Probability of kid in quintile 5 conditional on parent in quintile 1
count	double	%12.0g	Mean number of kids per cohort

-----

Sorted by:

## Generating new variables

Reminder: we are interested in mobility rates by college

`par_q1` gives fraction of parents in quintile 1 (low earnings)

`kq5_cond_parq1` gives probability of quintile 1 student moving to quintile 5

Together these allow us to measure the "mobility rate" of a given college

To do so, we need to use the `gen` command (short for generate)

## Syntax of generate command

```
gen newvar = exp
```

For example, if we have two variables `x1` and `x2`, we can create new variables that are functions of these two variables

- Addition: `gen sum_x1_x2 = x1 + x2`
- Subtraction: `gen diff_x1_x2 = x1 - x2`
- Multiplication: `gen mult_x1_x2 = x1*x2`
- Division: `gen div_x1_x2 = x1/x2`

In [2]:

```
* generate a new variable called mobility rate  
gen mobility_rate = par_q1*kq5_cond_parq1  
  
*summarize mobility rate  
sum mobility_rate
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
mobility_r~e	2,202	.0182738	.0131446	0	.1635797

Let's compare mobility rates of colleges in California vs. all other states

To do so we will create a binary variable that is equal to one if the college is in California and zero otherwise

- Binary because it only takes on two values (1 or 0)
- Sometimes referred to as indicator or dummy variable

Before we do so, let's look at the state variable

Before we do so, let's look at the state variable

```
In [3]: sum state
```

Variable	Obs	Mean	Std. Dev.	Min	Max
state	0				

No summary statistics because `state` is a **string** variable, not a numeric variable

No summary statistics because `state` is a **string** variable, not a numeric variable

In [4]:

```
%head state
```

	<b>state</b>
<b>1</b>	NY
<b>2</b>	NY
<b>3</b>	NY
<b>4</b>	NY
<b>5</b>	CA
<b>6</b>	NY
<b>7</b>	MA
<b>8</b>	NY
<b>9</b>	NY
<b>10</b>	NY

We want to create a variable equal to one if state is equal to CA

To do so we will again use the `gen` command

Useful to create many types of **relational** variables

- greater than: `gen newvar = (var>value)`
- greater than or equal to: `gen newvar = (var>=value)`
- less than: `gen newvar = (var<value)`
- less than or equal to: `gen newvar = (var<=value)`
- equals: `gen newvar = (var==value)`
- not equals to: `gen newvar = (var!=value)`

To create a variable equal to one if the college is in California we type:

To create a variable equal to one if the college is in California we type:

```
In [5]: gen CA = (state=="CA")
```

To create a variable equal to one if the college is in California we type:

```
In [5]: gen CA = (state=="CA")
```

A few lessons:

- If referencing values for string variables, use quotation marks
- If state had been a numeric code, for example 4=california, we would have typed `gen CA = (state==4)`

A useful way to describe binary (or categorical) variables is to use the `tab` command (short for tabulate)

Counts the number of observations in each cell defined by the binary (or categorical) variable (or the combination of multiple binary/categorical variables). For example:

A useful way to describe binary (or categorical) variables is to use the `tab` command (short for tabulate)

Counts the number of observations in each cell defined by the binary (or categorical) variable (or the combination of multiple binary/categorical variables). For example:

In [6]: `tab CA`

CA	Freq.	Percent	Cum.
0	2,034	92.37	92.37
1	168	7.63	100.00
Total	2,202	100.00	

A useful way to describe binary (or categorical) variables is to use the `tab` command (short for tabulate)

Counts the number of observations in each cell defined by the binary (or categorical) variable (or the combination of multiple binary/categorical variables). For example:

In [6]:

```
tab CA
```

CA	Freq.	Percent	Cum.
0	2,034	92.37	92.37
1	168	7.63	100.00
Total	2,202	100.00	

168 of the colleges in the dataset are in California.

Let's see how mobility rates compare across CA and other states

To do so we will make use of `if` statements

## IF statements in Stata

`if` option can be combined with most commands in Stata

Examples:

- `summarize if`
- `gen if`

In our example, we want to find the average mobility rate for California colleges and the average mobility rate colleges not in California

In [7]:

```
sum mobility_rate par_q1 kq5_cond_parq1 if CA==1  
sum mobility_rate par_q1 kq5_cond_parq1 if CA==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_r~e	168	.0275095	.0154819	0	.0991846
par_q1	168	.1443805	.0889914	.0321324	.4606968
kq5_cond_p~1	168	.2481358	.1638556	0	.8497473

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_r~e	2,034	.017511	.0126387	0	.1635797
par_q1	2,034	.1234973	.0880145	.0111896	.6097748
kq5_cond_p~1	2,034	.1917753	.1360376	0	.9192932

In [7]:

```
sum mobility_rate par_q1 kq5_cond_parq1 if CA==1
sum mobility_rate par_q1 kq5_cond_parq1 if CA==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_rate	168	.0275095	.0154819	0	.0991846
par_q1	168	.1443805	.0889914	.0321324	.4606968
kq5_cond_parq1	168	.2481358	.1638556	0	.8497473

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_rate	2,034	.017511	.0126387	0	.1635797
par_q1	2,034	.1234973	.0880145	.0111896	.6097748
kq5_cond_parq1	2,034	.1917753	.1360376	0	.9192932

## Interpretation

- Mobility rates are higher in CA (0.027 vs 0.17)
- This is driven by two factors
  - More students from low-income backgrounds (0.14 vs. 0.12)
  - Low-income students more likely to become high-income (0.25 vs. 0.19)

## Logical Operators in Stata

Suppose we want to know the mobility rate in both Oregon and Washington

We could combine our `if` statement with a logical `or`

Logical operators in Stata

- And: `&`
- Or: `|`
- Not: `!`

In [8]:

```
gen OR = (state=="OR")
gen WA = (state=="WA")

sum mobility_rate par_q1 kq5_cond_parq1 if OR==1|WA==1
sum mobility_rate par_q1 kq5_cond_parq1 if !(OR==1|WA==1)
sum mobility_rate par_q1 kq5_cond_parq1 if OR!=1&WA!=1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_rate	88	.0148811	.0055342	.0047164	.0352217
par_q1	88	.095565	.0511086	.0205435	.2776206
kq5_cond_parq1	88	.1927939	.1007786	.052114	.5241945

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_rate	2,114	.0184151	.0133497	0	.1635797
par_q1	2,114	.1263196	.0892515	.0111896	.6097748
kq5_cond_parq1	2,114	.1962118	.1405069	0	.9192932

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_rate	2,114	.0184151	.0133497	0	.1635797
par_q1	2,114	.1263196	.0892515	.0111896	.6097748
kq5_cond_parq1	2,114	.1962118	.1405069	0	.9192932

## **Interpretation**

- Mobility rates are lower in OR and WA (0.015 vs 0.18)
- This is mostly because of fewer students from low-income backgrounds
- Notice the two (equivalent) ways of calculating the other 48 states

Now let's look at the data for UCSD in particular!

## **Interpretation**

- Mobility rates are lower in OR and WA (0.015 vs 0.18)
- This is mostly because of fewer students from low-income backgrounds
- Notice the two (equivalent) ways of calculating the other 48 states

Now let's look at the data for UCSD in particular!

```
In [9]: sum mobility_rate par_q1 kq5_cond_parq1 if name == "University Of California, San Diego"
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
mobility_r~e	1	.0483275	.	.0483275	.0483275
par_q1	1	.0877724	.	.0877724	.0877724
kq5_cond_p~1	1	.5506001	.	.5506001	.5506001

```
In [9]: sum mobility_rate par_q1 kq5_cond_parq1 if name == "University Of California, San Diego"
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mobility_rate	1	.0483275	.	.0483275	.0483275
par_q1	1	.0877724	.	.0877724	.0877724
kq5_cond_parq1	1	.5506001	.	.5506001	.5506001

### Interpretation

- Mobility rates are quite a bit higher at UCSD relative to CA
  - Fewer students come from low-income backgrounds (0.09 vs. 0.14)
  - But those that do are very likely to become high earners (0.55 vs. 0.25)
- Overall, this latter effect dominates, leading to high mobility rates for UCSD

## Histograms

We now know that CA tends to have high mobility rates relative to the rest of the country and UCSD has high mobility rates relative to the rest of CA

A useful way to visualize the data is through a histogram

We will start with a simplified example and then see how this helps us understand variability across colleges in terms of mobility rates

## Histogram example

We want to understand the distribution of test scores within a class

For example, take the following example of a class with 15 students

	student	score
1	1	65
2	2	70
3	3	76
4	4	76
5	5	80
6	6	81
7	7	82
8	8	83
9	9	85
10	10	85
11	11	88
12	12	89
13	13	89
14	14	91
15	15	96

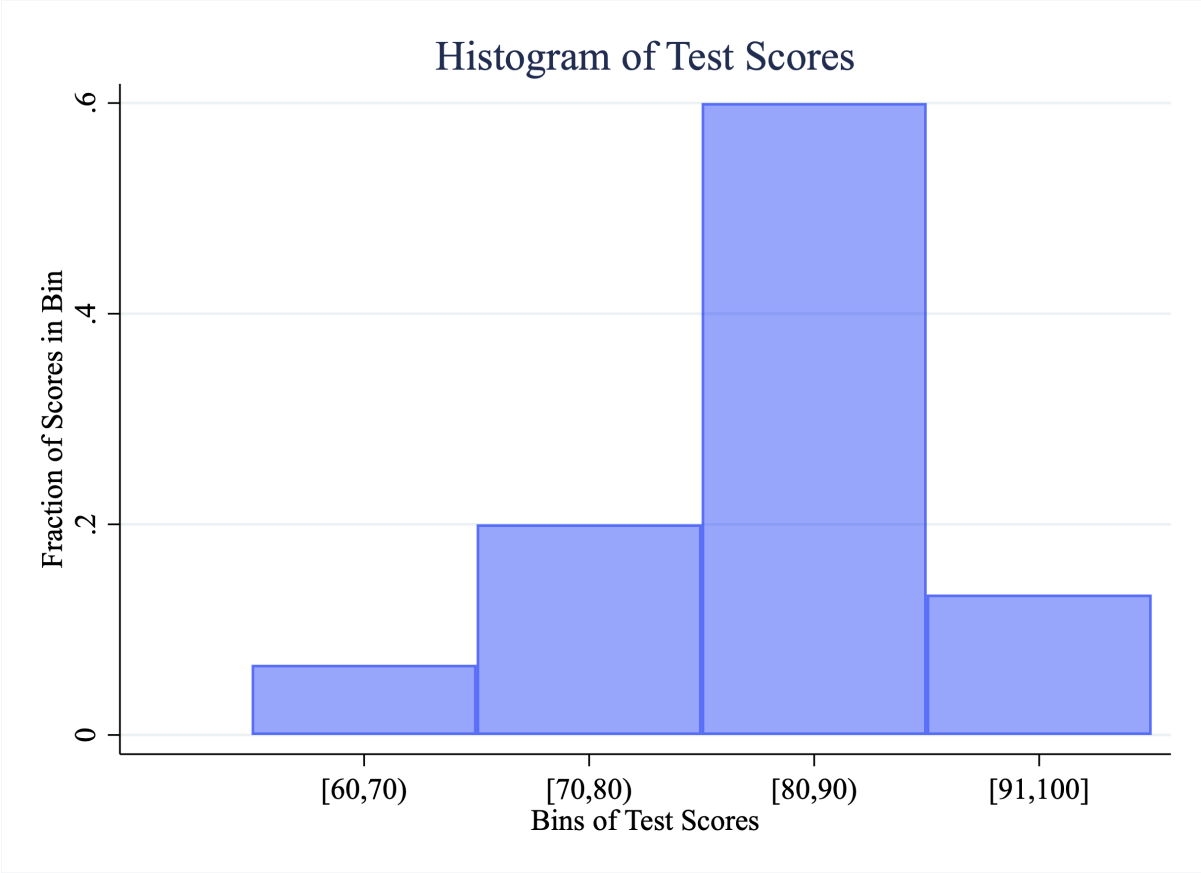
A **histogram** bins the data and then counts the frequency within each bin

There is no "right" way to bin a continuous variable

A **histogram** bins the data and then counts the frequency within each bin

There is no "right" way to bin a continuous variable

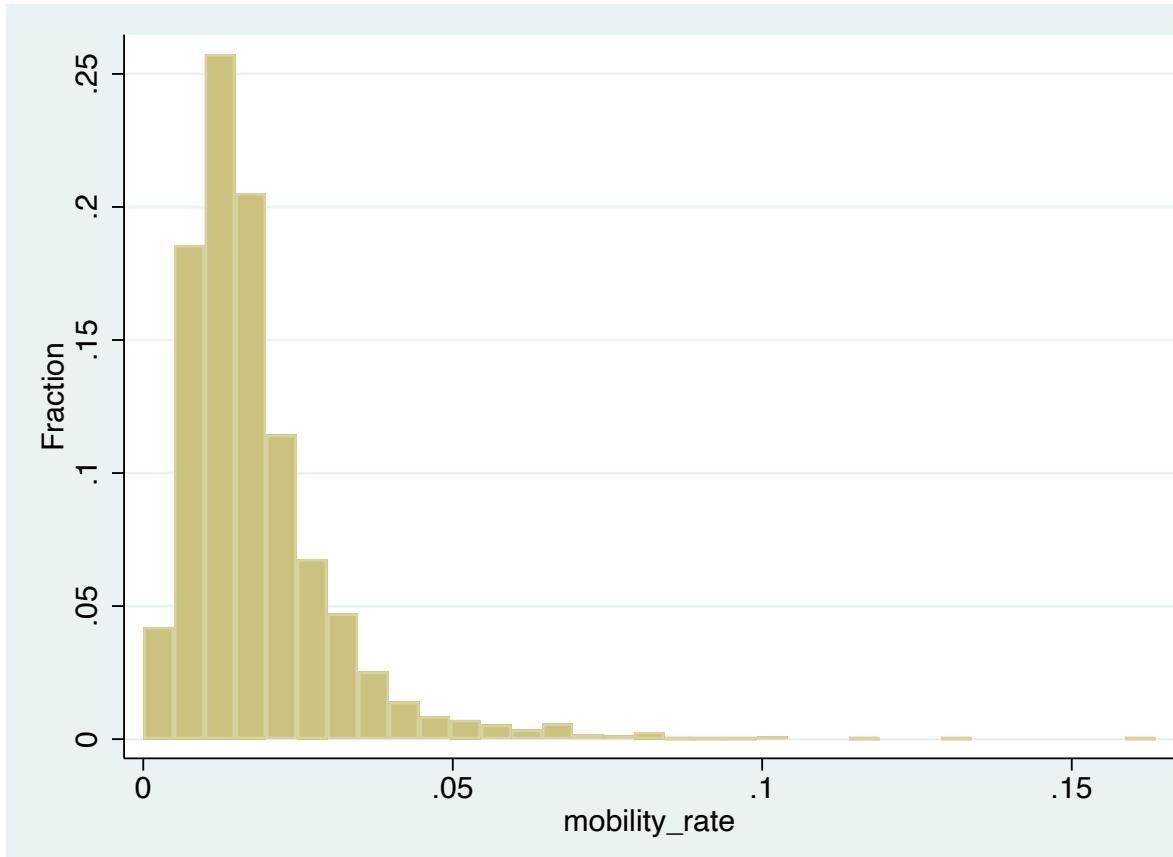
bin	frequency	fraction
[60,70)	1	.0666667
[70,80)	3	.2
[80,90)	9	.6
[90,100]	2	.1333333



So now let's go back and plot a histogram for mobility rates across colleges

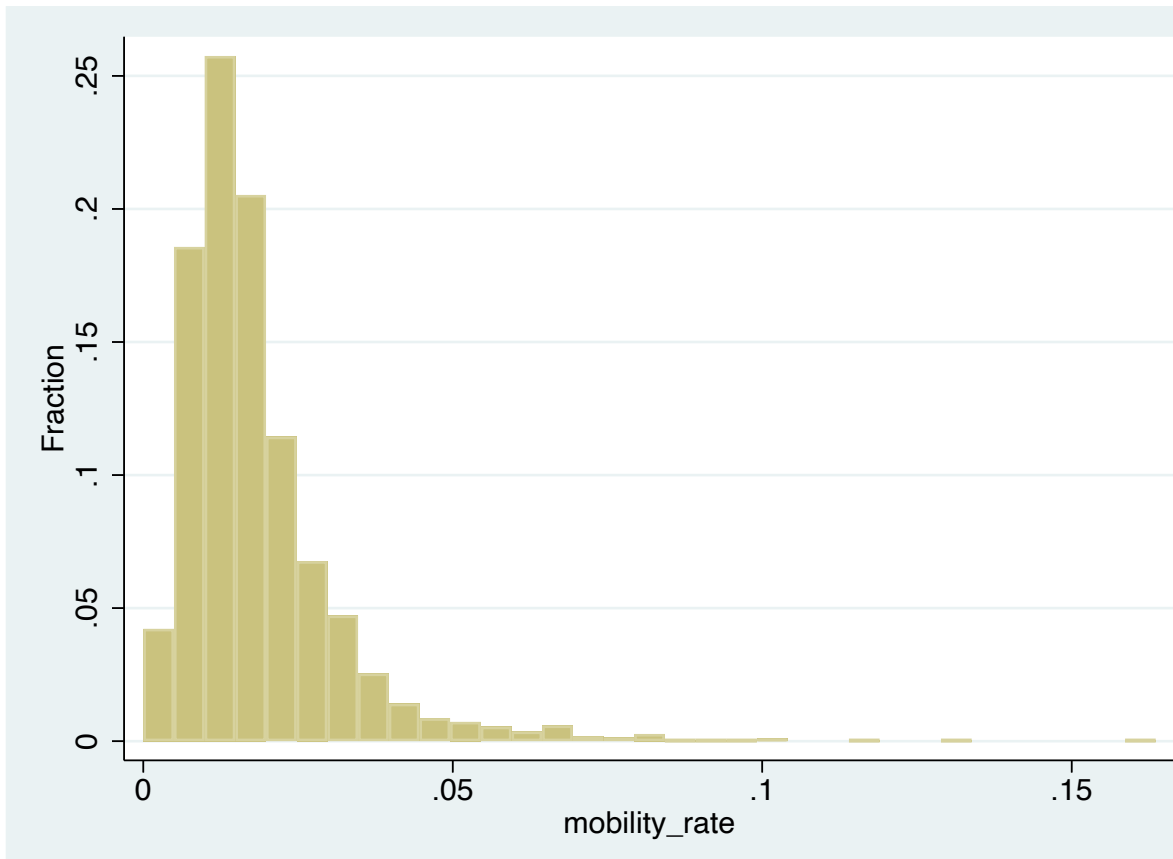
So now let's go back and plot a histogram for mobility rates across colleges

```
In [10]: graph twoway histogram mobility_rate, frac
```



So now let's go back and plot a histogram for mobility rates across colleges

```
In [10]: graph twoway histogram mobility_rate, frac
```

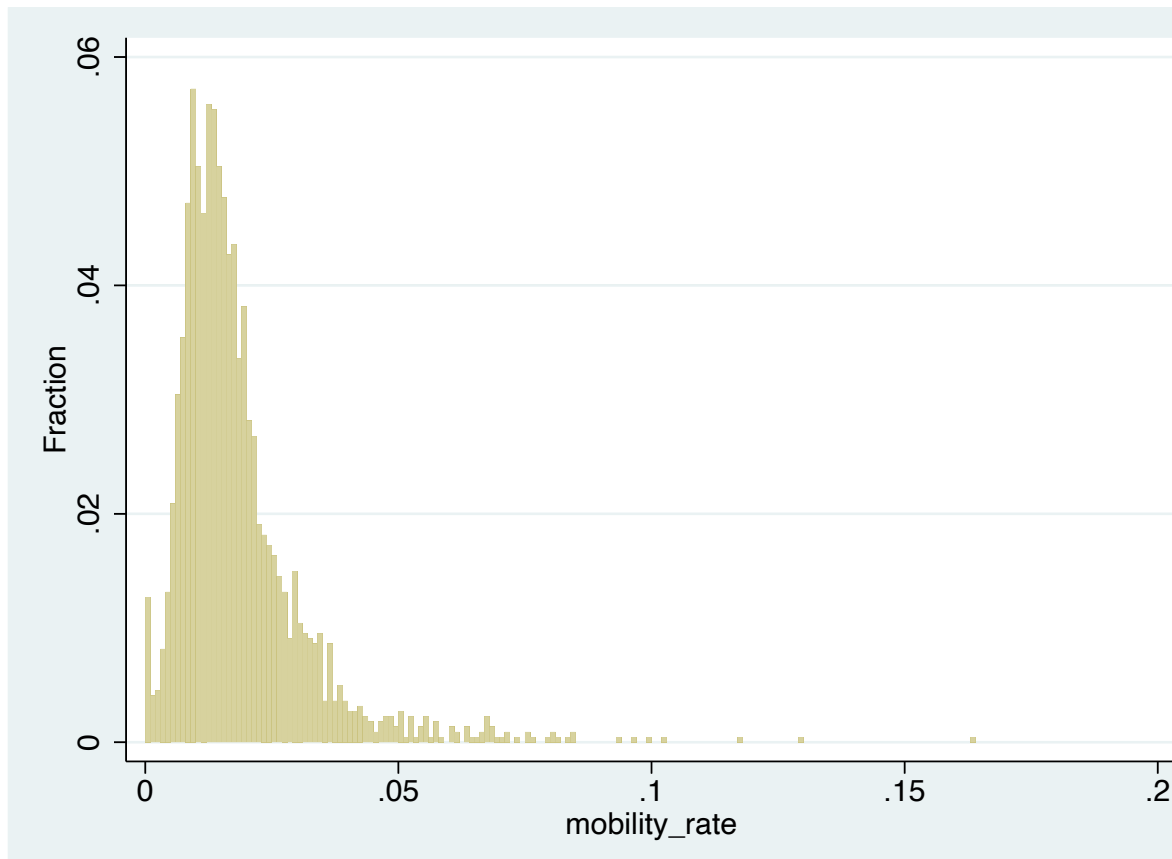


Interestingly, mobility has a "long right tail". While most mobility rates cluster in the 0.02-0.04 range, we have a few colleges with extremely high mobility rates relative to other colleges

We can customize our graph with more options. Type `help histogram` to see these. One option is to specify the `width()` of each bin.

We can customize our graph with more options. Type `help histogram` to see these. One option is to specify the `width()` of each bin.

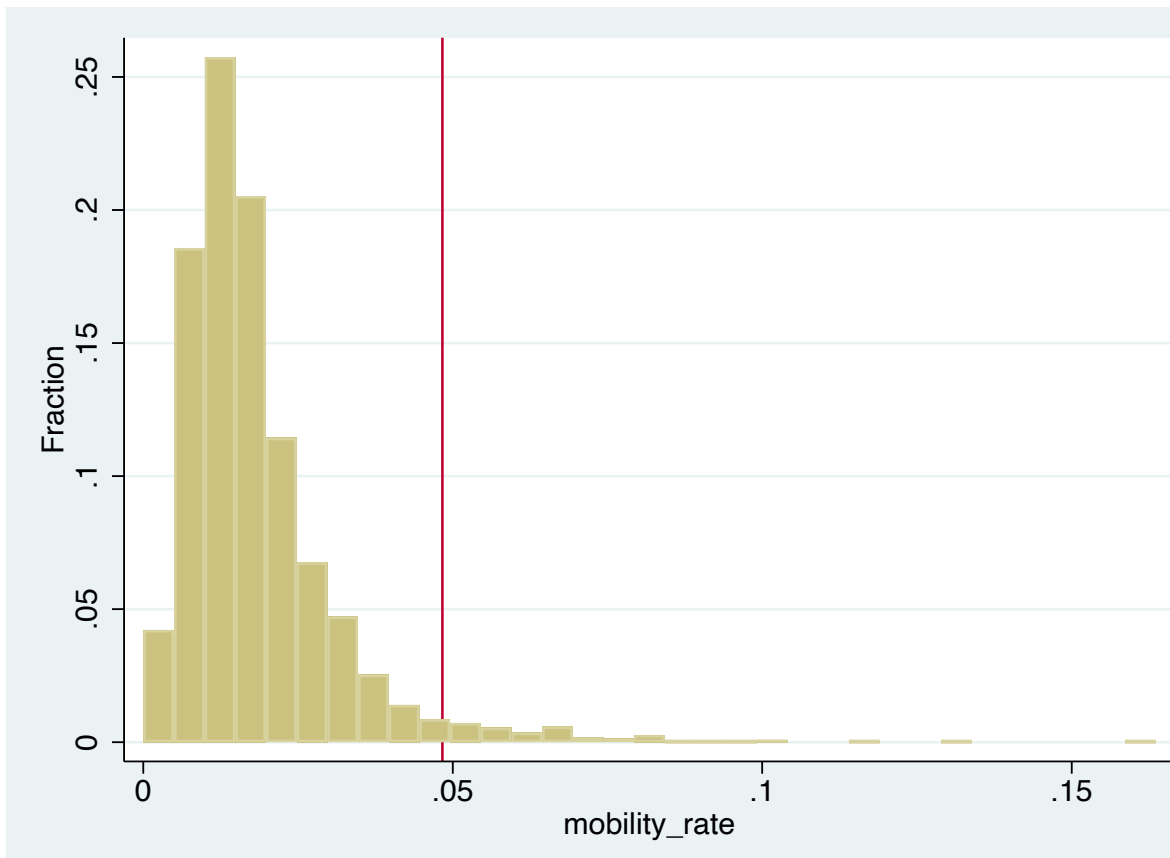
```
In [11]: graph twoway histogram mobility_rate, frac width(0.001)
```



Next, because we are from UCSD, let's add a vertical line for UCSD to the graph to see how we compare against the entire distribution. UCSD's mobility rate was equal to 0.0483275.

Next, because we are from UCSD, let's add a vertical line for UCSD to the graph to see how we compare against the entire distribution. UCSD's mobility rate was equal to 0.0483275.

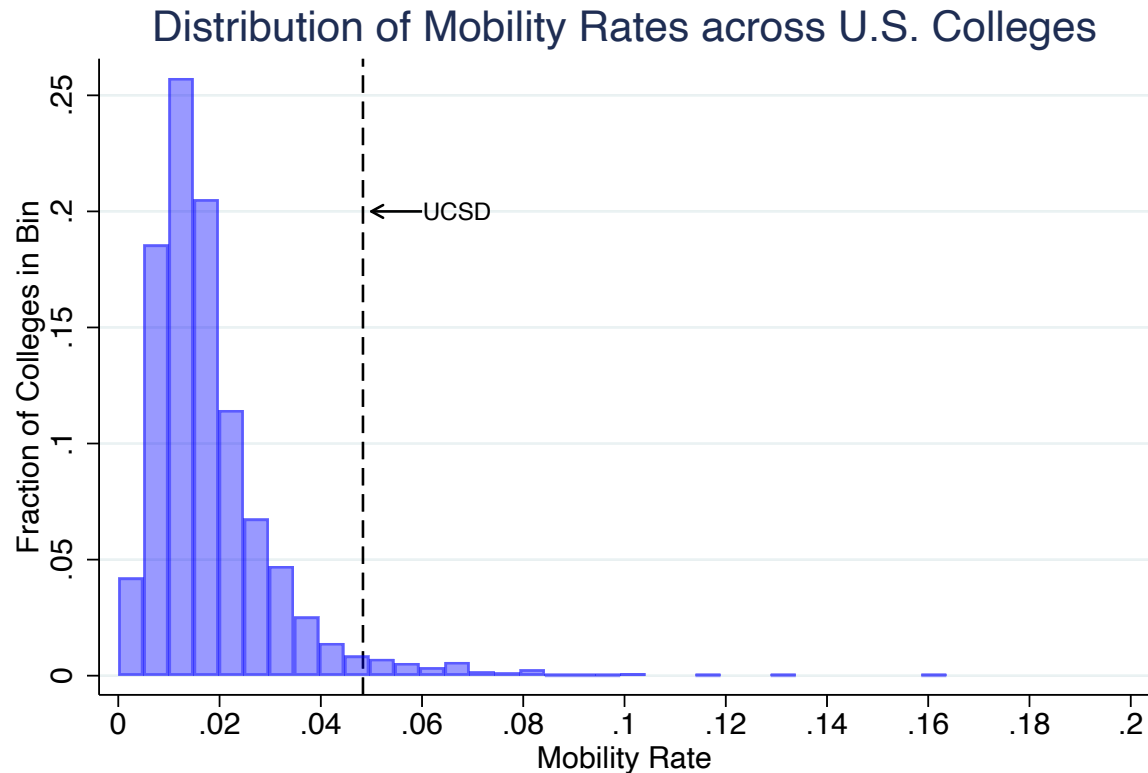
```
In [12]: graph twoway histogram mobility_rate, frac xline(0.0483275)
```



We can add (1) a descriptive title (2) labels and (3) customized aesthetics

In [13]:

```
graph twoway histogram mobility_rate, frac color(blue%40) ///
  || pcarrowi 0.2 0.06 0.2 0.05, lc(black) mcolor(black) ///
  text(0.2 0.06 "UCSD" , place(e) size(small) color(black) ) ///
  xline(0.0483275, lc(black) lp(dash)) ///
  xlabel(0(0.02)0.2) ///
  title("Distribution of Mobility Rates across U.S. Colleges") ///
  xtitle("Mobility Rate") ///
  ytitle("Fraction of Colleges in Bin") ///
  note("Source: Opportunity Insights") ///
  legend(off) ///
  graphregion(color(white) fcolor(white))
```



Source: Opportunity Insights

## Conclusion

### Today's Application

- The degree of intergenerational mobility varies across colleges
- Further questions to think about? What characteristics of colleges do you think are correlated with the ability to promote intergenerational mobility