Statistics Bootcamp Day 1

16 September 2019

Welcome to bootcamp!

Our goals:

1. Increase students' understanding of and confidence with basic statistical concepts.
2. Build students' programming intuition and data management skills.
3. Encourage collaboration and camaraderie among the graduate student cohort.

Who are we?

Amy	3rd year PhD student in Sociology Studies Millennials, mental health, and social media Drinks a lot of coffee Will be your TA for 381 and 382!	aljohnson@stanford.edu
Rebecca	5th year PhD student in Sociology Studies school discipline, race, and policing Also drinks a TON of coffee Loves her dog and Bachelor in Paradise	rgleit@stanford.edu
Nick	3rd year PhD student in Sociology Studies nonprofits and He's not here □	nsherefkin@stanford.edu □

Who are you?

Name & program

Favorite thing you did over the summer

Most boring fact about yourself

And now a word from our sponsors...

https://iriss.stanford.edu/

Overview of the week

Monday: mindset, descriptive & inferential statistics, summary statistics, and Stata workshop

Tuesday: graphing, exponents/logarithms, sampling distributions, and statistical significance

Wednesday: probability basics, file structure and data workflow

Thursday: variable types, functions, lines of best fit, prediction equations

Friday: matrix algebra basics, reading calculus

A few things to keep in mind

We do not expect you to already know ANYTHING that we're going to cover this week. We will start from the very beginning!

> Do not be alarmed because it says "matrix algebra" and "calculus!" We don't expect you to have a background in math. :)

Please let us know if you're going to miss a day/session of bootcamp so we can adjust our lunch ordering.

Today’s learning objectives

...understand the concept of growth mindset and how it applies to math.

...explain the difference between descriptive and inferential statistics.

...calculate mean, median, mode, and standard deviation.

Today's learning objectives

...understand the concept of growth mindset and how it applies to math.

...explain the difference between descriptive and inferential statistics.

...calculate mean, median, mode, and standard deviation.

…understand how data are stored in Stata

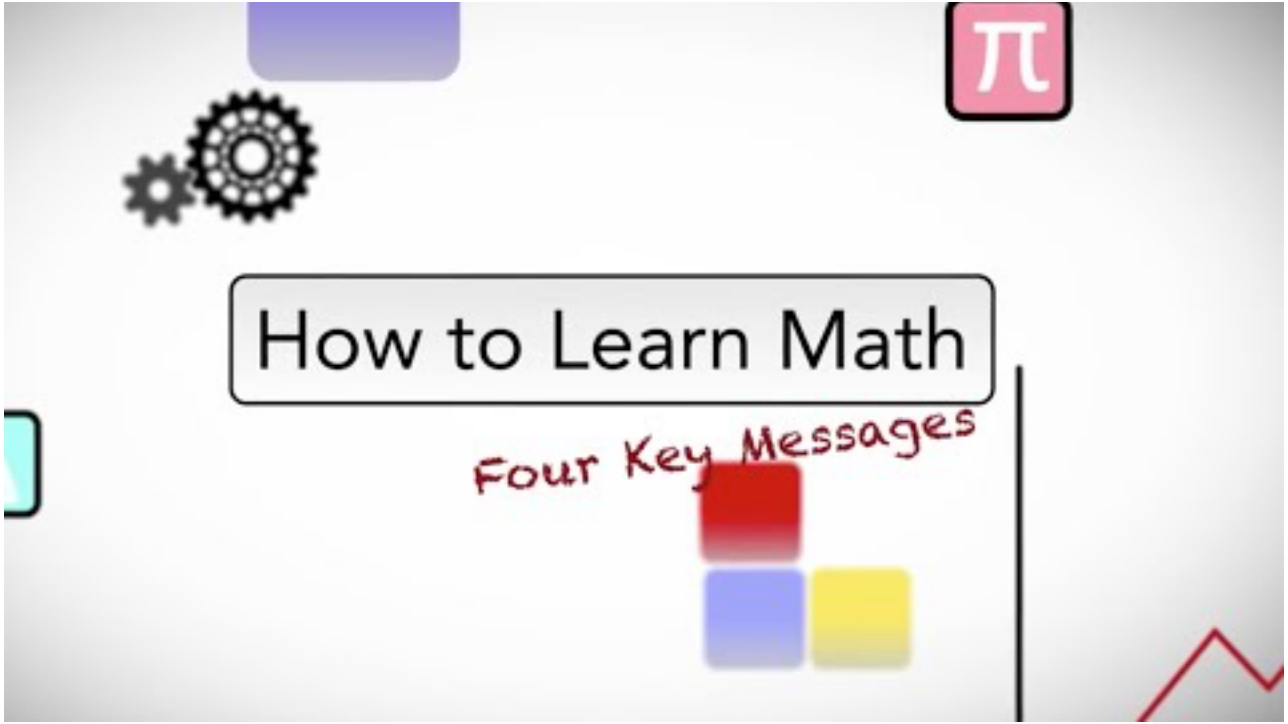
...use logical if-statements to subset data

...use a .do file to write reproducible code

...begin to use functions to manipulate data (e.g. variable creation)

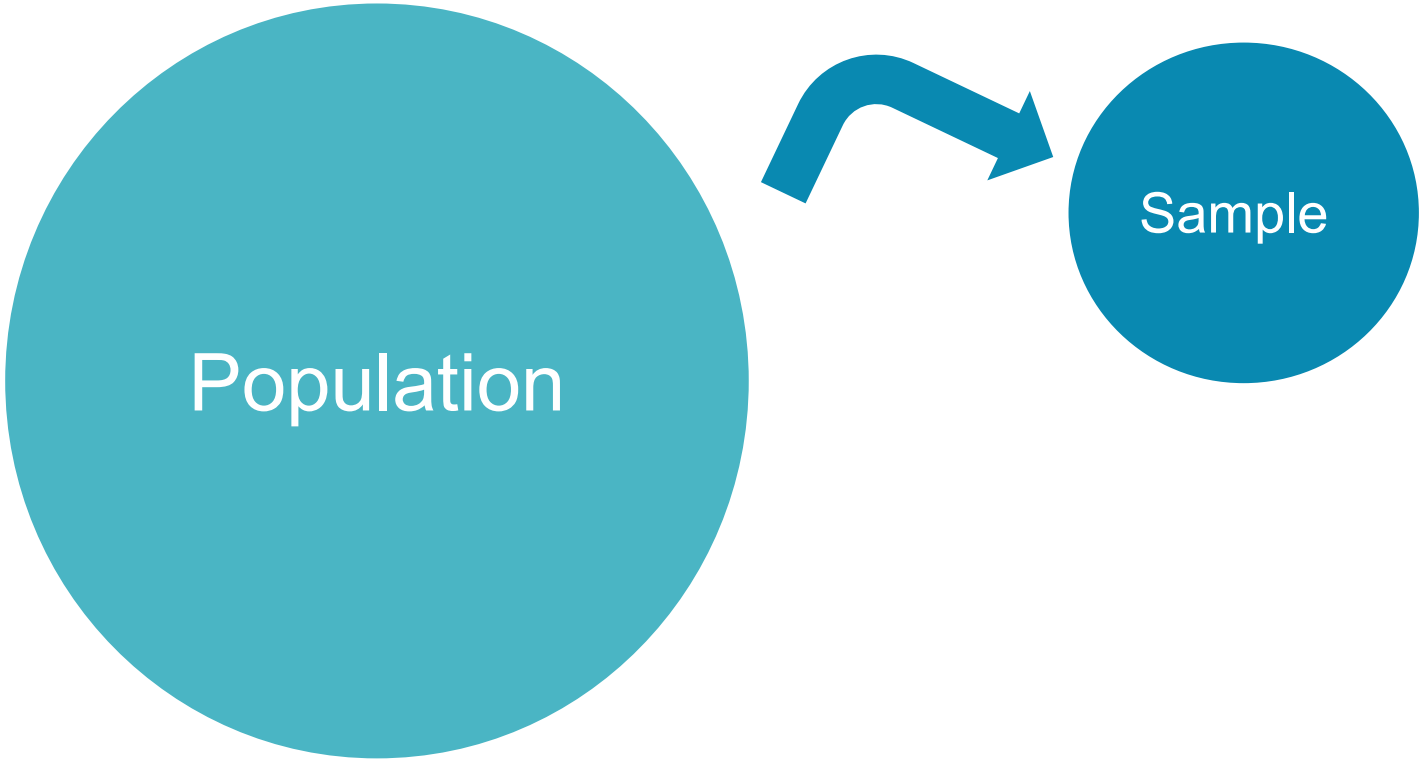
} Stata Workshop

Mindset

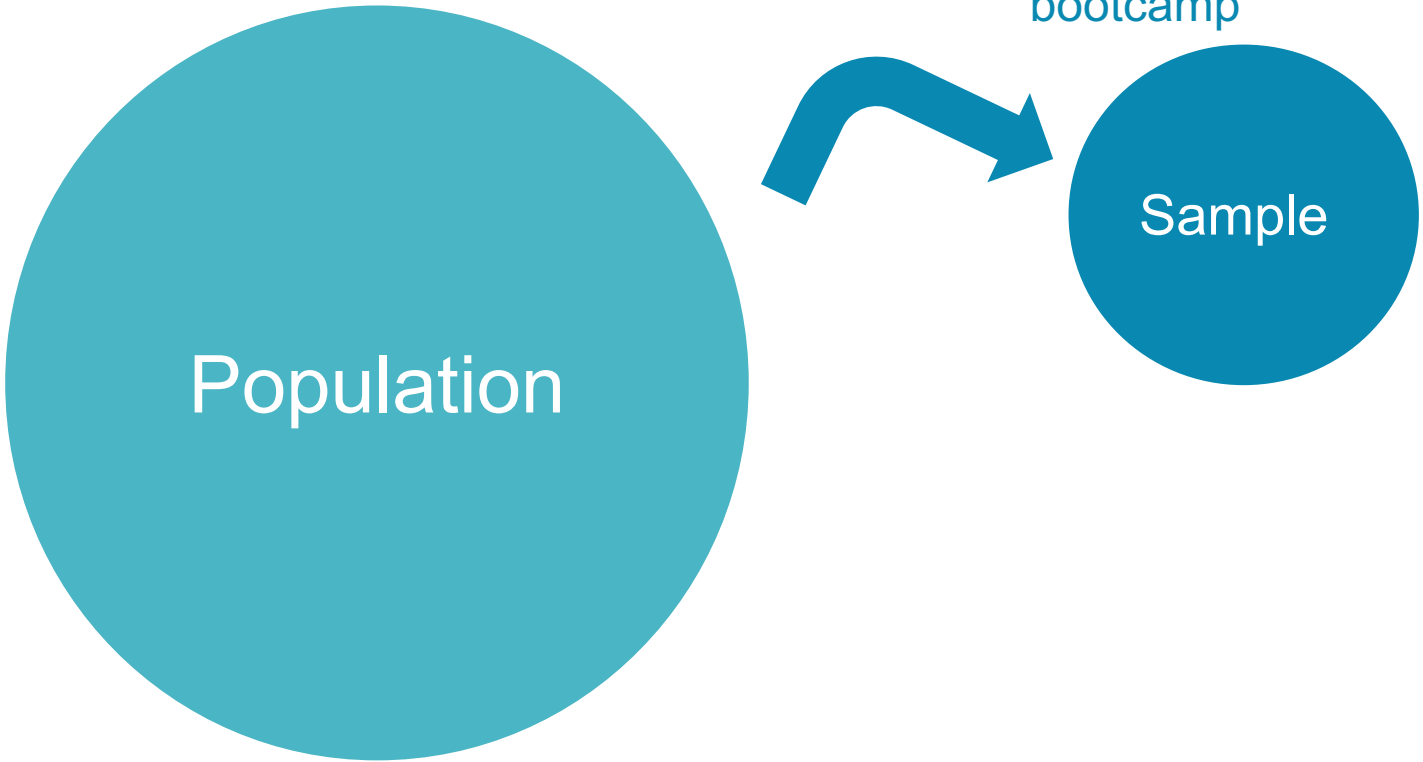

How to Learn Math
Four Key Messages

Descriptive & Inferential Statistics

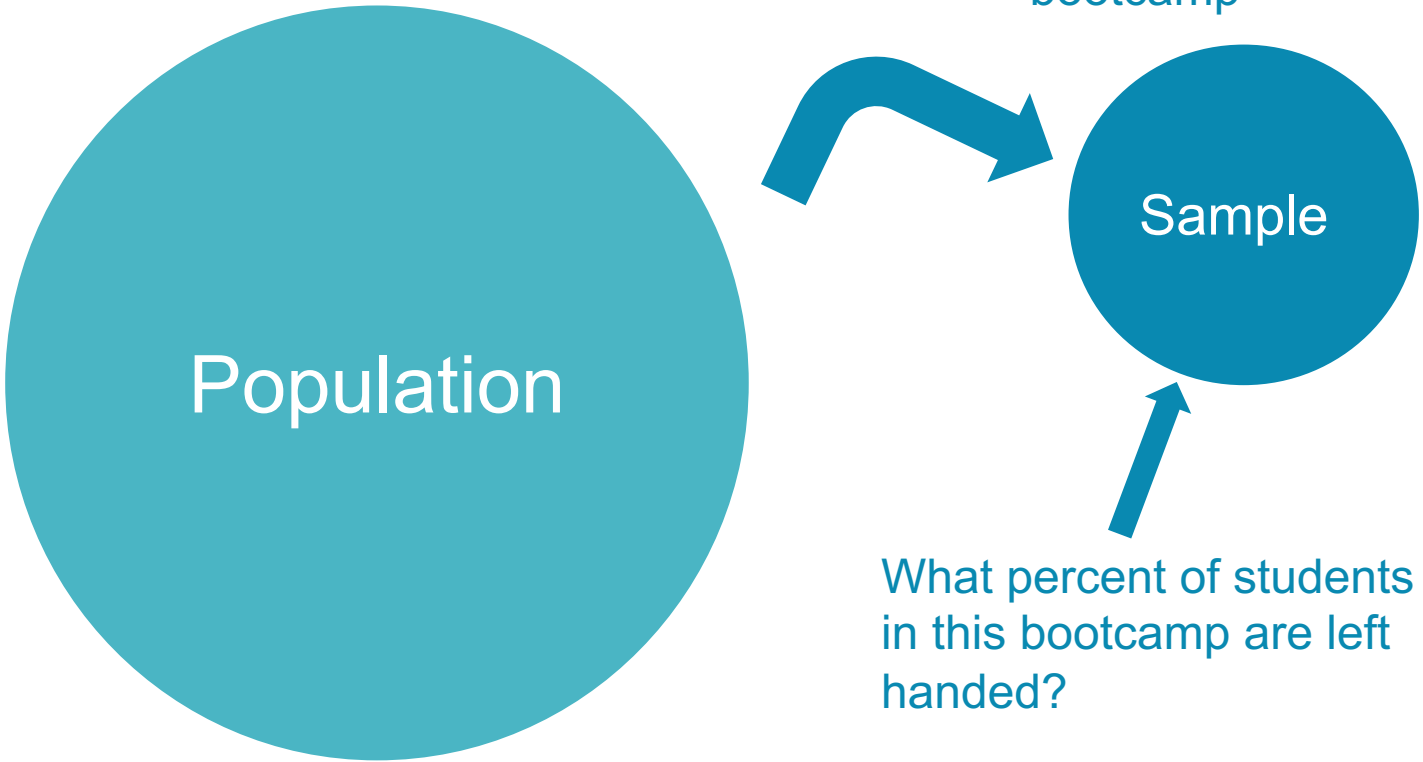
Sample vs. population



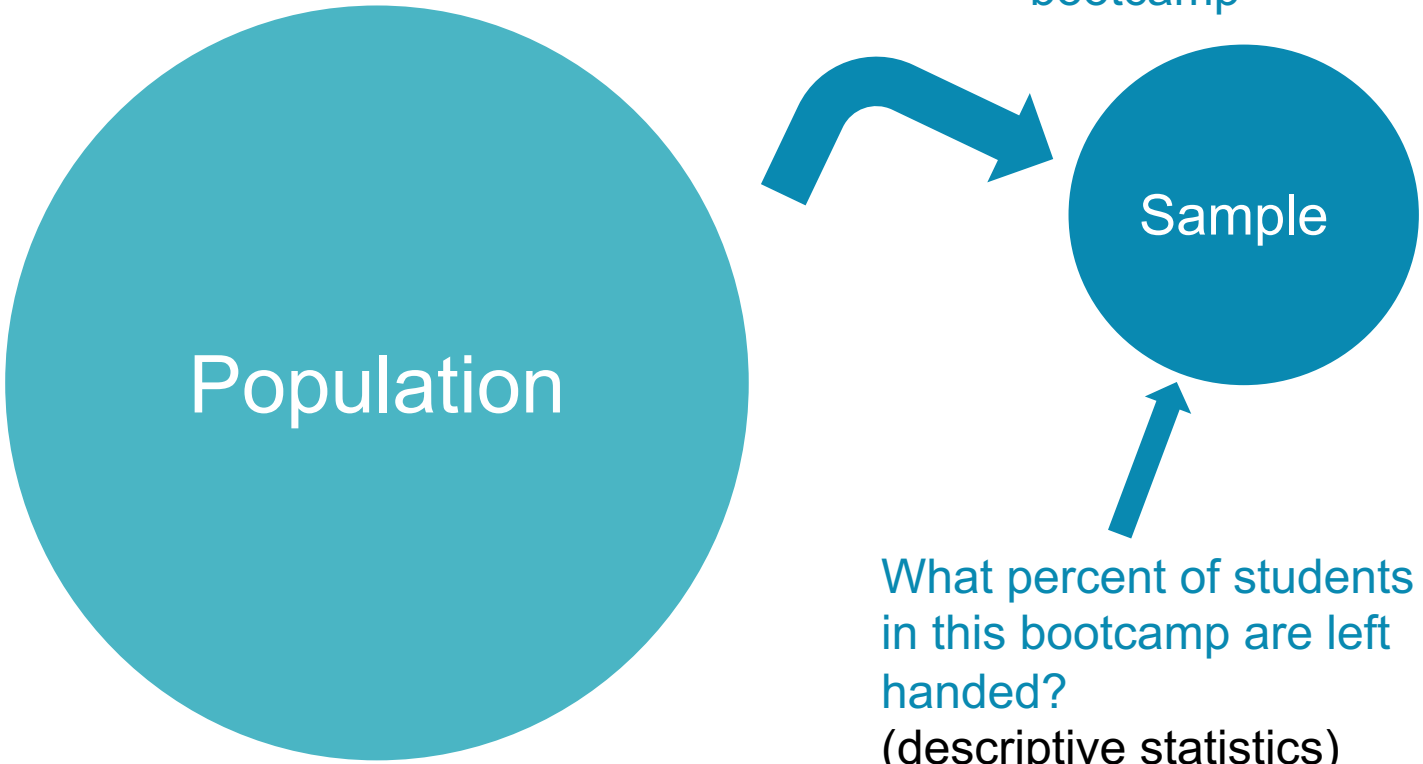
Sample vs. population



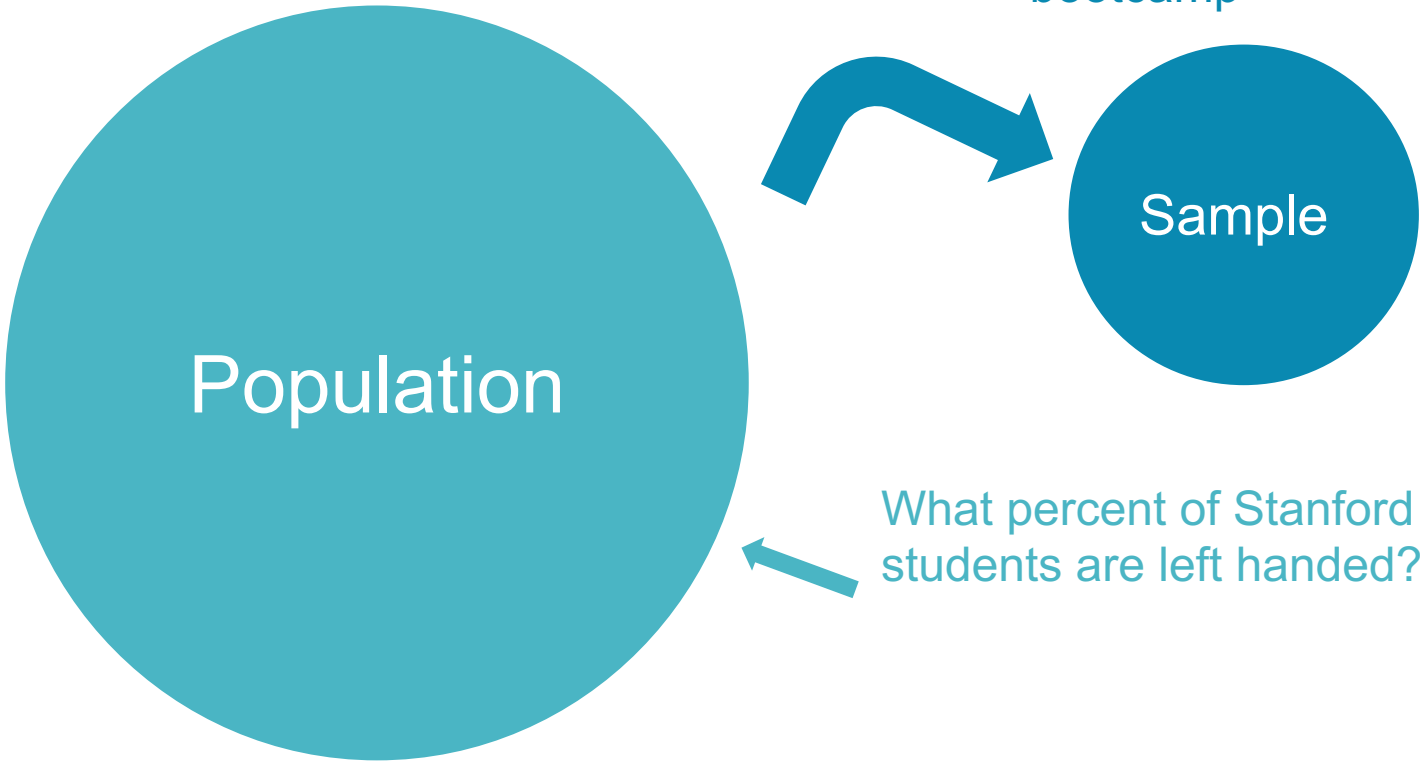
Sample vs. population



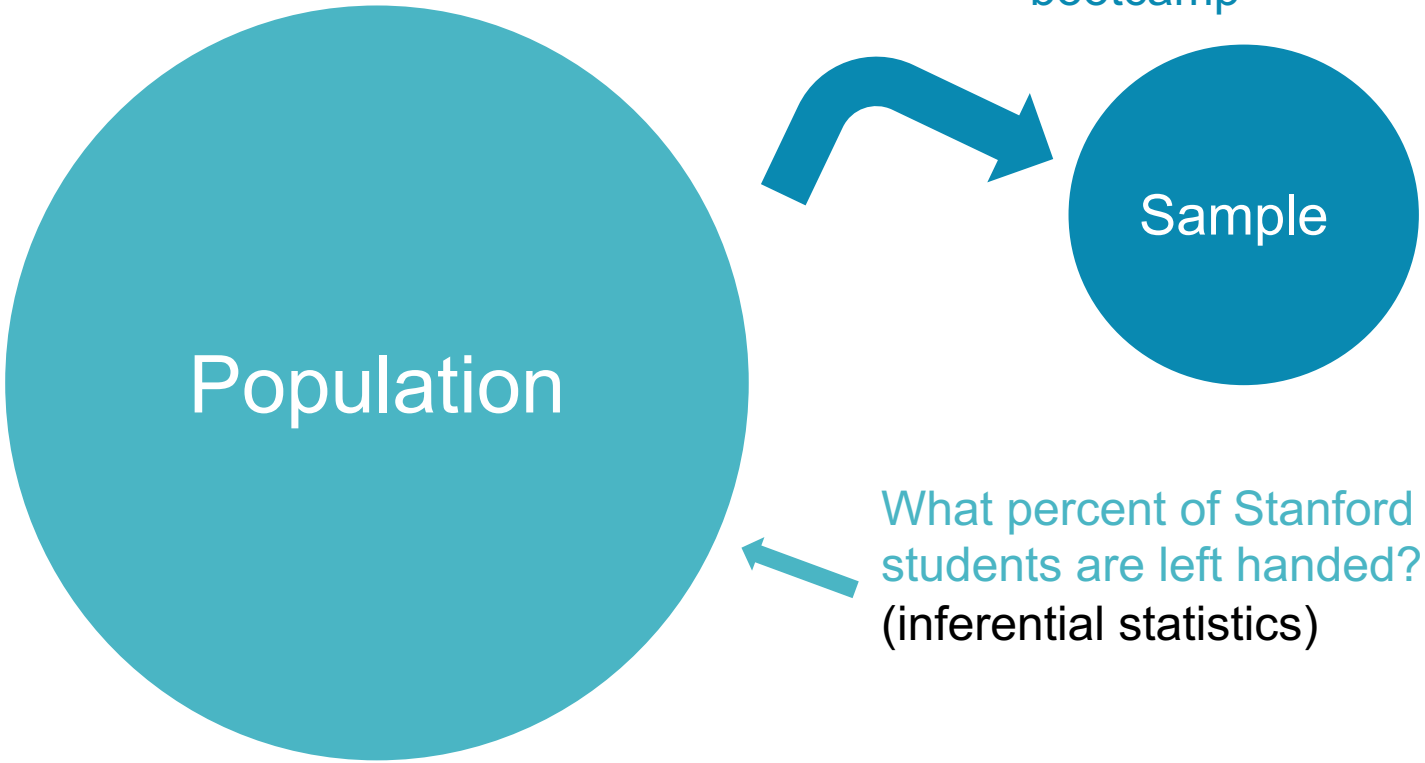
Sample vs. population



Sample vs. population



Sample vs. population



Descriptive vs. Inferential Statistics

Type	Definition	Example
Descriptive	Procedures that help organize and describe data from a sample or a population	
Inferential	The logic and procedures concerned with making predictions or inferences about a population from observations and analyses of a sample	

Descriptive vs. Inferential Statistics

Type	Definition	Example
Descriptive	Procedures that help organize and describe data from a sample or a population	Mean Median Mode Standard Deviation
Inferential	The logic and procedures concerned with making predictions or inferences about a population from observations and analyses of a sample	Hypothesis testing Regression analysis

Descriptive vs. Inferential Statistics

These are called “summary statistics” and are used to “summarize” the data.

Type	Definition	Example
Descriptive	Procedures that help organize and describe data from a sample or a population	Mean Median Mode Standard Deviation
Inferential	The logic and procedures concerned with making predictions or inferences about a population from observations and analyses of a sample	Hypothesis testing Regression analysis

Descriptive Statistics

Statistic	Meaning	Pros & Cons
Mean		
Median		
Mode		
Standard Deviation		

Descriptive Statistics

Statistic	Meaning	Pros & Cons
Mean	The middle	-More prone to being skewed by outliers -Easily understood
Median		
Mode		
Standard Deviation		

Descriptive Statistics

Statistic	Meaning	Pros & Cons
Mean	The middle	-More prone to being skewed by outliers -Easily understood
Median	The middle	-Less prone to being skewed by outliers -Easily misunderstood
Mode		
Standard Deviation		

Descriptive Statistics

Statistic	Meaning	Pros & Cons
Mean	The middle	-More prone to being skewed by outliers -Easily understood
Median	The middle	-Less prone to being skewed by outliers -Easily misunderstood
Mode	The most	-Sometimes not useful
Standard Deviation		

Descriptive Statistics

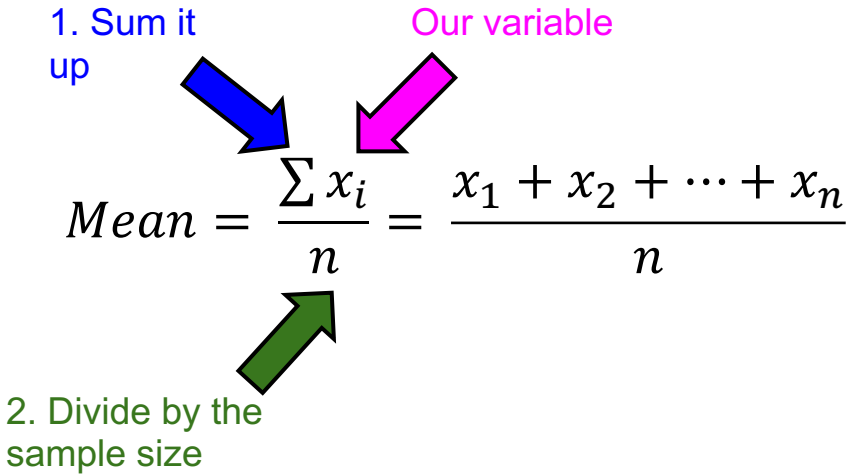
Statistic	Meaning	Pros & Cons
Mean	The middle	-More prone to being skewed by outliers -Easily understood
Median	The middle	-Less prone to being skewed by outliers -Easily misunderstood
Mode	The most	-Sometimes not useful
Standard Deviation	Average distance from the mean	-Most meaningful when provided with the mean -Easily misunderstood

Mean

So if we had a dataset of ages:

- 8
- 6
- 11
- 8
- 7

The mean would be…



In other words: add all of the data points up, and divide by the number of data points.

Median

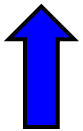
- Put the numbers in order from smallest to largest
- The median is the number in the middle
  - If you have an even number of data points (meaning there is no “middle”), the median is the mean of the middle two numbers

So if we had a dataset of ages:

- 8
- 6
- 11
- 8
- 7

The median would be…

6, 7, 8, 8, 11



The median

Mode

- Determine which number is observed most often

So if we had a dataset of ages:

- 8
- 6
- 11
- 8
- 7

The mode would be…

Signal v. Noise

Signal: our estimate based on the data we have

E.g. mean, median

Noise: the uncertainty of our estimate

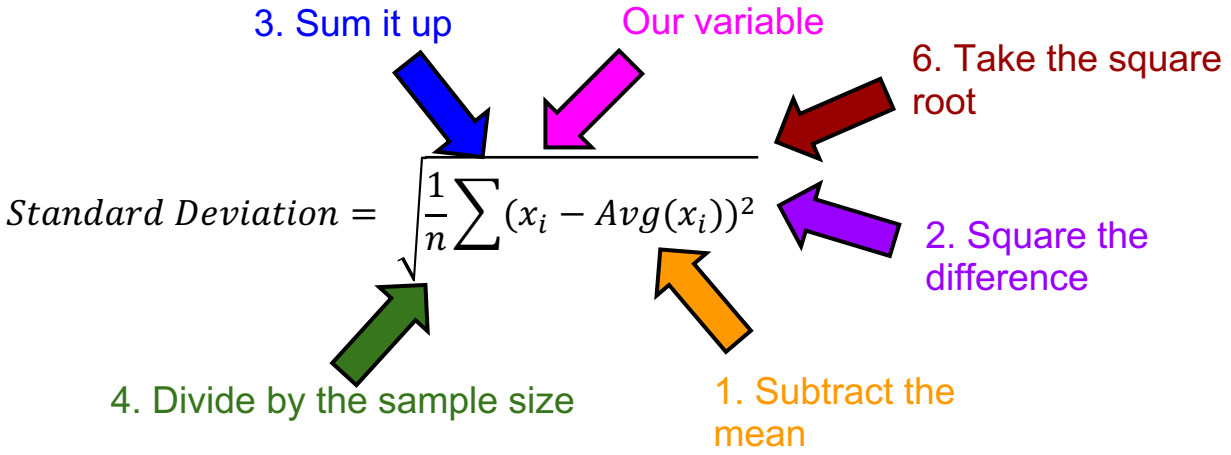
E.g. standard deviation

Standard Deviation

So if we had a dataset of ages:

- 8
- 6
- 11
- 8
- 7

The standard deviation would be…



In other words: subtract the mean from each data point and square it; add them all up; divide by the number of data points; take the square root.

A quick note on variables

- A **variable** is something we measure for every **unit** in our sample.
  - Units are the things we are studying, e.g. people

Variable type	**Description**
Dummy variable	Can take on two values, yes or no (usually coded as 1/0) e.g. female, immigrant
Categorical variable	Can take on a finite number of values, called categories e.g. marital status, race
Continuous/numeric	Can take on any numeric value e.g. age, height, income

lunch 

Introduction to Stata

Outline

Part 1: Data Organization

Part 2: Data Manipulation

Part 3: Self-Directed with Stata

Part 1: Data Organization

Data file: Friends.dta

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66
2	Math	Sophomore	Midwest,West	1	64
3	Sociology	Grad student	Northeast,Midwest,West	3	69
4	Sociology	Grad student	Northeast,Midwest,West	1	65
5	Sociology	Grad student	Northeast,West	4	65
6	Sociology	Grad student	Northeast,West	2	83
7		Co-term		0	77
8		Sophomore	South	1	88
9	Sociology of Education	Grad student	Northeast,West	1	63
10	Undeclared	Freshman	Midwest	.	38
11	Sociology!	Grad student	West	1	68
12	Sociology	Grad student	Midwest,West	1	70
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66
2	Math	Sophomore	Midwest,West	1	64
3	Sociology	Grad student	Northeast,Midwest,West	3	69
4	Sociology	Grad student	Northeast,Midwest,West	1	65
5	Sociology	Grad student	Northeast,West	4	65
6	Sociology	Grad student	Northeast,West	2	83
7		Co-term		0	77
8		Sophomore	South	1	88
9	Sociology of Education	Grad student	Northeast,West	1	63
10	Undeclared	Freshman	Midwest	.	38
11	Sociology!	Grad student	West	1	68
12	Sociology	Grad student	Midwest,West	1	70
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66

Each row is a different person

regions[1]	Northeast,West			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

regions[1]	Northeast,West

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

year_school[1]	3

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

regions[1]	Northeast,West

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

year_school[1]	3

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

height[1]	66

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

String

regions[1]	Northeast,West			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

Numeric, with labels

year_school[1]	3			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

Numeric, without labels

height[1]	66			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

Ways to Store Information

1. **String:**

Data are stored and appear as text

Ways to Store Information

1. **String:**

Data are stored and appear as text

2. Numeric

Ways to Store Information

1. **String:**

Data are stored and appear as text

2. Numeric

2A. **With labels:**

Data appear to be text, but are actually stored in the computer as numbers

Ways to Store Information

1. **String:**

Data are stored and appear as text

2. Numeric

2A. **With labels:**

Data appear to be text, but are actually stored in the computer as numbers

2B. **Without labels:**

Data are stored and appear as numbers

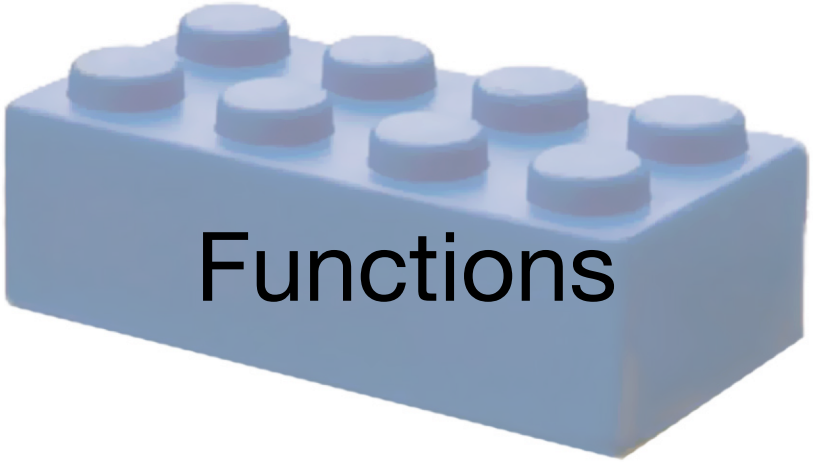
ACTIVITY: Determine the type of variable (string, numeric with labels, numeric without labels) for each variable.

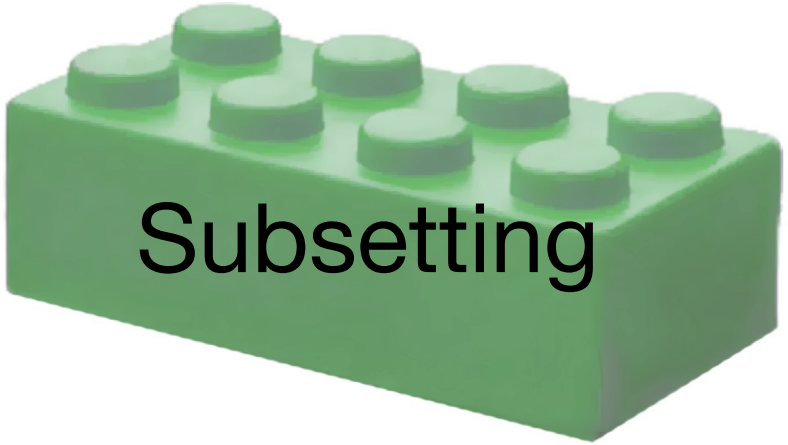
	major	year_school	regions	siblings	height	temp	F_C	cheese
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
7		Co-term		0	77	0	C	Gouda
8		Sophomore	South	1	88	.	.	
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya

Vars: 8 Order: Dataset Obs: 13

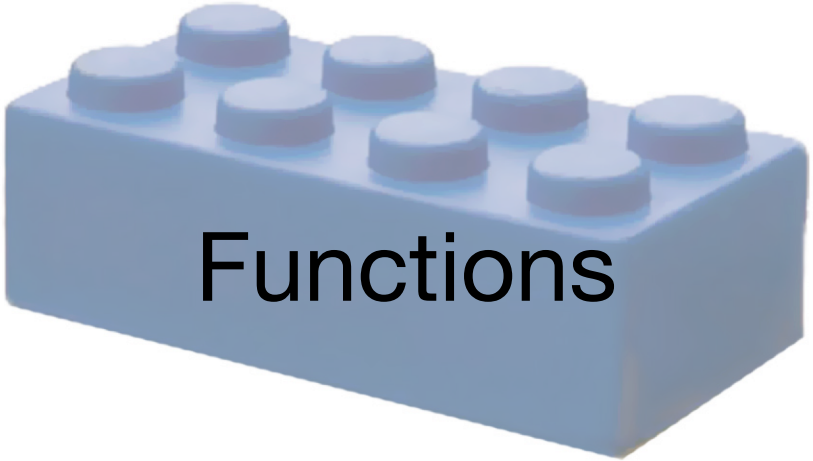
Part 2: Data Manipulation

The building blocks of data manipulation:

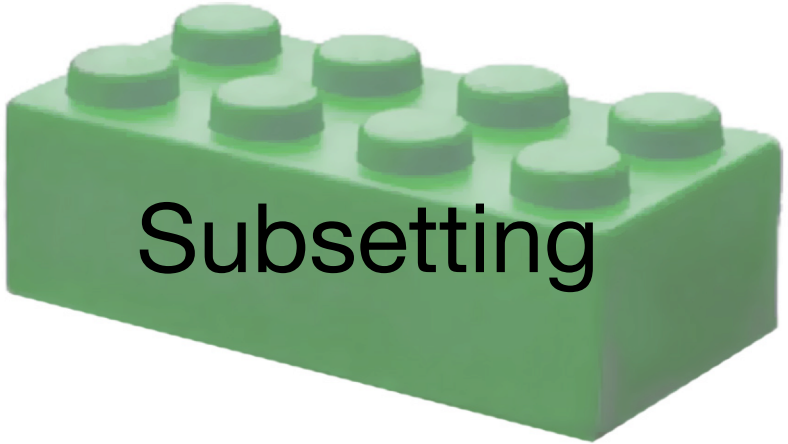




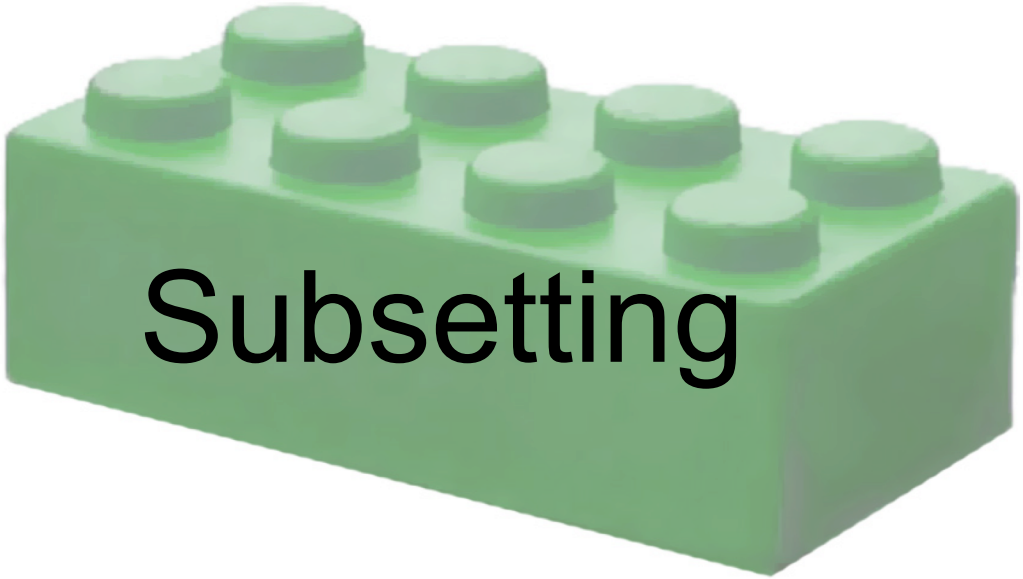
The building blocks of data manipulation:



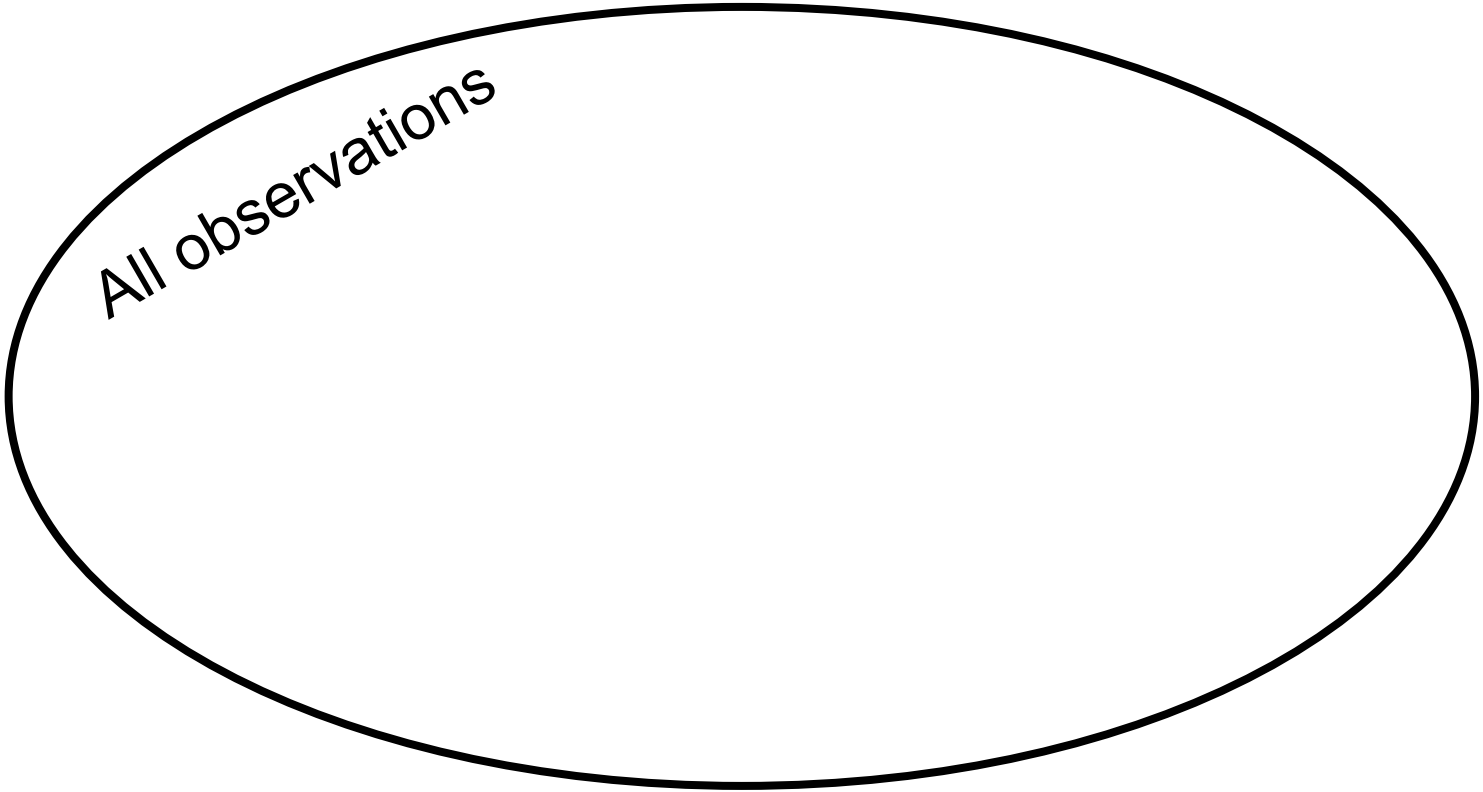
what we want to do

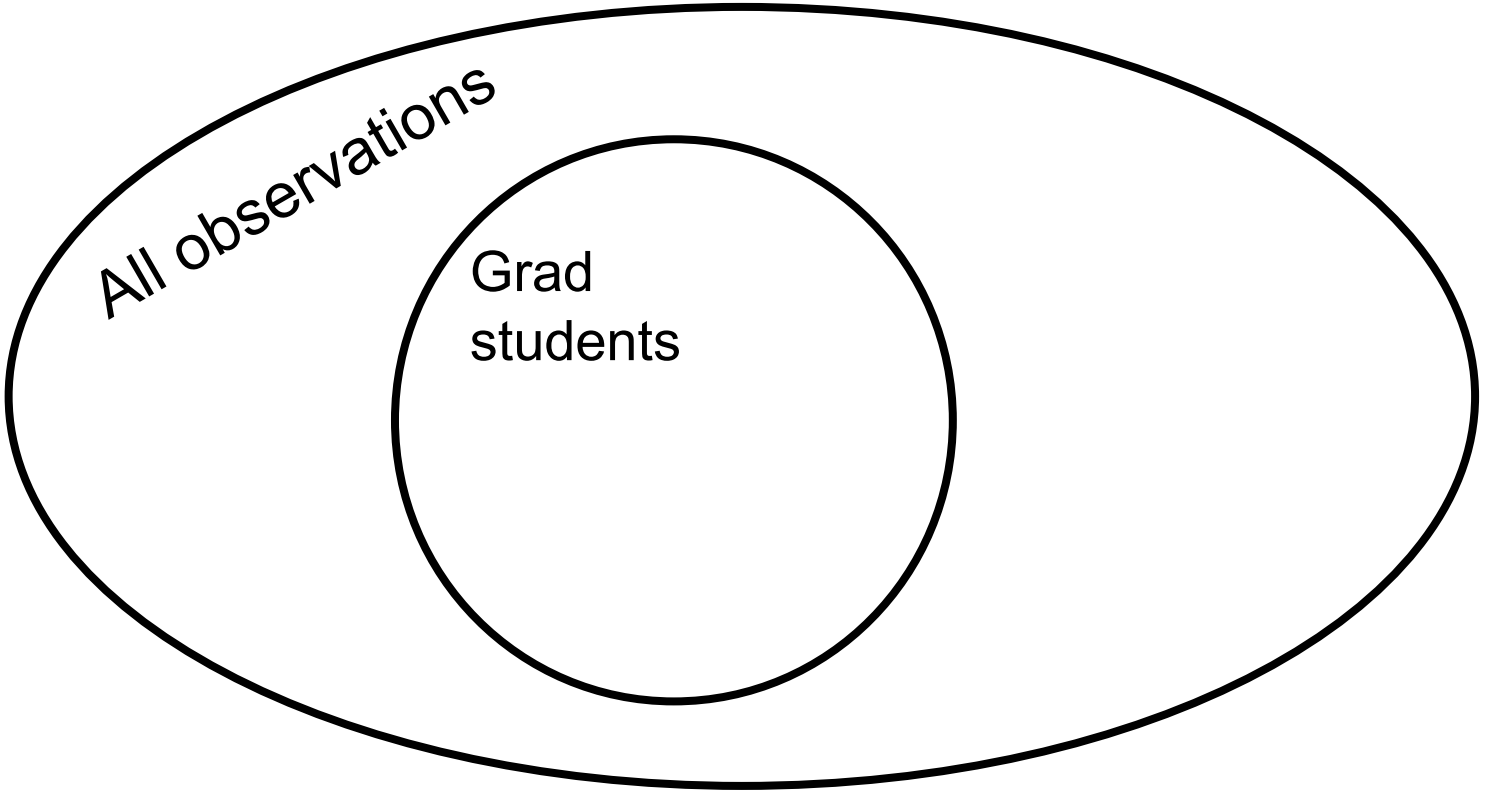


which observations we want to use the function on

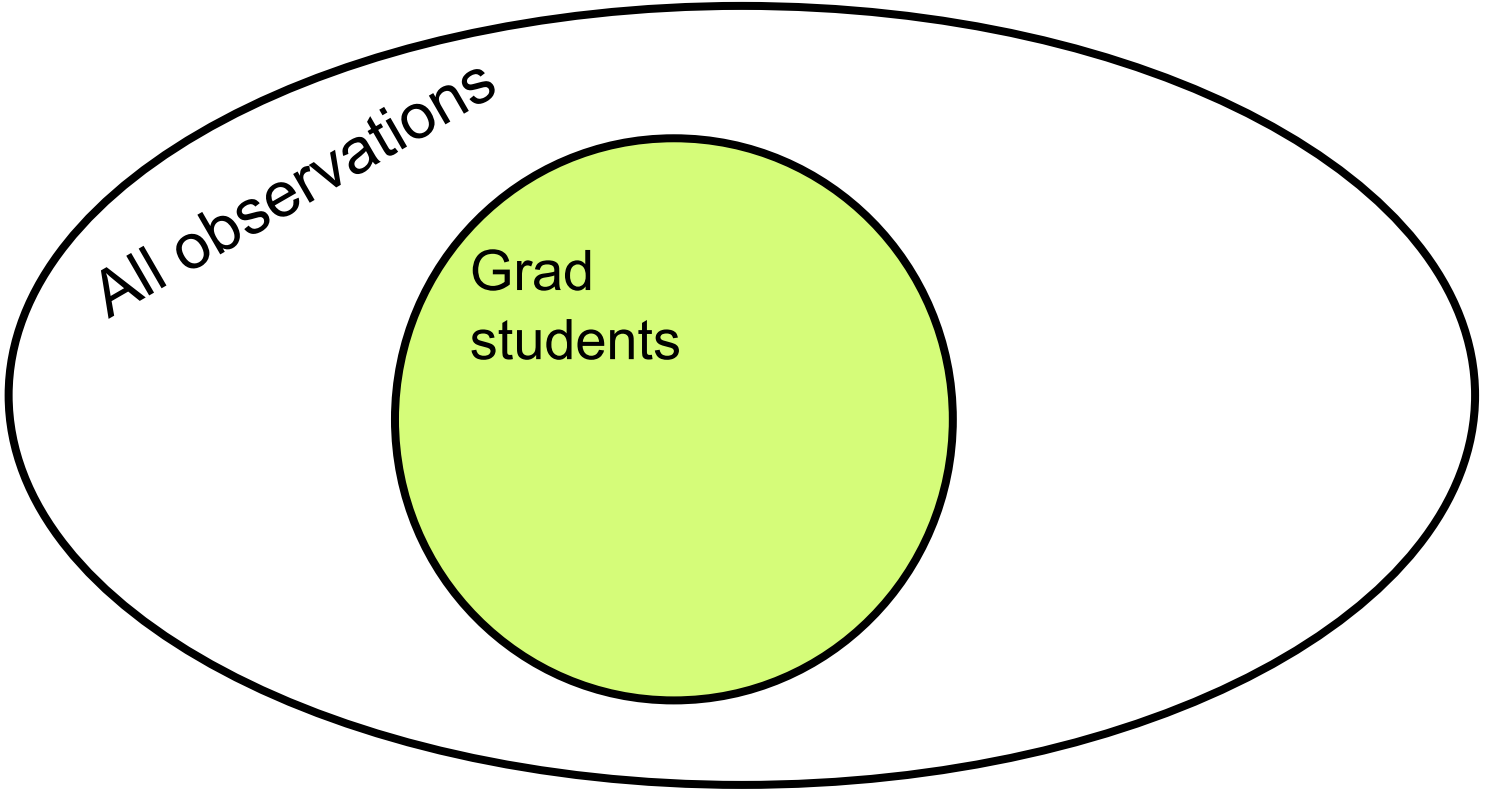

Subsetting

Using logical if-statements to subset

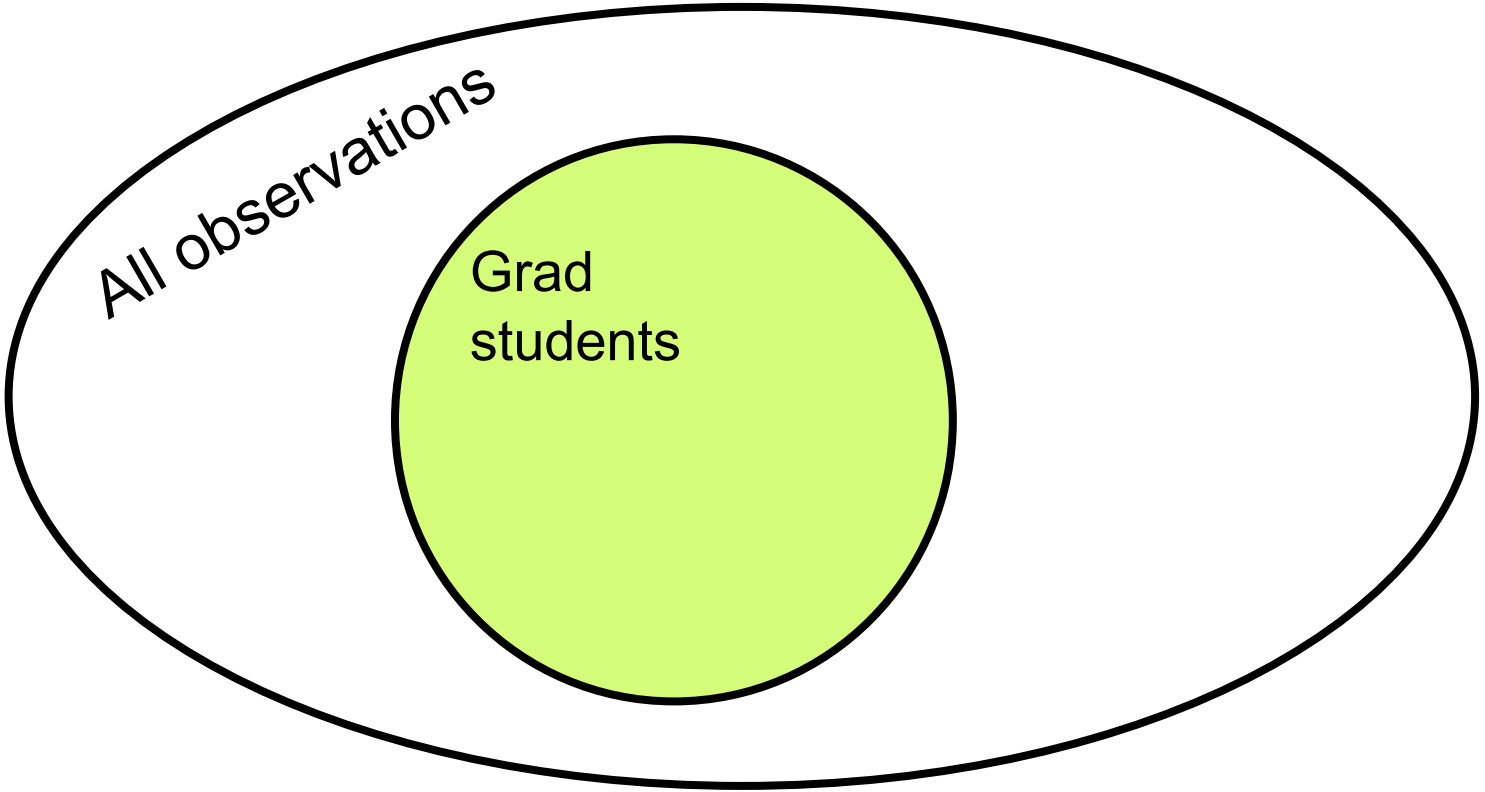



All observations
Grad
students

Look at people **if** they are a grad student

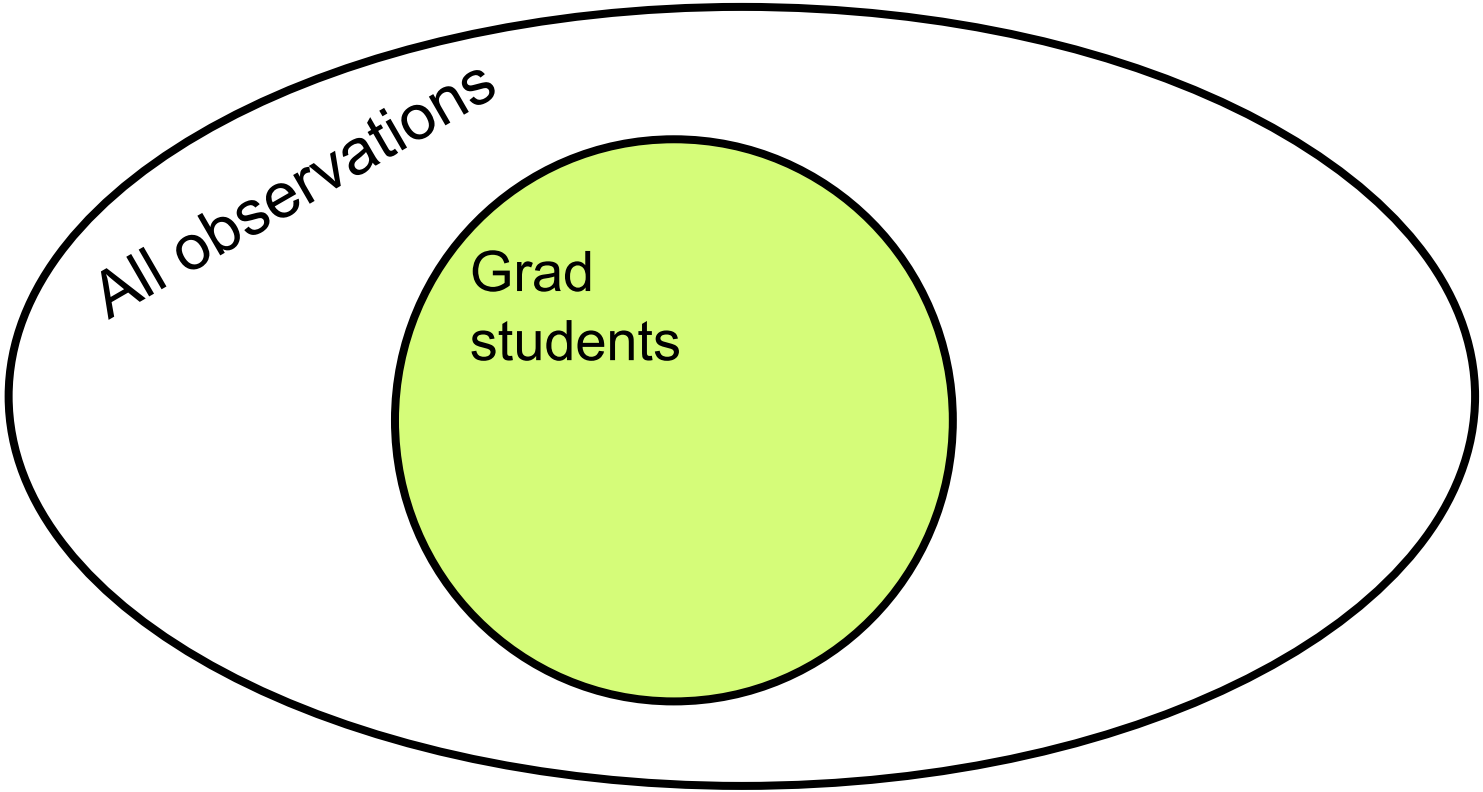


if year_school is equal to 6 ← (words)

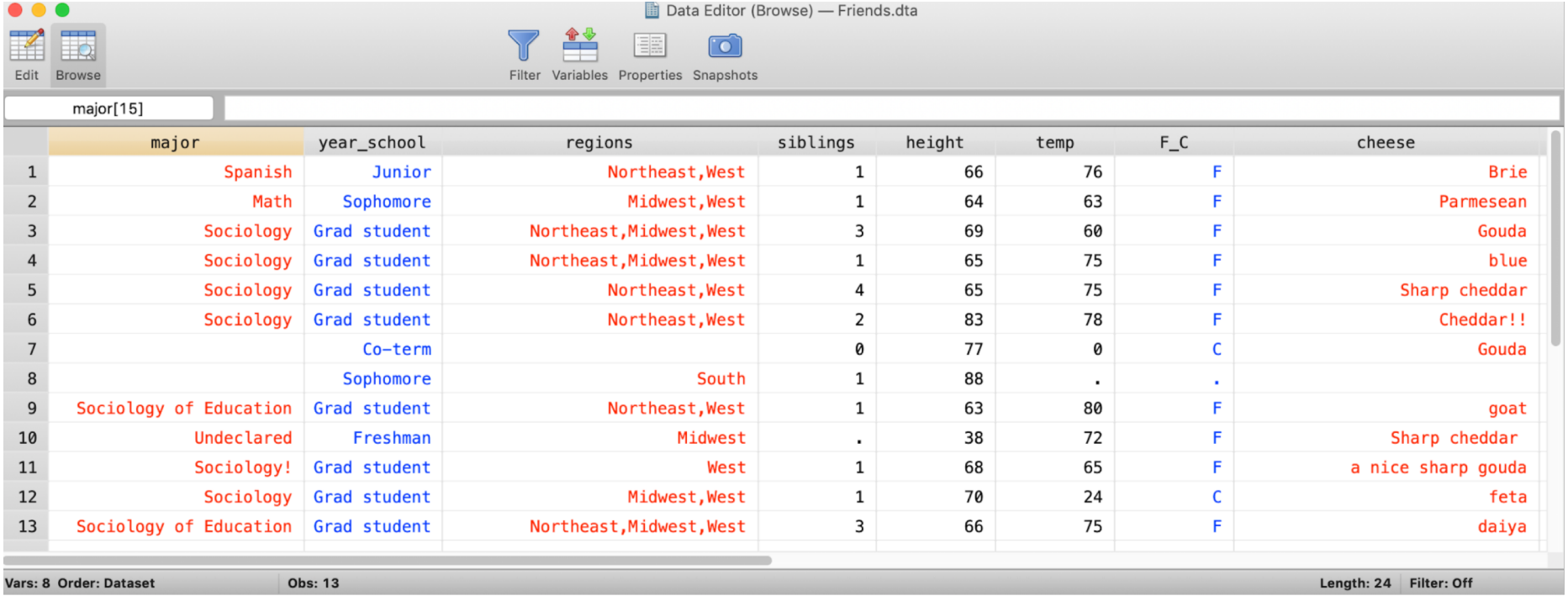


if year_school is equal to 6 ← (words)

```
if year_school==6
```
← (Stata syntax)

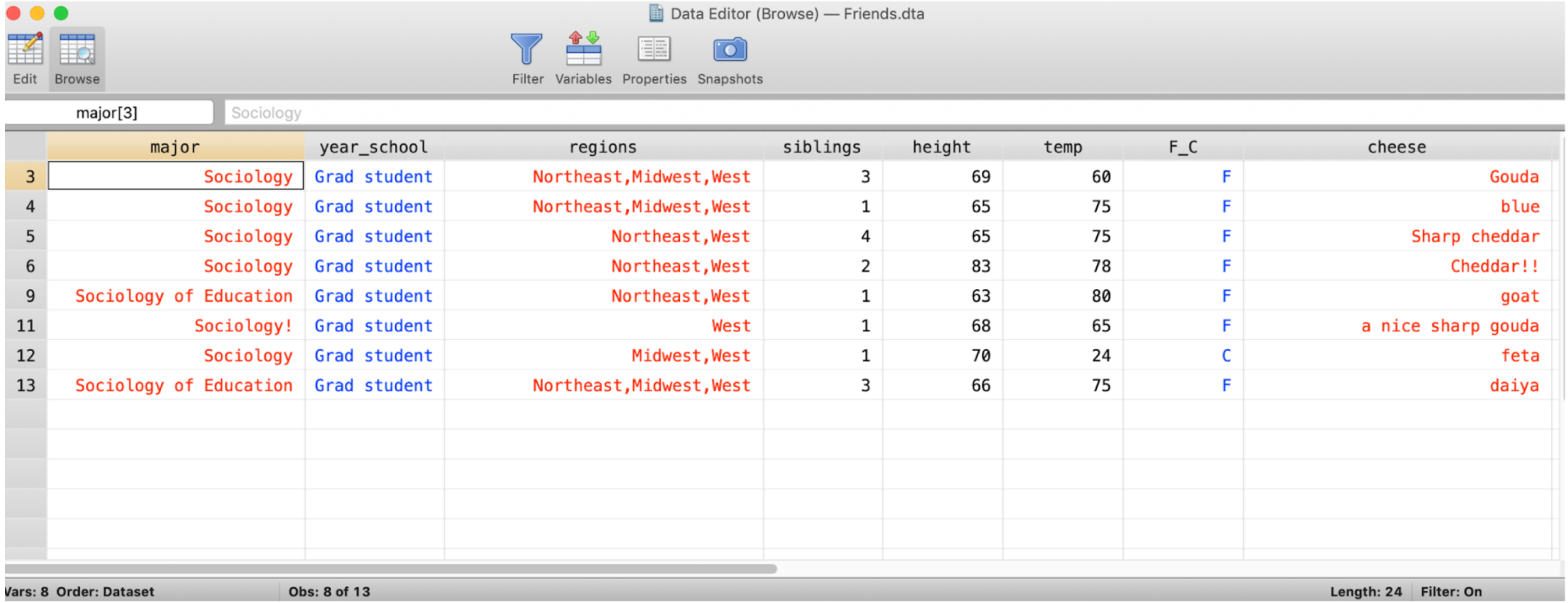


browse

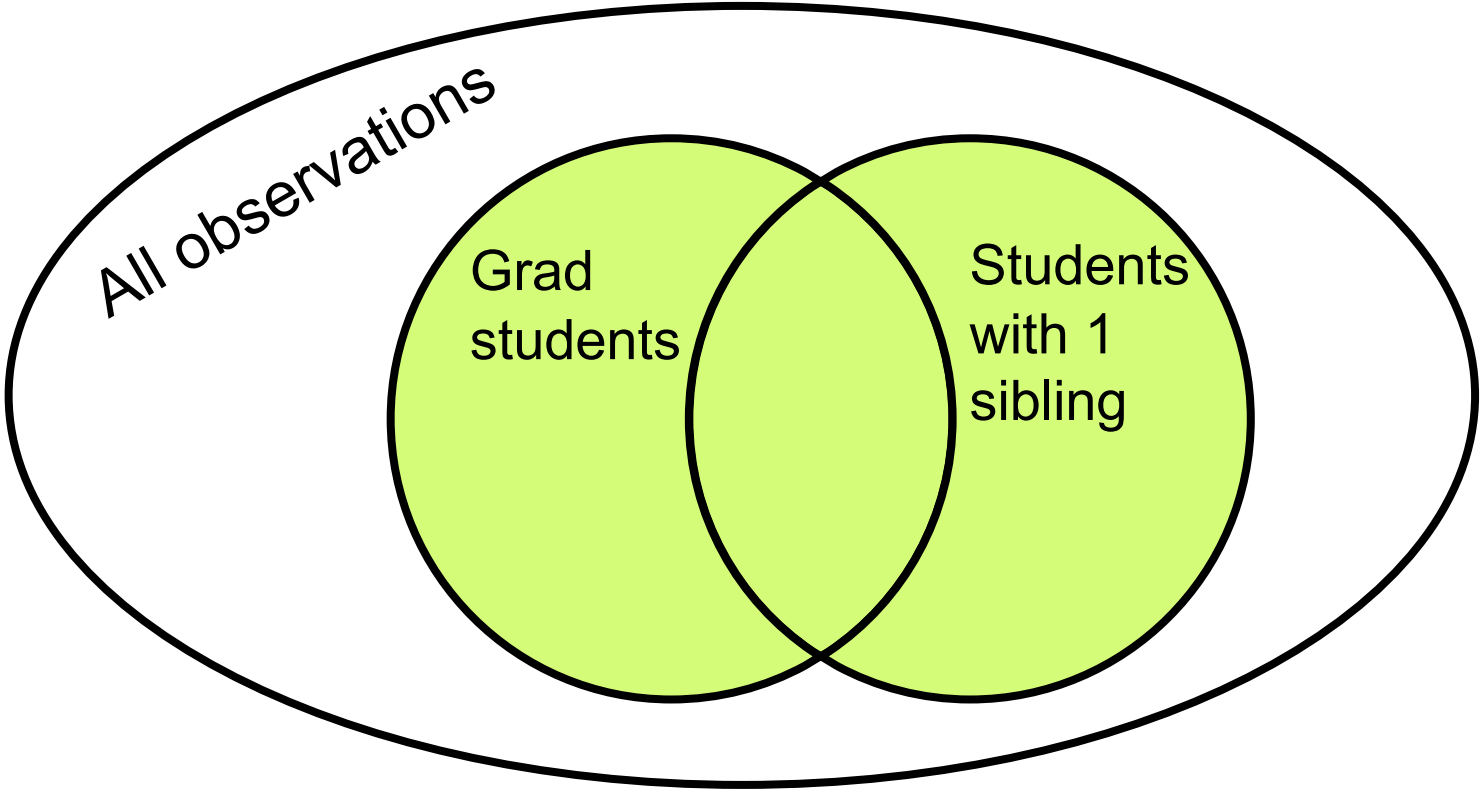


	major	year_school	regions	siblings	height	temp	F_C	cheese
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
7		Co-term		0	77	0	C	Gouda
8		Sophomore	South	1	88	.	.	
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya

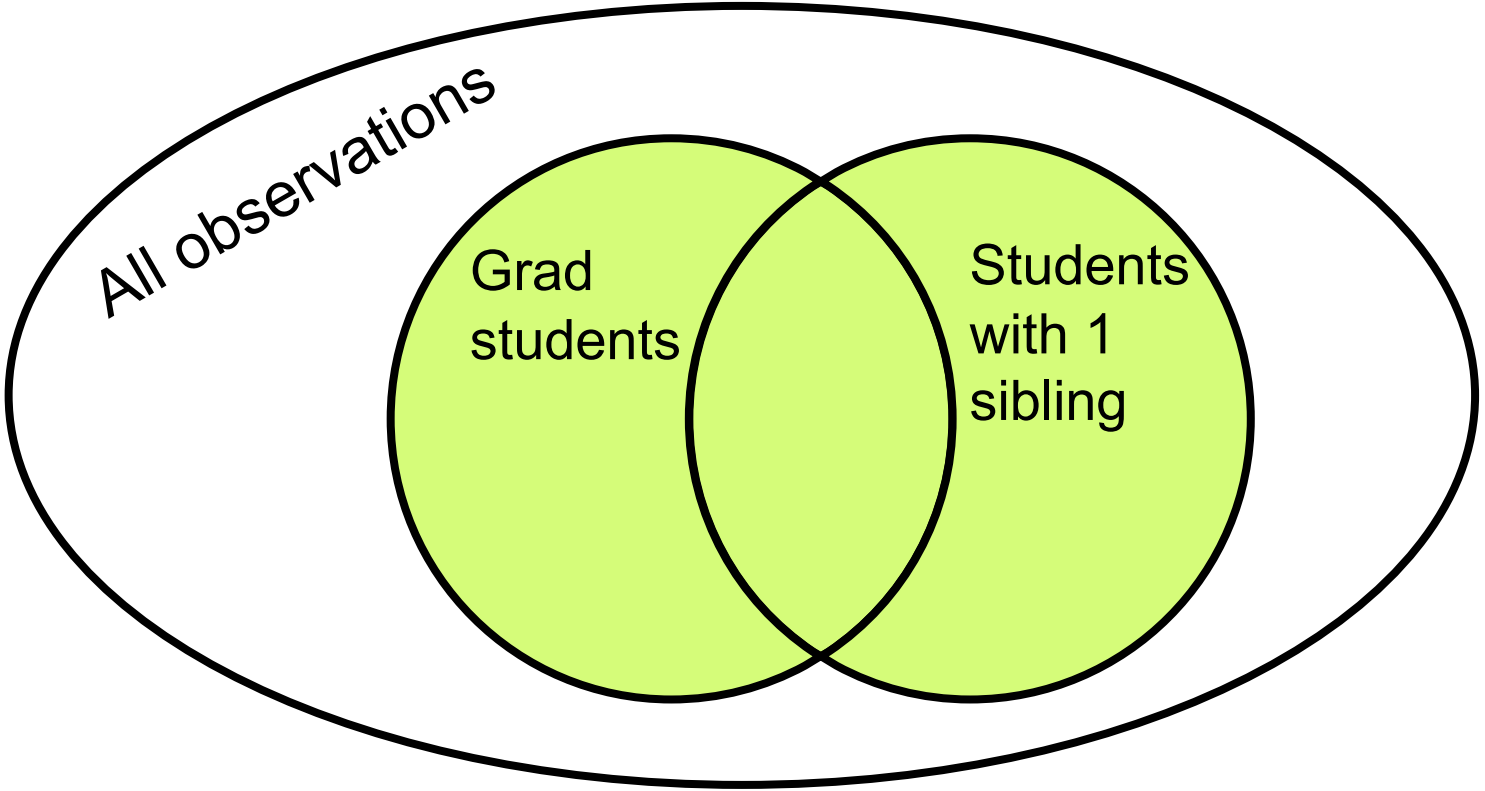
browse if year_school==6



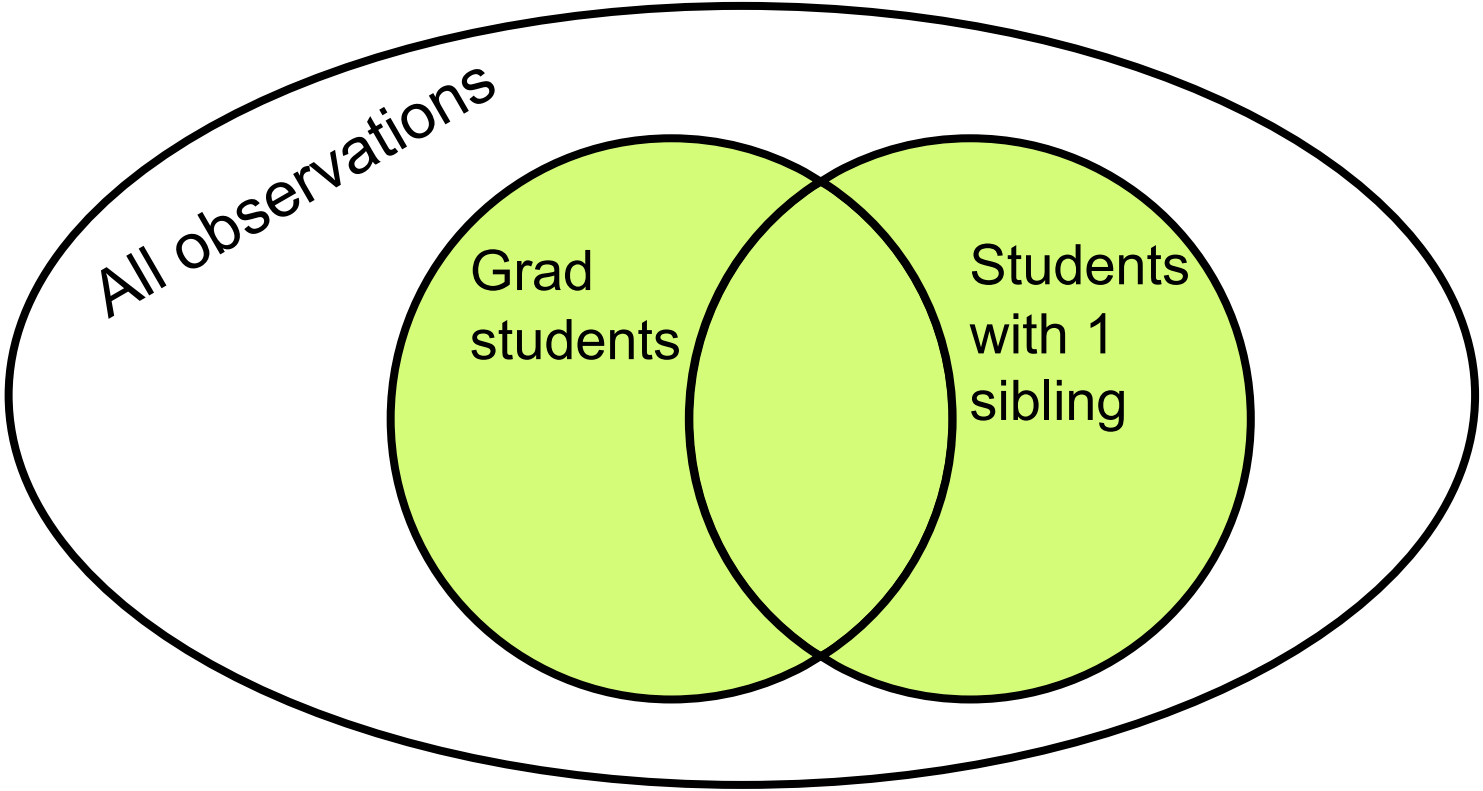
	major	year_school	regions	siblings	height	temp	F_C	cheese
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya


All observations
Grad students
Students with 1 sibling

Look at people **if** they are a grad student OR they have 1 sibling

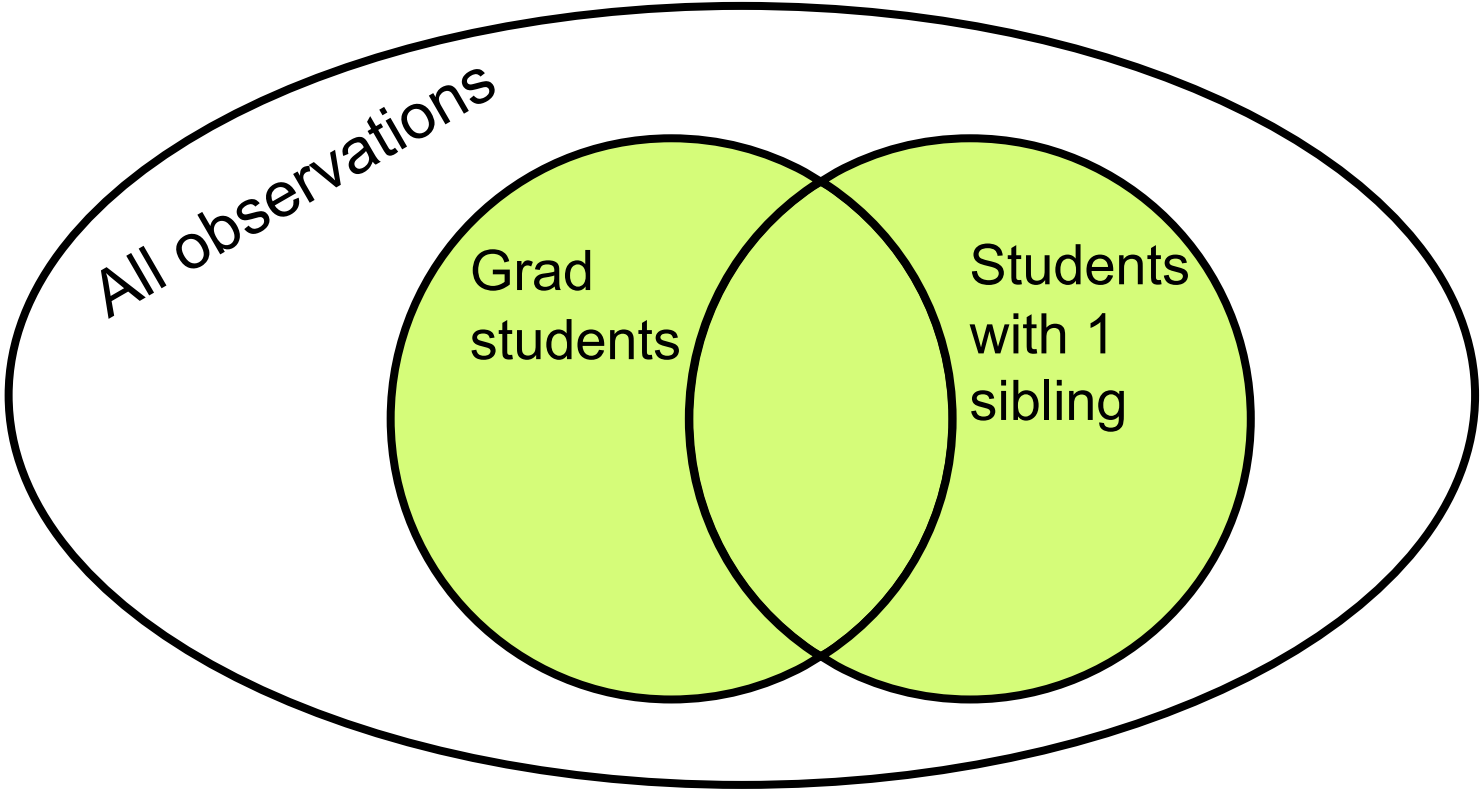


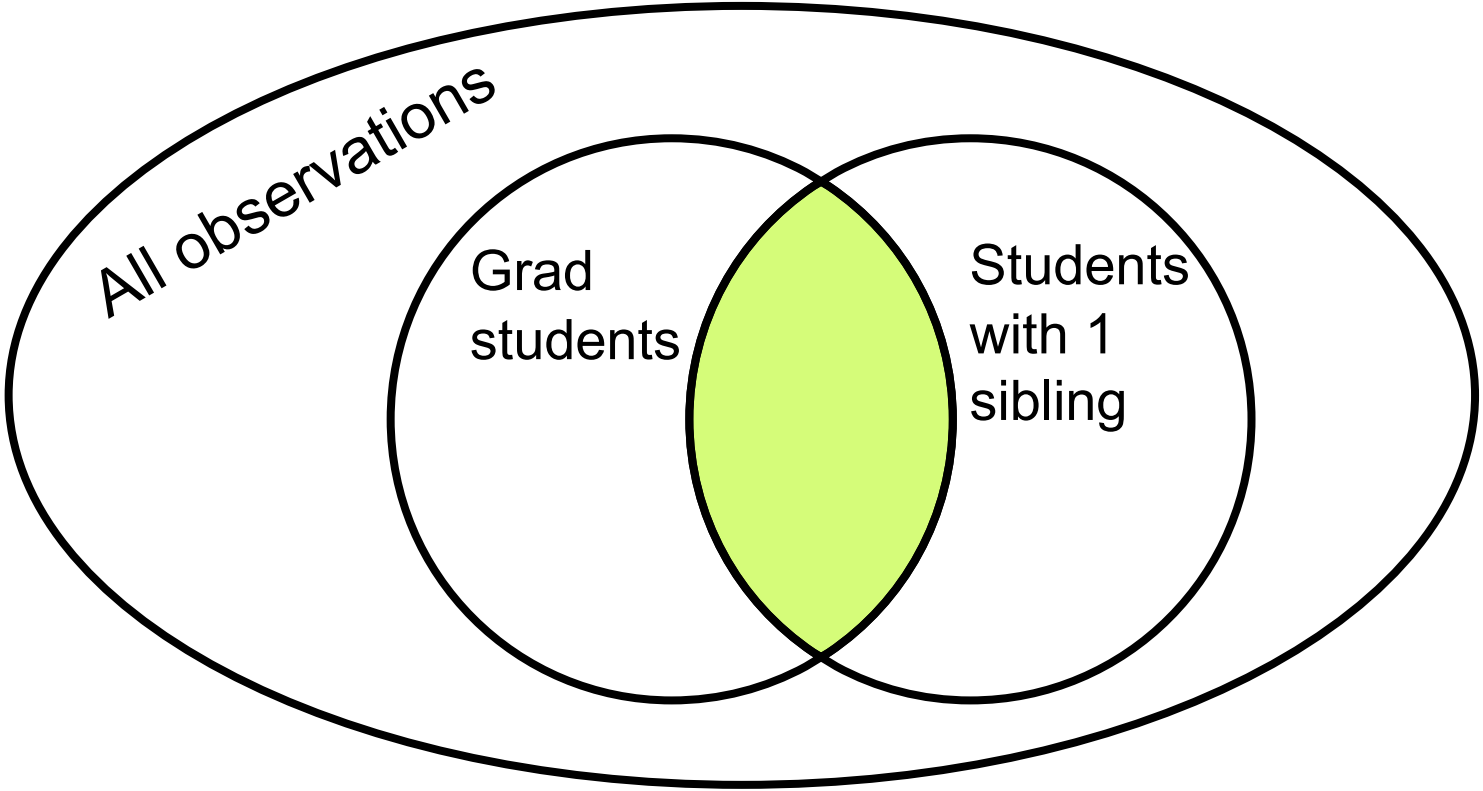
if year_school is equal to 6 OR siblings is equal to 1



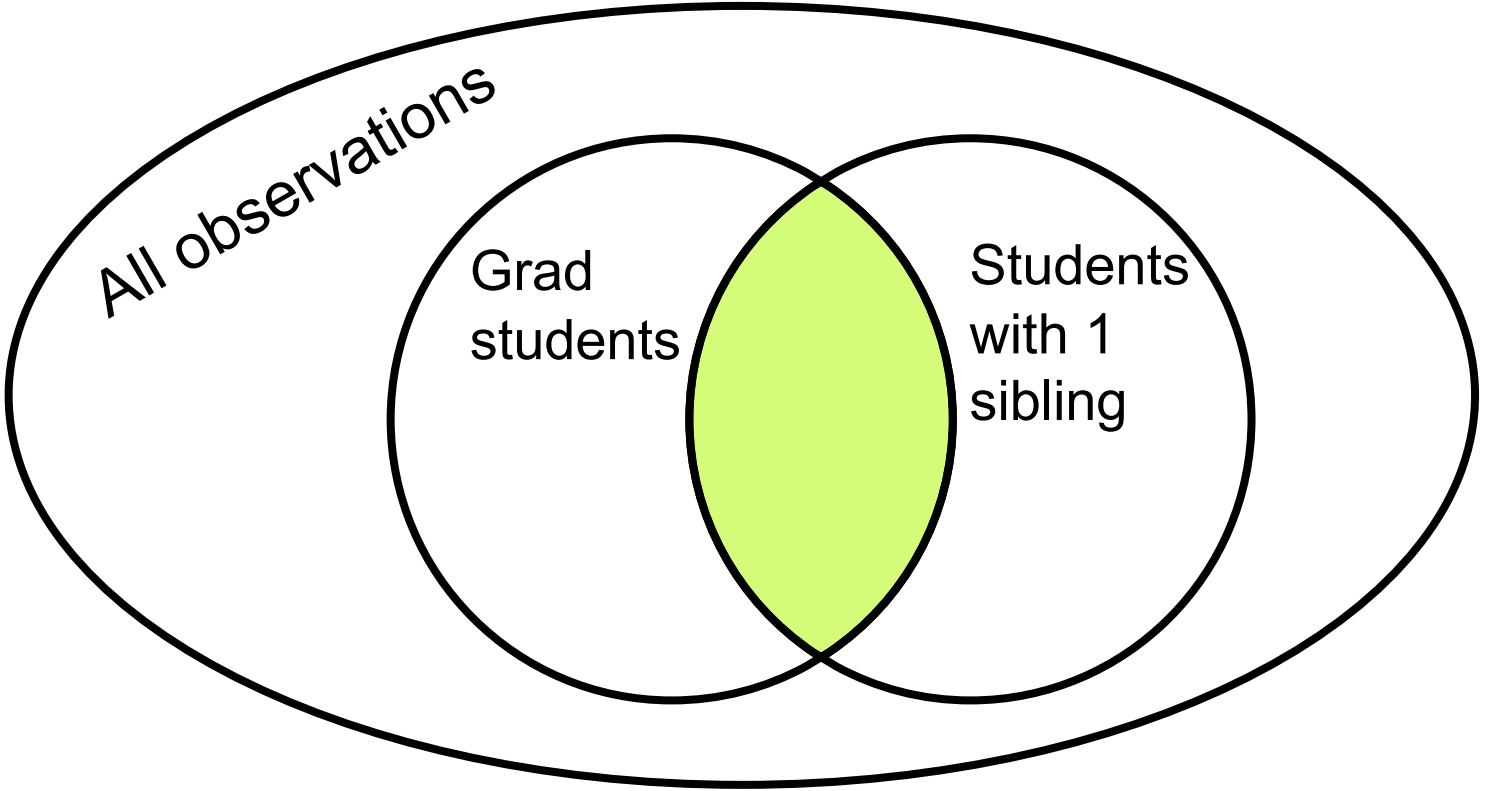
if year_school is equal to 6 OR siblings is equal to 1

```
if year_school==6 | siblings==1
```

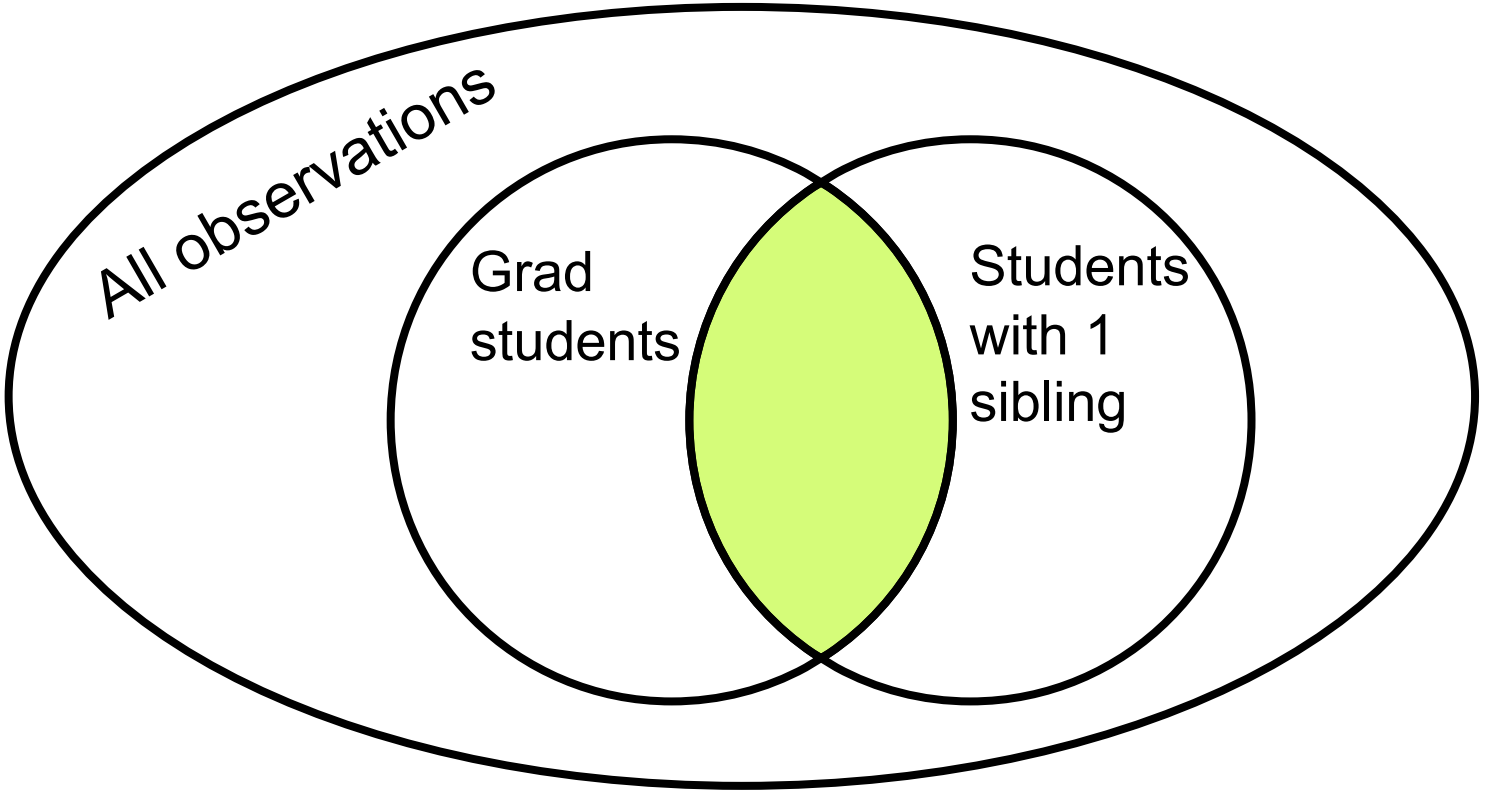



All observations
Grad students
Students with 1 sibling

Look at people **if** they are a grad student AND they have 1 sibling

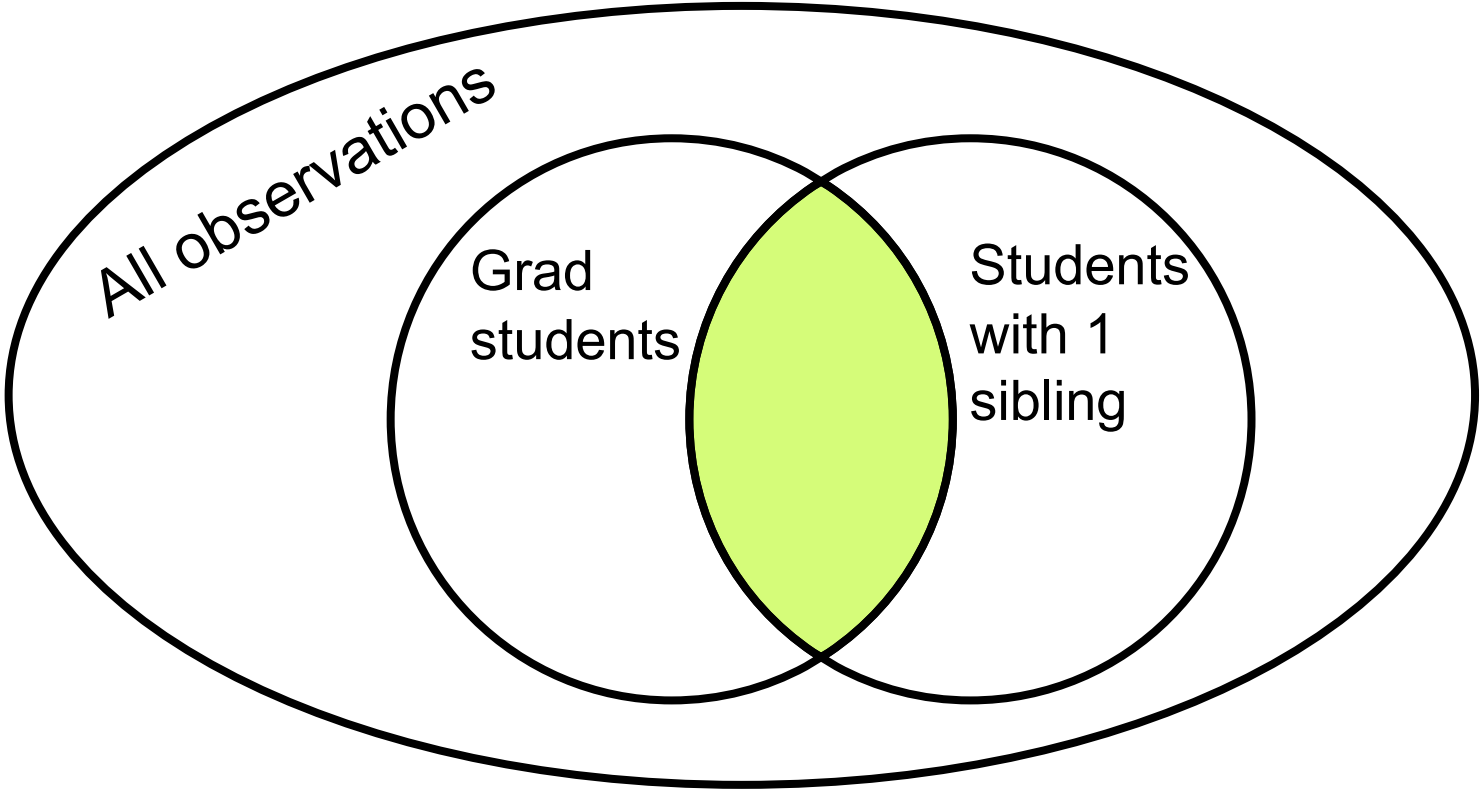


if year_school is equal to 6 AND siblings is equal to 1



if year_school is equal to 6 AND siblings is equal to 1

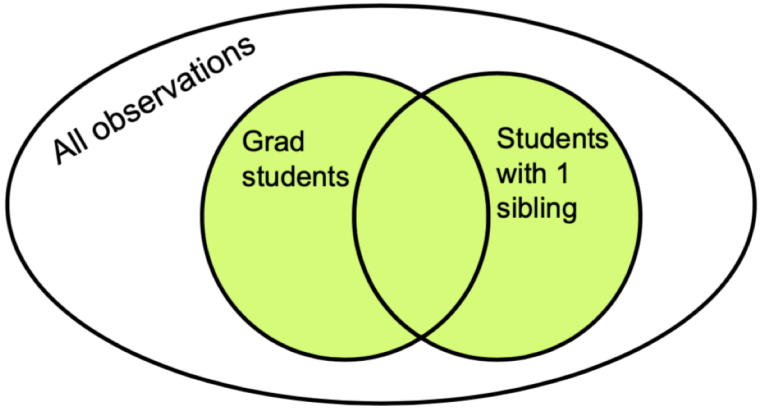
```
if year_school==6 & siblings==1
```



SUMMARY: Using logical statements to subset

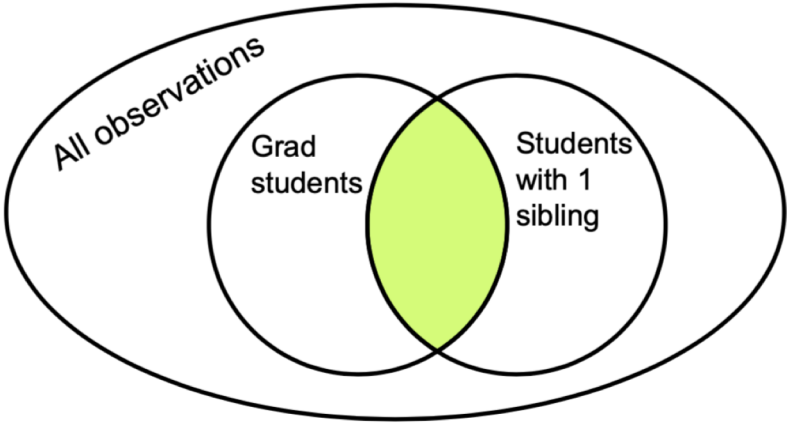
"OR": |

Must meet *at least 1* criteria



"AND": &

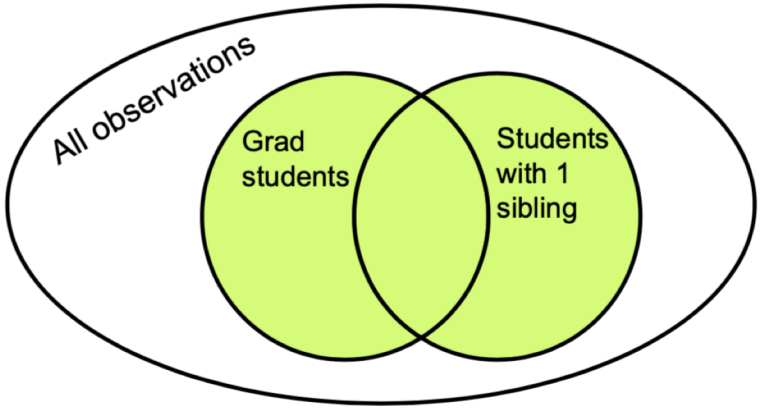
Must meet *all* criteria



SUMMARY: Using logical statements to subset

"OR": |

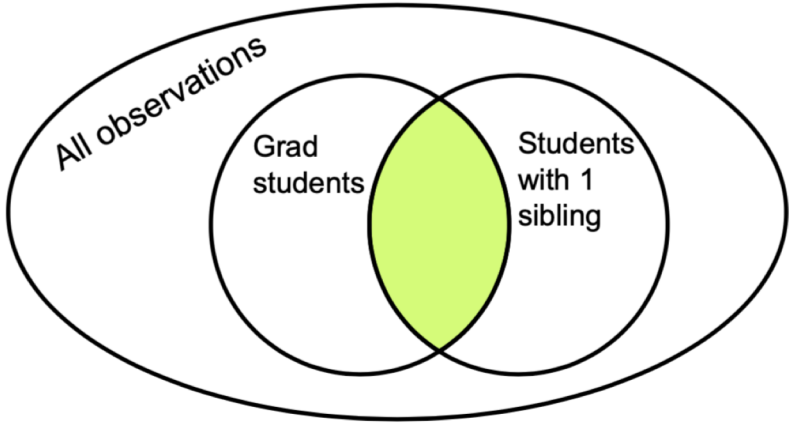
Must meet *at least 1* criteria



think: ***all*** *areas*

"AND": &

Must meet *all* criteria

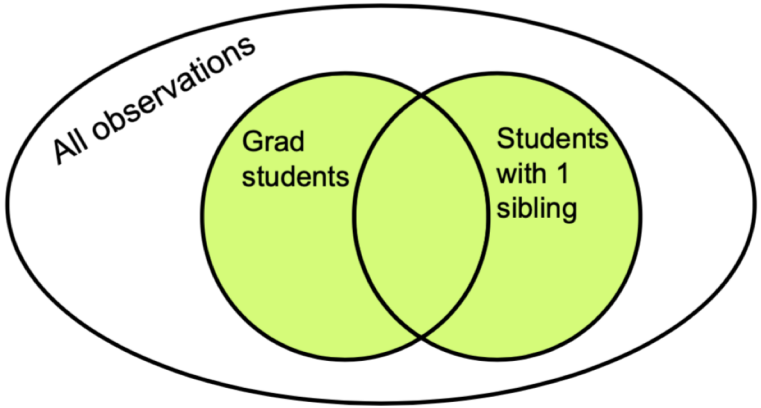


think: ***overlapping*** *areas*

ACTIVITY: “Practice subsetting observations” #1-13

“OR”: |

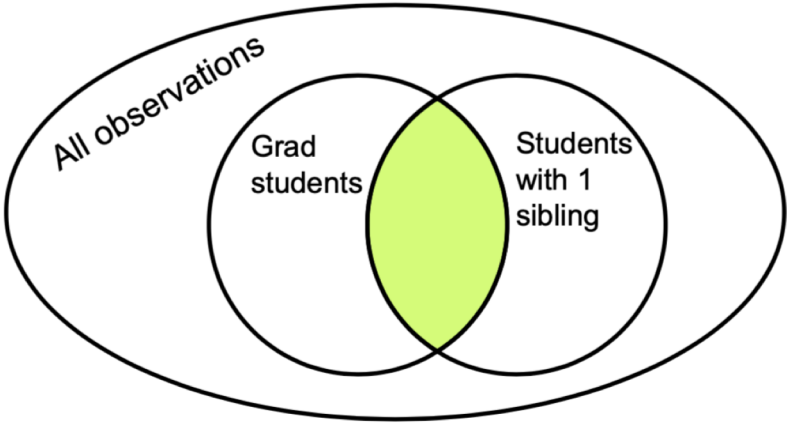
Must meet *at least 1* criteria



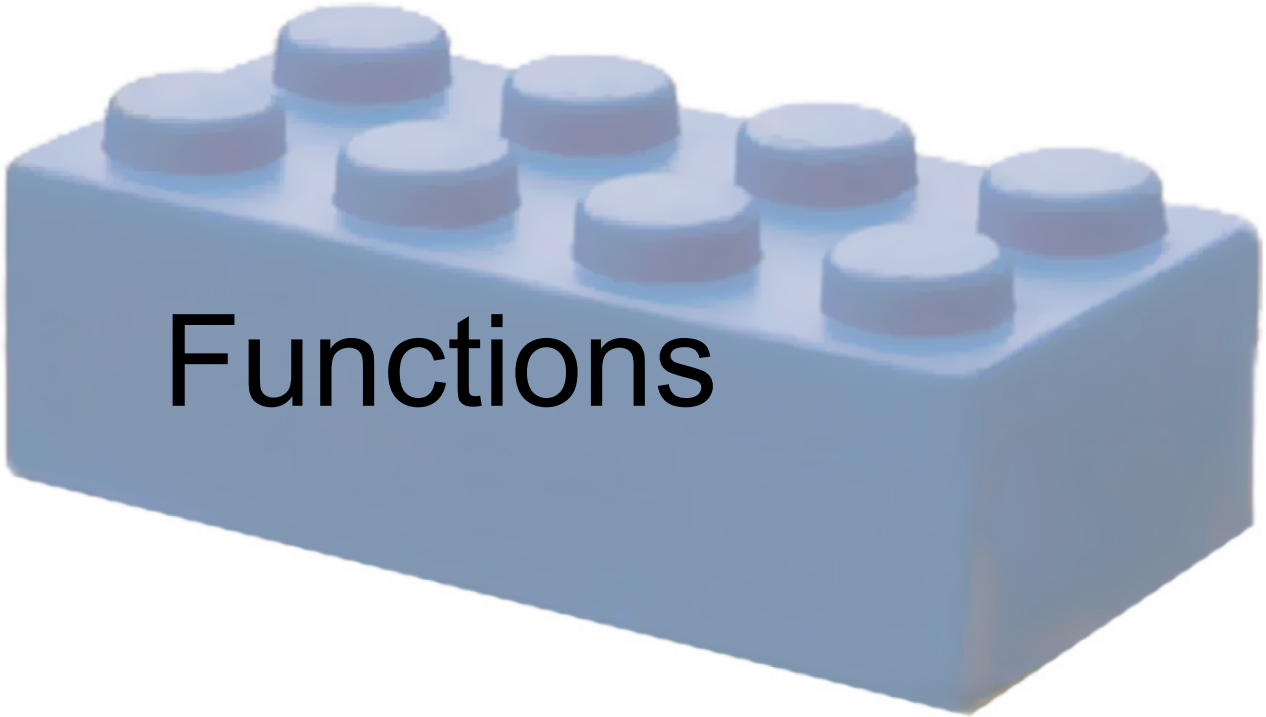
think: ***all*** *areas*

“AND”: &

Must meet *all* criteria



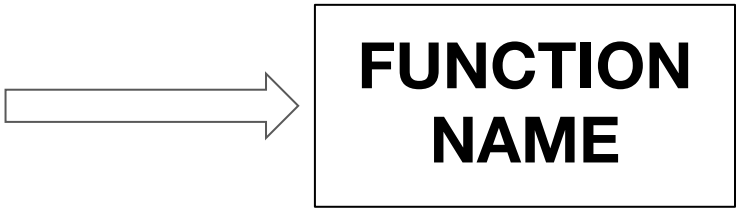
think: ***overlapping*** *areas*


Functions

Functions

Functions

Information included in input code



Functions


Information included in input code
FUNCTION NAME
Information included in the output

Example. We want to create a new variable that tells us whether someone is an only child. How would we do this?

Example. We want to create a new variable that tells us whether someone is an only child. How would we do this?

Answer: Use the variable *siblings* to create a different variable called *onlychild* where 1 = “is an only child” and 0 = “not an only child.”
(this is called a dummy variable)

To create new variables, we use two functions: **generate** and **replace**.

	major	year_school	regions	siblings	height	temp	F_C	cheese
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
7		Co-term		0	77	0	C	Gouda
8		Sophomore	South	1	88	.	.	
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya

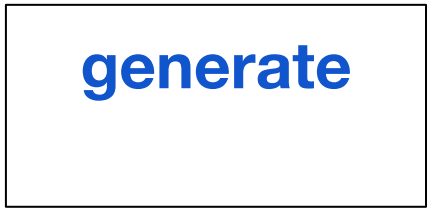
generate onlychild = 0

	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	0
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

replace onlychild = 1 if siblings==0

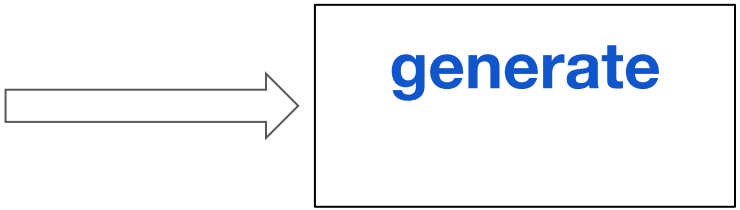
	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	1
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

generate



`generate onlychild`

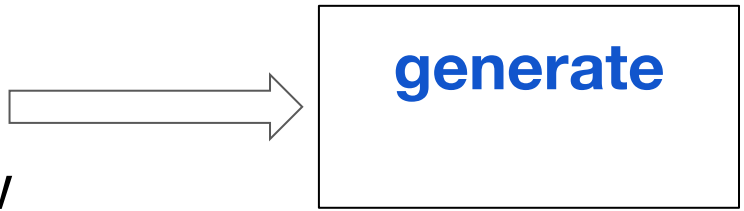
New variable name



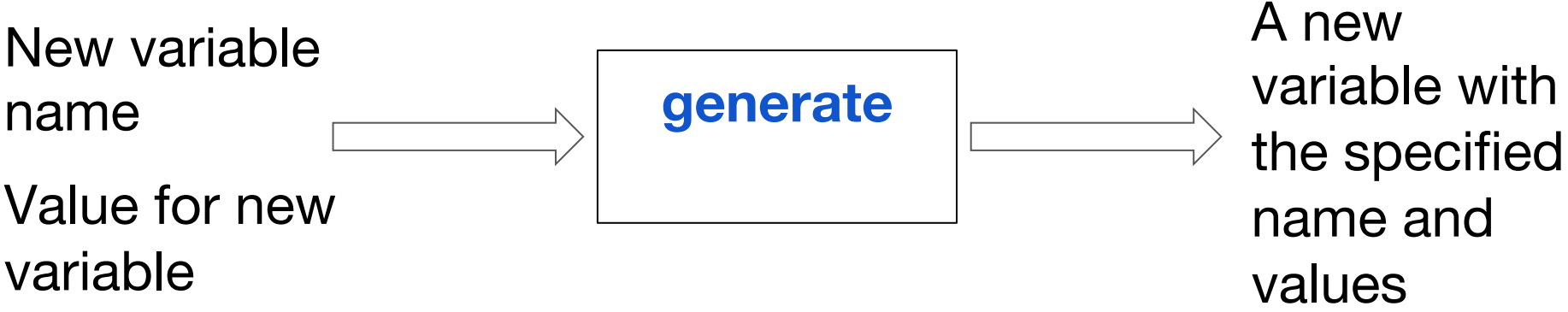
```
generate onlychild = 0
```

New variable name

Value for new variable



```
generate onlychild = 0
```



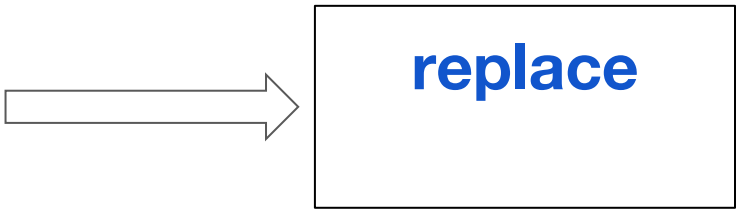
generate onlychild = 0

	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	0
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

replace

`replace onlychild`

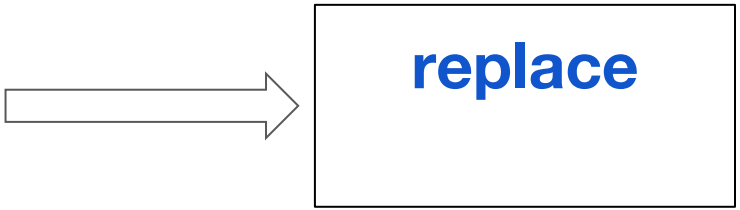
Variable we want to change



```
replace onlychild = 1
```

Variable we want to change

New value

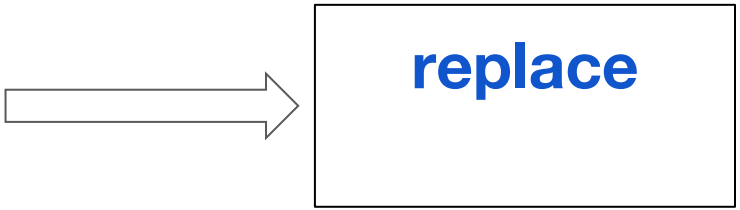


```
replace onlychild = 1 if siblings==0
```

Variable we want to change

New value

If statement specifying a subset



```
replace onlychild = 1 if siblings==0
```

Variable we want to change

New value

If statement specifying a subset

The specified subset of that variable will take on the new value

replace onlychild = 1 if siblings==0

	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	1
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

ACTIVITY: “Generate and replace” #14-18

New variable name

Value for new variable

A new variable with the specified name and values

Variable we want to change

New value

If statement specifying a subset

The specified subset of that variable will take on the new value

Part 3. Self-Directed Worksheet