

2. Describing Univariate Distributions (I)

LPO.8800: Statistical Methods for Education Research

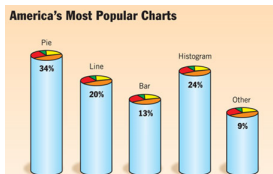
Sean P. Corcoran

Last time

- Descriptive vs. inferential statistics
- Basic concepts: outcomes, variables, unit of observation, population, sample
- Measurement scales
 - ▶ quantitative or categorical
 - ▶ nominal, ordinal, interval, ratio
 - ▶ discrete vs. continuous
- Sampling methods

Today

- Stata introduction (in brief—see video for more)
- Describing univariate distributions: categorical and quantitative data
- Choice of statistical tools used to describe a variable depends in part on how it is measured



Today

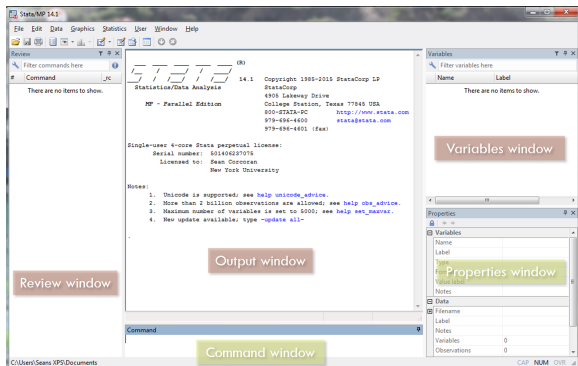
- Describing categorical variables
 - ▶ Frequency (and relative frequency) distributions
 - ▶ Bar graphs
 - ▶ Pie graphs
- Describing quantitative variables
 - ▶ Frequency (and relative frequency) distributions, possibly grouped
 - ▶ Histograms
 - ▶ Stem-and-leaf plot
 - ▶ Measures of central tendency

Stata introduction

Interacting with Stata

- Command window (*interactive mode*) vs. **do-file editor** (*batch mode*)
- Review window
- Variables and properties windows
- Results window
- Menu and task bar commands

Stata windows



Basic Stata tasks

- Opening and saving data (.dta) files
- Removing a data file from memory
- Data browser and editor
- Using Stata as a calculator
- Getting help
- *Variables* vs. cases/observations
- *Variable* labels vs. *value* labels

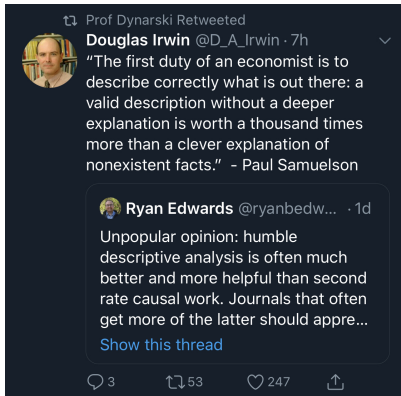
Stata command syntax

Example:

```
summarize varlist [if] [in] [weight] [,options]
```

- `summarize`: the command
- *varlist*: terms in italics are things you provide
- Syntax in brackets [] is *optional*
- [,options]: Most commands have other options that are specified *after* a comma, at the very end
- [if]: execute the command only if a certain condition is true
- [in]: execute the command for a certain subset of observation numbers

Importance of descriptive analyses



Importance of descriptive analyses

See the IES report by Loeb et al. (2017), which describes the role of descriptive analysis in education and social science.

Key Themes

- Descriptive analysis characterizes the world or a phenomenon—answering questions about who, what, where, when, and to what extent. Whether the goal is to identify and describe trends and variation in populations, create new measures of key phenomena, or describe samples in studies aimed at identifying causal effects, description plays a critical role in the scientific process in general and education research in particular.
- Descriptive analysis stands on its own as a research product, such as when it identifies socially important phenomena that have not previously been recognized. In many instances, description can also point toward causal understanding and to the mechanisms behind causal relationships.
- No matter how significant a researcher's findings might be, they contribute to knowledge and practice only when others read and understand the conclusions. Part of the researcher's job and expertise is to use appropriate analytical, communication, and data visualization methods to translate raw data into reported findings in a format that is useful for each intended audience.

Importance of descriptive analyses

Some examples of excellent, influential descriptive studies in education:

- Arnold et al. (2009) - on “summer melt”
- Scott-Clayton (2012) - changes over time in undergraduates’ propensity to work while in school
- Reardon (2011) - on changes in the academic achievement gap between high- and low-income students
- Lankford, Loeb, & Wyckoff (2002) - on the distribution of teacher qualities across New York State districts and schools
- Hoxby & Avery (2003) - on the “missing one-offs”
- Murnane (2013) - on long-run trends in U.S. HS graduation rates

Importance of descriptive analyses

DATA



SORTED



ARRANGED



PRESENTED
VISUALLY



EXPLAINED
WITH A STORY



Frequency distributions

- A **frequency distribution** is a table showing the number (count) of occurrences of each unique outcome in the data.
- The **relative frequency** of an outcome or category is the *proportion* or *percentage* of all observations in that category. (Must sum to one, or 100%).

Frequency distributions

Example: test scores for 16 students

Table: Classroom test scores

Score	Frequency (Count of students)
5	2
10	3
12	1
15	4
20	4
25	2
Total	16

Frequency distributions

Example: test scores for 16 students

Table: Classroom test scores

Score	Frequency (Count of students)	Relative Frequency (Percentage)
5	2	$2/16 * 100 = 12.5\%$
10	3	$3/16 * 100 = 18.75\%$
12	1	$1/16 * 100 = 6.25\%$
15	4	$4/16 * 100 = 25.0\%$
20	4	$4/16 * 100 = 25.0\%$
25	2	$2/16 * 100 = 12.5\%$
Total	16	$16/16 * 100 = 100\%$

Frequency distributions

Side note on terminology:

- The fractions in the above relative frequency distribution: $2/16$, $3/16$, $1/16$ —or, 0.125, 0.1875, 0.0625, etc.—are **proportions**: the frequency of cases in a given category divided by the total number of cases in all categories. Ranges between zero and one.
- Multiply by 100 to get **percentages** (12.5%, 18.75%, 6.25%, etc.)
- Relative frequencies can be expressed either way

Pro tip: see handout on **rounding conventions**. Percentages are typically rounded to one decimal place; proportions typically rounded to three.

Frequency distributions

- Because frequency distributions list every distinct value in the data, they are only practical for variables with a limited number of unique values
 - ▶ Categorical variables
 - ▶ Discrete quantitative variables
- It is possible to *group* variables with many distinct outcomes into smaller categories (shown later)
 - ▶ Continuous quantitative variables

Frequency distributions

Easy to generate in Stata using `tabulate`

. tabulate region			
geographic region of school	Freq.	Percent	Cum.
northeast	106	21.20	21.20
north central	151	30.20	51.40
south	150	30.00	81.40
west	93	18.60	100.00
Total	500	100.00	

Frequency distributions

Note the *cumulative* column is not very meaningful here (a categorical, non-ordered variable).

. tabulate region			
geographic region of school	Freq.	Percent	Cum.
northeast	106	21.20	21.20
north central	151	30.20	51.40
south	150	30.00	81.40
west	93	18.60	100.00
Total	500	100.00	

Frequency distributions

In SPSS: “valid percent” expresses relative frequency as a percentage of all *non-missing* observations. In this example there is a **missing value code** of 98.

Parents' Marital Status in Eighth Grade					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Divorced	26	5.2	5.5	5.5
	Widowed	6	1.2	1.3	6.7
	Separated	8	1.6	1.7	8.4
	Never Married	5	1.0	1.0	9.4
	Marriage-Like Relationship	5	1.0	1.0	10.5
	Married	427	85.4	89.5	100.0
	Total	477	95.4	100.0	
Missing	98	23	4.6		
	Total	500	100.0		

Frequency distributions

To show the count of missing values in Stata, include the missing option:
`tabulate varname, missing`

```
. tabulate parmar18, missing
```

parents' marital status in eighth grade	Freq.	Percent	Cum.
divorced	26	5.20	5.20
widowed	6	1.20	6.40
separated	8	1.60	8.00
never married	5	1.00	9.00
marriage-like relationship	5	1.00	10.00
married	427	85.40	95.40
.	23	4.60	100.00
Total	500	100.00	

Frequency distributions

Also see the `fre` command which shows the relative frequency with and without missing values:

```
. fre parmar
```

parmar18 — parents' marital status in eighth grade

	Freq.	Percent	Valid	Cum.
Valid 1 divorced	26	5.20	5.45	5.45
2 widowed	6	1.20	1.26	6.71
3 separated	8	1.60	1.68	8.39
4 never married	5	1.00	1.05	9.43
5 marriage-like relationship	5	1.00	1.05	10.48
6 married	427	85.40	89.52	100.00
Total	477	95.40	100.00	
Missing .	23	4.60		
Total	500	100.00		

`fre` is also nice in that it shows you both variable *values* and *labels*.

Frequency distributions

The `table` command provides a simpler frequency distribution (without the percent or cumulative percent). This command allows for many options for customizing the contents of the table.

```
. table parmar18
```

parents' marital status in eighth grade	Freq.
divorced	26
widowed	6
separated	8
never married	5
marriage-like relationship	5
married	427

Frequency distributions

Note the categories used in each of these variables are mutually exclusive and collectively exhaustive:

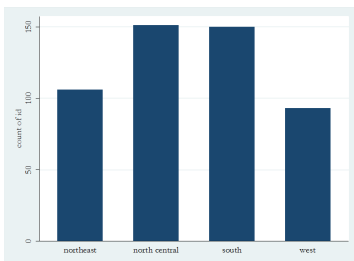
- **mutually exclusive:** being in one category precludes being in another
- **collectively exhaustive:** all possible categories are represented by the defined categories

Bar graphs

- **Bar graphs** are a visual way to display frequency (or relative frequency) distributions
- In Stata: `graph bar` can be used for frequency (or relative frequency) distributions. An alternative is `histogram` with the `discrete` option.
- Try: `region`, `parmarl8`, `advmath8`

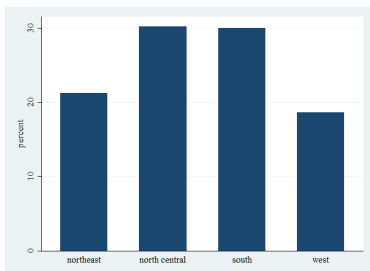
Bar graphs

`graph bar (count), over(region)`



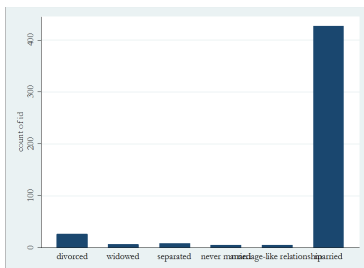
Bar graphs

`graph bar (percent), over(region)`



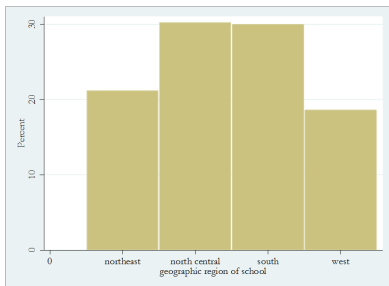
Bar graphs

`graph bar (count), over(parmar18)`



Bar graphs

```
histogram region, discrete percent gap(2) xlabel( ,  
valuelabel)
```



Bar graphs

- Word of caution with bar graphs: always check where vertical axis begins—does it begin at zero? Avoid misleading scales
- Note “bar graphs” (as opposed to histograms) have gaps between the bars, suggestive of distinct categories
- Note The `gap(2)` option in `histogram` forces a gap of size 2

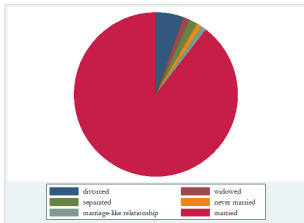
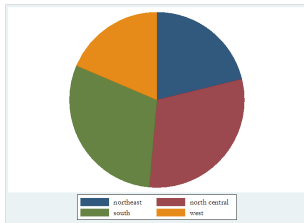
Pie graphs

Pie graphs can be used to show the relative frequency of a variable

- These only make sense when the variable has *collectively exhaustive* (and a limited number of) categories
- In Stata: `graph pie`
- Try: `region`, `parmarl8`

Pie graphs

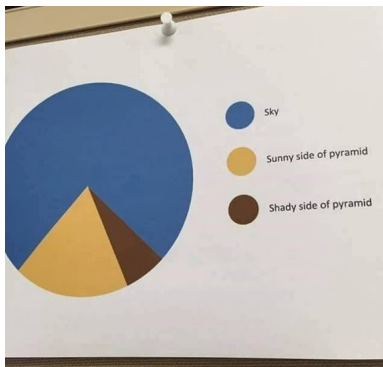
`graph pie, over(region)`



Pie graphs



Pie graphs



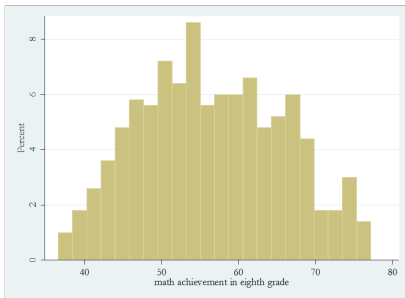
Histograms

A **histogram** is a bar graph where the height of each bar represents the count (or percent) of observations within a given *range* of values, called an **interval** or **bin**

- Number of bins determined by default in Stata, but can be adjusted
- Obviously makes sense only for interval (or ratio) measured variables, where a range of values is meaningful
- In Stata: `histogram`

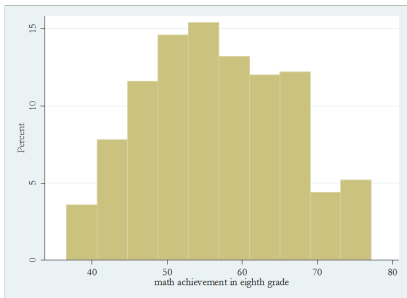
Histograms

`histogram achmat08, percent`



Histograms

```
histogram achmat08, percent bin(10)
```



Stem-and-leaf plot

A **stem-and-leaf** plot is similar to a histogram, but provides a bit more detail on specific values in the data

- Leading digits are called “stems”
- Trailing digits are called “leaves”
- The number of leaves corresponds to the frequency of a particular value

Stem-and-leaf plot

Stem-and-leaf plot for *achmat08* (math achievement in eighth grade)

achmat08 rounded to nearest multiple of .1
plot in units of .1

```
36* 6
37* 122
38* 478
39* 0357
40* 0225555799
41* 34689
42* 0335577
43* 000133445689
44* 00012244467788
45* 334467778899
46* 3445555667788889
47* 002344456677888
48* 0000013867889
49* 022233444666889
50* 001112356677888899
51* 0000111233334455779
52* 111112223334455668889
53* 00002456677889
54* 00011223334445555778888999
55* 00000003345678899
56* 11223346777888889
57* 0000112233344444778
58* 0111334677888899999
59* 034446679
60* 00122334556677888
61* 000112333345689
62* 0001111233345555889
63* 2223677777779
64* 01122566788899
65* 2234567789
66* 0001122333344667799
67* 122333456677889
68* 0022346666899
69* 012244679
70* 0114899
71* 389
72* 146
73* 3335788
74* 02233334
75* 00014
76*
77* 222222
```

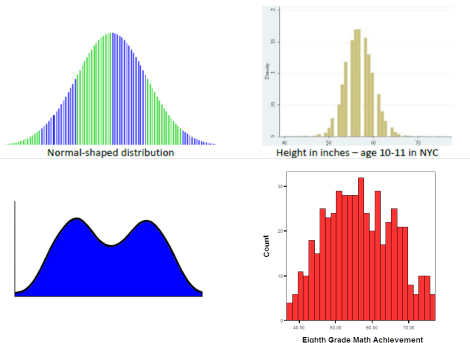
Stem-and-leaf plot

- In the above example, Stata rounded to the nearest 0.1
- First two digits used as stem, and final digit used as leaf
- Leaves correspond to frequency of particular values
- Compare the shape of the stem-and-leaf plot for *actmat08* to that of the histogram

Shape of distributions

- The histogram (and stem-and-leaf plots) are revealing about the *shape* of a distribution—i.e. which values tend to be more or less frequent
- A distribution is **symmetric** if the distribution of outcomes is identical (or approximately identical) on either side of its central value
- Examples: normal distribution (bell curve), U-shaped distribution, bi-modal distribution, uniform distribution

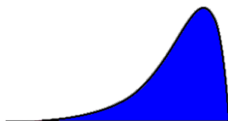
Shape of distributions



Shape of distributions

- A distribution is **skewed left** (or **negatively skewed**) if the distribution has a long tail to the left of its central value
- A distribution is **skewed right** (or **positively skewed**) if the distribution has a long tail to the right of its central value

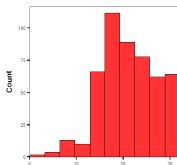
Left-skewed distributions



Left-skewed distribution

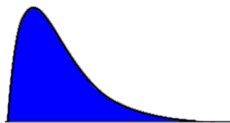


Texas standardized math scores, 5th grade (2000)

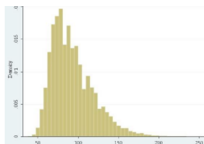


Eighth Grade Self-Concept

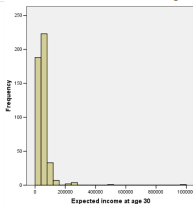
Right-skewed distributions



Right-skewed distribution



Weight in pounds – age 10-11 in NYC



Grouped frequency distributions

- Again, frequency distributions are less useful for variables with *many* distinct possible values (e.g. continuous variables)
- For continuous variables, one could create a smaller number of equal width groups (bins, or intervals) and then create a frequency distribution for this grouped (“re-coded”) variable. (This is what the histogram does behind the scenes).

Grouped frequency distributions

- NELS example: *unitmath* is the number of units of high school math taken, and ranges from 1-6. Includes many fractional units.
- Can set up groups or intervals, for example:
 - ▶ $1 \leq x < 2$
 - ▶ $2 \leq x < 3$
 - ▶ $3 \leq x < 4$, and so on
- Lecture 3 will show how to re-code variables in this way (a kind of data transformation)

Measures of central tendency

Measures of central tendency characterize the “center” or “location” of a distribution, its “typical” or “expected” value. Examples:

- Mean
- Median
- Mode

Mean

The **mean** (or *average*) adds all of the observed values and divides by the number of observations n

- Let $x_1, x_2, x_3, \dots, x_n$ represent the n values of a variable x (x_i is the i th observation, and i is the *index*)
- Then the **mean** is: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

Mean

Example: calculating the mean number of wins for baseball teams

Table 2.2: American League Standings, July 28, 2013

East	W	L	PCT
Tampa Bay	62	42	0.596
Boston	62	43	0.590
Baltimore	58	47	0.552
NY Yankees	54	50	0.519
Toronto	47	56	0.456
Central	W	L	PCT
Detroit	58	45	0.563
Cleveland	55	48	0.534
Kansas City	50	51	0.495
Minnesota	45	56	0.446
Chi White Sox	40	61	0.396
West	W	L	PCT
Oakland	61	43	0.587
Texas	56	48	0.538
Seattle	49	55	0.471
LA Angels	48	54	0.471
Houston	35	68	0.340
$\sum W$	780		

Mean

Example: calculating the mean number of wins for baseball teams

$$\overline{W} = \frac{\sum W_i}{n} = \frac{780}{15} = 52$$

Mean

The mean in Stata can be calculated using several commands, including `summarize` (or `sum`).

```
. sum achmat08
```

Variable	Obs	Mean	Std. Dev.	Min	Max
achmat08	500	56.59102	9.339608	36.61	77.2

```
. sum expinc
```

variable	obs	Mean	Std. Dev.	Min	Max
expinc30	459	51574.73	58265.76	0	1000000

Mean

The mean of a categorical variable is usually meaningless, except in the case of a *dichotomous* variable coded 0-1, in which case the mean is the proportion equal to 1:

```
. sum advmath8
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
advmath8	491	.4602851	.4989286	0	1

Mean

The mean is highly influenced by extreme values or **outliers**: observations that fall well above or well below the bulk of the data.

- Example 1 ($n = 15$)
 - ▶ 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5
 - ▶ mean = $(55/15) = 3.67$
- Example 2 ($n = 15$)
 - ▶ 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5000000
 - ▶ mean = $(5000050/15) = 333,336.67$

Mean

The mean can be characterized as the “center of gravity,” or balance point, of the distribution. It is the point at which the sum of the distances to the mean from observations *above* the mean equal the sum of the distances to the mean from observations *below* the mean

Mean

Deviations from the mean:

Table 2.3: Deviations above and below the mean

	W	$W - \bar{W}$	Totals
Tampa Bay	62	10	
Boston	62	10	
Oakland	61	9	
Baltimore	58	6	
Detroit	58	6	
Texas	56	4	
Cleveland	55	3	
NY Yankees	54	2	50
Kansas City	50	-2	
Seattle	49	-3	
LA Angels	48	-4	
Toronto	47	-5	
Minnesota	45	-7	
Chi White Sox	40	-12	
Houston	35	-17	-50

Mean

The **least squares principle**:

- The *average deviation* of x from its mean will always be zero. That is, the sum of negative deviations from the mean will always equal the sum of positive deviations from the mean.
- The mean is the point in a distribution around which the variation is minimized (as indicated by the *squared* differences): $\sum (x_i - \bar{x})^2$ (proof later)

Median

The **median** is the observation that falls in the middle of the data, when the observations are ordered from lowest to highest values.

- When n is *odd*: a single value will fall in the middle
- When n is *even*: the median is the midpoint of the two middle values
- Alternatively, index the ordered n values from 1 to n . The median will be the value with index $(n + 1)/2$

Median

- Example 1 ($n = 15$):
 - ▶ 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5
 - ▶ median = 4
- Example 2 ($n = 6$):
 - ▶ 0, 1000, 1000, 5000, 6000, 8000
 - ▶ median = $(1000 + 5000)/2 = 3000$

Median

The median in Stata can be calculated using several commands, including `summarize` (or `sum`) with the `detail` option:

```
. sum achmat08, detail
```

math achievement in eighth grade			
Percentiles		Smallest	
1%	38.55	36.61	
5%	41.89	37.14	
10%	44.185	37.2	obs
25%	49.42	37.24	Sum of wgt. 500
50%	56.18		Mean 56.59102
75%	63.74	Largest 77.2	Std. Dev. 9.339608
90%	68.935	77.2	variance 87.22827
95%	73.33	77.2	skewness .1133238
99%	77.2	77.2	kurtosis 2.242742

Median

Because the median is simply the middle value, it is insensitive (“robust”) to extreme values or outliers in the distribution

- Example 1 ($n = 15$):
 - ▶ 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5000000
 - ▶ median = 4

Mode

The **mode** is the outcome that occurs most often in a distribution

- Most appropriate for highly discrete variables, such as categorical variables
- Variables with lots of unique values (such as continuous variables) tend to have few repeats, and thus the mode is not that meaningful
- There is no command in Stata specifically for obtaining the mode. However, can use a frequency distribution or other combinations of commands (like `egen`).

Mode

```
. ** Find mode using tabulate with sort (for descending sort)
```

```
. tabulate famsize, sort
```

family size	Freq.	Percent	Cum.
4	199	39.80	39.80
5	142	28.40	68.20
6	55	11.00	79.20
3	52	10.40	89.60
7	21	4.20	93.80
9	13	2.60	96.40
2	9	1.80	98.20
8	9	1.80	100.00
Total	500	100.00	

```
. ** Find mode using egen (note you will get a message if >1 mode)
```

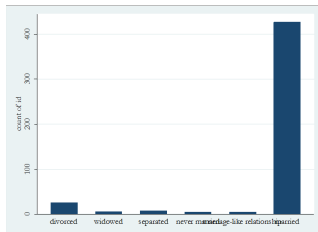
```
. egen mode=mode(famsize)
```

```
. table mode
```

mode	Freq.
4	500

Mode

“Married” is the modal parents’ marital status. One might say the “typical” 8th grade student in the NELS has married parents.



Mode

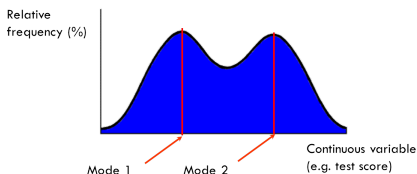
North Central is (technically) the modal region. But region here is more accurately described as *bimodal*—it has a distribution with two values that occur most often (North Central and South).

```
. tabulate region
```

geographic region of school	Freq.	Percent	Cum.
northeast	106	21.20	21.20
north central	151	30.20	51.40
south	150	30.00	81.40
west	93	18.60	100.00
Total	500	100.00	

Mode

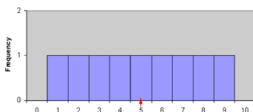
Another *bimodal* distribution



Mode

Some problems with using the mode:

- It is not very useful for “flat” distributions (e.g. the *uniform* distribution)
- Example 1: 1, 1, 2, 2, 3, 3, 4, 4, 5, 5
- Example 2:



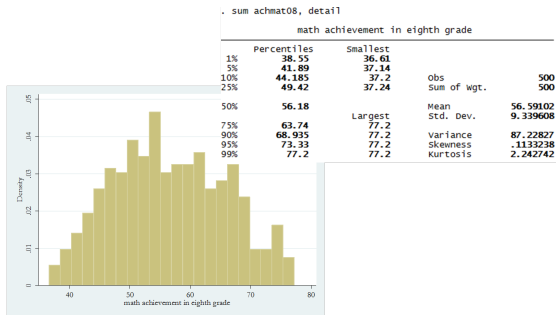
Comparing measures of central tendency

How will the mode, median, and mean usually compare? It depends on the shape of the distribution.

- For *symmetric* distributions the median \approx mean
- If symmetric and *unimodal*, median \approx mean \approx mode (when the mode is meaningful, distributions with a limited number of unique values)

Comparing measures of central tendency

Example: note the mode of this distribution is 72 (not shown). Illustrates the problem of using the mode with a continuous variable.



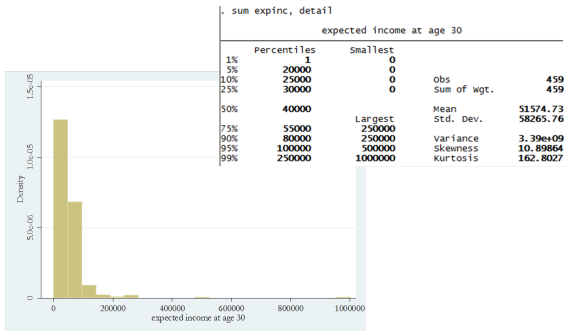
Comparing measures of central tendency

How will the mode, median, and mean usually compare? It depends on the shape of the distribution.

- For *right-skewed* distributions the mean $>$ median
- For *left-skewed* distributions the mean $<$ median

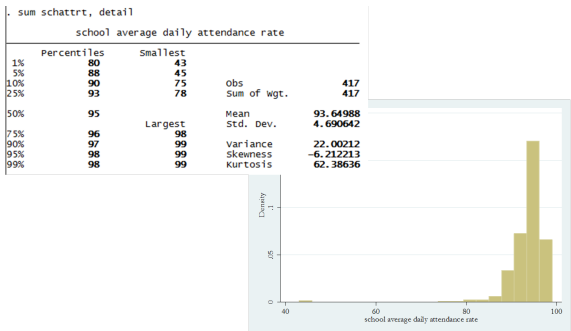
Comparing measures of central tendency

A right-skewed variable with some large positive outliers:



Comparing measures of central tendency

A left-skewed variable with some low-value outliers:



Measuring the Black-white wealth gap

The True Cost of Closing the Racial Wealth Gap

<https://www.nytimes.com/2021/04/30/business/racial-wealth-gap.html>

Using data from the 2019 Survey of Consumer Finances:

- Median Black household wealth: \$24,100
- Median white household wealth: \$188,200
- Gap: **\$164,100**
- Mean Black household wealth: \$142,500
- Mean white household wealth: \$983,400
- Gap: **\$840,000**
- 97% of white households' total wealth is held by households above the median

Alternative command

An alternative command in Stata for measures of central tendency (and other statistics):

- `tabstat achrrdg*, stat(mean p50 n)`
- `tabstat achrrdg*, stat(mean p50 n) col(stat)`

```
. tabstat achrrdg*, stat(mean p50 n)
```

stats	achr ^{rdg} 08	achr ^{rdg} 10	achr ^{rdg} 12
mean	56.04906	56.11404	55.60188
p50	56.445	57.545	57.005
N	500	500	500

```
. tabstat achrrdg*, stat(mean p50 n) col(stat)
```

variable	mean	p50	N
achr ^{rdg} 08	56.04906	56.445	500
achr ^{rdg} 10	56.11404	57.545	500
achr ^{rdg} 12	55.60188	57.005	500

A bit more on the summation operator

The mean is written as: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

- \sum is the summation operator
- i is the index of summation
- 1 and n are the lower and upper limit of the summation (i.e., summing the numbers x_i for all values of i from 1 to n)

Three properties of the summation operator

The summation operator has three properties:

- For any constant c : $\sum_{i=1}^n c = nc$
- For any constant c : $\sum_{i=1}^n cx_i = c \sum_{i=1}^n x_i$
- For any constants a and b : $\sum_{i=1}^n (ax_i + by_i) = a \sum_{i=1}^n x_i + b \sum_{i=1}^n y_i$

Example: \bar{x} is least squares

Given x_1, x_2, \dots, x_n , what value a minimizes the sum of squared differences between the x_i and a ?

$$\begin{aligned} & \min_a \sum_{i=1}^n (x_i - a)^2 \\ & \min_a \sum_{i=1}^n (x_i^2 - 2ax_i + a^2) \\ & \min_a \left(\sum_{i=1}^n x_i^2 - 2a \sum_{i=1}^n x_i + na^2 \right) \end{aligned}$$

Example: \bar{x} is least squares

Take the derivative with respect to a and solve for a :

First order condition:

$$\begin{aligned} -2 \sum_{i=1}^n x_i + 2na &= 0 \\ 2na &= 2 \sum_{i=1}^n x_i \\ a &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \end{aligned}$$

What not to do with the summation operator

Note the following, which are **not** properties of the summation operator:

$$\sum_{i=1}^n \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i}$$

$$\sum_{i=1}^n x_i^2 \neq \left(\sum_{i=1}^n x_i \right)^2$$

Example of a double summation

Consider two sets of numbers x_1, \dots, x_n and y_1, \dots, y_n . Use the index of summation i for x and the index j for y . The following is an example of a double summation:

$$\sum_{i=1}^n \sum_{j=1}^n x_i y_j$$

This can be written:

$$\sum_{i=1}^n x_i \sum_{j=1}^n y_j = x_1(y_1 + \dots + y_n) + x_2(y_1 + \dots + y_n) + \dots$$

Or:

$$x_1 y_1 + x_1 y_2 + x_1 y_3 + \dots + x_2 y_1 + x_2 y_2 + x_2 y_3 + \dots$$

Next lecture

- Univariate descriptive statistics, continued: measures of variability/dispersion, and skewness
- Measures of position in a distribution (e.g., quantiles, z-scores)
- Data transformations, and effects on descriptive statistics